TARGET-COHERENT RISK ENVELOPES: SHARP COMPLETION BOUNDS FOR OVERLAPPING FORECAST ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rolling multi-horizon forecasters issue new predictions before earlier targets are observed. Forecasts from different origins can therefore refer to the same absolute target. Standard itemwise worst-case evaluation treats these repeated predictions as if their shared label could vary independently, allowing mutually incompatible outcomes and overstating the remaining excess of issued corrections. We introduce the Target-Coherent Forecast-Space Risk Envelope (TC-FSRE), a safety layer that groups pending loss terms by absolute target and optimizes over one legal label for each group. Under the stated completion model, this construction gives the sharp bound for a fixed pending set. For normalized absolute loss, we further characterize exactly when the itemwise bound is conservative: the gap is positive when corrections sharing a target require conflicting labels to attain their separate worst cases. Online, TC-FSRE projects each proposed correction onto the feasible action interval. If adaptation is later disabled, previously issued corrections remain within the cumulative-excess constraint as their labels arrive. Across 11 FreTS-H192 and PatchTST-H720 dataset–backbone units, target-coherent evaluation reduced the same-action bound by 77.23% on average and increased mean issued action by 0.18184. All 72 certificate paths remained valid in matched tests with TAFAS, DSOF, and COSA candidates. The guarantee concerns cumulative excess under the specified loss and completion model; candidate forecast quality remains a separate empirical question.

1 INTRODUCTION

Pretrained time-series forecasters encounter nonstationarity after deployment, while retraining the full model can be costly. Recent test-time methods instead use newly observed data to update a calibration module, fit a lightweight online predictor, or correct the forecast directly in output space (Kim et al., 2025; Lee et al., 2025; Im & Kwon, 2026). These methods address how to construct a candidate forecast for the current distribution. The quality of that candidate, however, varies with the dataset, backbone, and form of drift. An adapter that helps on average can still cause marked negative transfer on particular tasks or intervals.

Rolling multi-horizon prediction adds a temporal complication. A forecast issued at time t receives its H targets over subsequent time steps, while later forecasts are issued before that feedback is complete. The system therefore holds issued, irreversible forecast corrections with unobserved labels. Delayed-feedback theory primarily provides regret or tracking guarantees as feedback is postponed (Joulani et al., 2013; Ryabchenko et al., 2025); its theorem object is not the remaining deterministic excess of an already-issued overlapping forecast set.

Delay length alone is insufficient: sharp completion evaluation also requires *target identity*. Forecast terms from different origins can refer to the same absolute time and channel. Maximizing each term separately allows that shared target to take a different value for each forecast. The resulting product completion is a valid relaxation, but it contains label assignments that cannot occur together and can be loose. Conversely, merging genuinely distinct targets can underestimate the excess and produce a constructive violation. The correct completion rule therefore groups unresolved terms by the target they share and maximizes over one legal label per target.

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

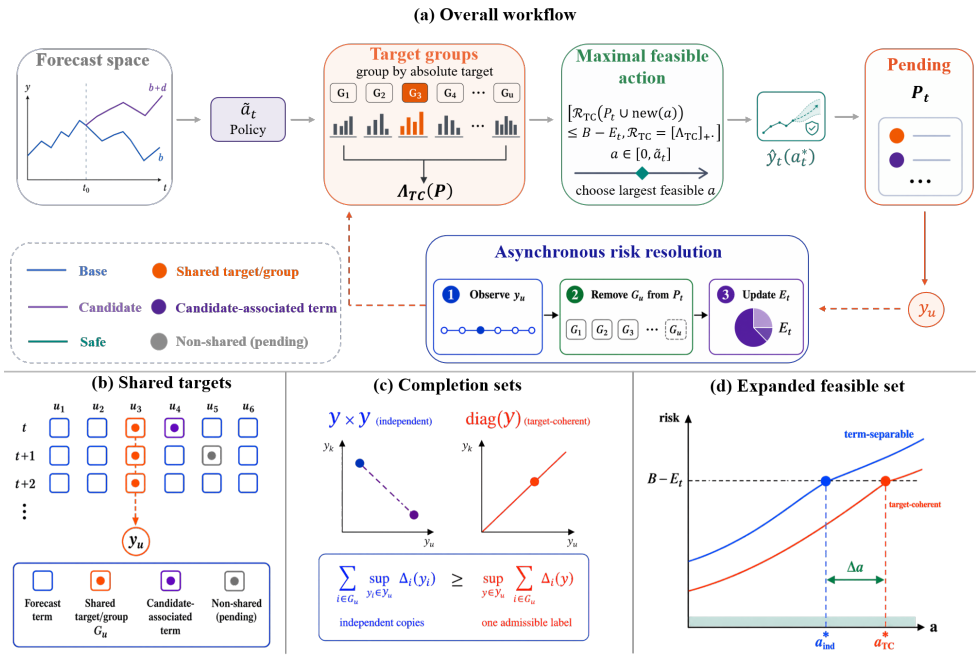


Figure 1: Overview of target-coherent forecast-space adaptation. **a**, A frozen forecast, candidate direction, and proposed action enter the risk layer; arriving labels update pending target groups. **b**, Rolling origins create multiple terms for one absolute target. **c**, Product and shared-target completion sets yield different sharp values. **d**, The smaller target-coherent value admits a larger action under the same cumulative-excess threshold.

We answer this question with a Forecast-Space Risk Envelope (FSRE) and its target-coherent realization, TC-FSRE. A frozen forecaster supplies a base trajectory, an interface-compatible adapter supplies a candidate direction, and an FTL, Hedge, or AdaHedge rule supplies a proposed action. TC-FSRE evaluates the joint worst-case excess of the pending set, projects the proposed action onto the feasible interval, and removes resolved terms when their targets arrive. This separation distinguishes candidate quality from how much adaptation can be applied under a cumulative-excess constraint.

This target grouping yields a sharper certificate and can admit more of a proposed adaptation without changing the cumulative excess constraint. We characterize when the itemwise relaxation is strict and lift the fixed-set result to arbitrary finite sequences of issue and feedback events.

Our contributions are as follows.

- **Completion geometry.** We define the sharp value obtained by assigning one label to all unresolved terms for the same target. For NMAE, an exact gap decomposition characterizes itemwise conservatism; partition results show why target identity is necessary for sharpness.

- **Online certificate.** Exact joint solvers for NMAE and clipped-nMSE support target-indexed updates and feasible action projection. A sequential invariant gives cumulative-excess and fallback guarantees under the stated settlement assumptions.

- **Empirical validation.** Controlled overlap and partition experiments isolate shared-target dependence. Across backbones, datasets, and candidate generators, the tighter bound permits larger issued actions while preserving every reported certificate path.

2 Related Work

Forecasting and test-time adaptation. Long-horizon forecasting has progressed from decomposition-based attention to PatchTST, TimesNet, iTransformer, and FreTS (Wu et al., 2021; Nie et al., 2023; Wu et al., 2023; Liu et al., 2024; Yi et al., 2023), and more recently to pretrained models such as TimesFM, Chronos, Moirai, and MOMENT (Das et al., 2024; Ansari et al., 2024; Woo et al., 2024; Goswami et al., 2024). TAFAS, OneNet, DSOF, ELF, and COSA use deployment-time observations to update calibrators, lightweight models, or output-space corrections (Kim et al., 2025; Zhang et al., 2023; Lau et al., 2025; Lee et al., 2025; Im & Kwon, 2026). These methods construct or improve a candidate forecast. We study a subsequent decision object: the deterministic remaining excess of candidate corrections that have already been issued but whose labels have not arrived. Forecast reconciliation adjusts forecast values to satisfy aggregation or coherence constraints (Wickramasuriya et al., 2019), while stability-aware multi-horizon methods compare predictions from different origins (Ma et al., 2026). TC-FSRE instead leaves forecast values unchanged and imposes coherence only on latent label completions: all excess terms for one absolute target share its single realized label.

Neighboring guarantees. Conservative bandits and safe online optimization constrain reward relative to a baseline or enforce per-round action feasibility (Wu et al., 2016; Kazerouni et al., 2017; Hutchinson & Alizadeh, 2024). Delayed online learning studies regret as feedback is postponed (Joulani et al., 2013; Quanrud & Khashabi, 2015; Ryabchenko et al., 2025). Conformal risk control and online conformal prediction target expected risk, coverage, or quantiles under exchangeable or sequential data (Angelopoulos et al., 2024b; Bhatnagar et al., 2023; Angelopoulos et al., 2024a; Areces et al., 2025). Recent monitors use sequential tests or confidence sequences to detect risk violations for static or continuously adapted models (Timans et al., 2025; Schirmer et al., 2025). They control statistical false alarms; our certificate is deterministic and indexed by the shared targets of issued point forecasts. We define this certificate for an already-issued pending set; item-separable evaluation is the natural relaxation that omits the target partition. Appendix Table 2 separates the theorem objects. Robust optimization supplies the general principle that a worst-case value depends on its uncertainty set (Bertsimas et al., 2011; Wiesemann et al., 2014); here, absolute target identity determines the legal completion set, and we characterize the value lost when that partition is omitted.

3 Problem Setup: Forecast Adaptation Under a Cumulative Excess Constraint

3.1 Frozen base forecast, candidate interface, and action

At issue time t, a frozen forecaster produces an H-step, C-channel base forecast $b_t \in \mathbb{R}^{H\times C}$, and an adapter exposing the forecast-space interface produces c_t. We write the candidate direction as $d_t = c_t - b_t$. An empirical policy uses the information available at time t to propose $\widetilde{a}_t \in [0,1]$, and the risk layer approves

$$a_t \in [0, \widetilde{a}_t], \qquad p_t = b_t + a_t d_t. \tag{1}$$

The risk layer does not inspect the candidate's parameters, gradients, hidden states, or training procedure. Setting $a_t = 0$ recovers the frozen base and is always a feasible fallback.

3.2 Multi-horizon feedback protocol

The label for lead $k \in \{0, \dots, H-1\}$ becomes observable no earlier than logical time $t+k+1+\delta$, where $\delta \geq 0$ is an additional feedback delay. Every quantity that affects $b_t, c_t, \widetilde{a}_t$, the scale $s_t > 0$, or a_t must be measurable when the forecast is issued. Future labels cannot enter the decision. Each lead retains the normalization denominator of the full forecast, so the H termwise losses sum exactly to the mean loss of the trajectory.

Algorithm 1: Event-driven TC-FSRE update

State: realized excess E, pending terms P, budget B.
1. Settle arriving target groups: add realized group excess to E and remove them from P.
2. Recompute $\Lambda_{\mathrm{TC}}(P) = \sum_u \sup_y \sum_{i \in G_u} \Delta_i(y)$ and $\mathcal{R}_{\mathrm{TC}}(P) = [\Lambda_{\mathrm{TC}}(P)]_+$.
3. Receive $(b_t, d_t, \widetilde{a}_t)$; choose the largest $a_t \le \widetilde{a}_t$ with $E + \mathcal{R}_{\mathrm{TC}}(P \cup \mathrm{new}(a_t)) \le B$.
4. Issue $p_t = b_t + a_t d_t$ and index its terms by target; fallback sets later $a_t = 0$ while settlement continues.

3.3 Loss and prefix constraint

Fix a controlled loss ℓ. For an observed forecast term (t, k), define its realized excess over the frozen forecast as

$$\Delta_{t,k} = \ell_{t,k}(p_t, y) - \ell_{t,k}(b_t, y). \tag{2}$$

Let $\mathcal{M}_q$ and $\mathcal{P}_q$ denote the observed and pending terms at event prefix q, and let $B \ge 0$ be the cumulative excess threshold. To evaluate the worst-case effect of pending corrections, imagine disabling further adaptation and completing every currently missing label. The largest resulting excess is the *signed sharp completion value* $\Lambda(P)$ of the fixed pending set P. The forecaster itself continues with $a_t = 0$. The online constraint uses its positive part, the *nonnegative risk amount* $\mathcal{R}(P) = [\Lambda(P)]_+$. We maintain

$$\sum_{i \in \mathcal{M}_q} \Delta_i + \mathcal{R}(\mathcal{P}_q) \le B \tag{3}$$

at every issue or feedback event. We call this condition *prefix stop-and-complete safety*: at each prefix, the realized excess plus the worst legal completion of pending corrections does not exceed B. Individual Δ_i may still be positive.

Proposition 1 (Meaning of the prefix constraint)**.** *If Eq. (3) holds at event prefix q, then disabling subsequent nonzero adaptation and completing the current fixed pending set under any legal label assignment yields final cumulative excess at most B.*

Argument. Once nonzero adaptation is disabled, future base forecasts add zero excess. The final excess of the current pending set is at most $\Lambda(\mathcal{P}_q) \le \mathcal{R}(\mathcal{P}_q)$, so the total cannot exceed B. Thus each event prefix certifies an operational fallback in which the base forecaster continues while existing adapted forecasts mature.

4 Target-Coherent Forecast-Space Risk Envelopes

Algorithm 1 gives the event-level interface before the fixed-set values and solvers are derived below. It separates settlement of irrevocable past actions from projection of a newly proposed action.

4.1 Exact worst-case excess at issue time

Consider a single forecast correction. For normalized absolute loss and any displacement h,

$$\sup_x \{|x + h| - |x|\} = |h|. \tag{4}$$

The exact worst-case NMAE excess of action a is therefore

$$L_{\mathrm{NMAE}}(a, d, s) = a \ \mathrm{mean}\, |d/s|. \tag{5}$$

Exactness means that the bound is attainable; it requires no label distribution, bounded support, or moment condition.

Our second loss uses $\phi_\tau(x) = \min(x^2, \tau^2)$. Its single-term worst-case excess also has an attainable closed form, and the full clipped-nMSE excess is the coordinate average of that form evaluated at $a|d/s|$. The piecewise expression and its derivation are given in Appendix A.8. The guarantee applies only to the normalized clipped loss at the specified τ.

4.2 An item-separable bound

The simplest valid evaluation ignores shared targets and gives every pending term its own worst-case label.

Suppose each pending term i is assigned $\lambda_i = \sup_y \Delta_i(y)$. For the two losses considered here, $\lambda_i \geq 0$, and hence

$$\Lambda_{\mathrm{ind}}(P) = \sum_{i \in P} \lambda_i, \qquad \mathcal{R}_{\mathrm{ind}}(P) = [\Lambda_{\mathrm{ind}}(P)]_+ = \Lambda_{\mathrm{ind}}(P), \qquad S = B - E - \mathcal{R}_{\mathrm{ind}}(P), \quad (6)$$

where E is the realized excess and S is the feasibility margin. A forecast is issued only if S remains nonnegative after its terms are added. When a label arrives, its term is removed from the pending set, $\mathcal{R}_{\mathrm{ind}}$ is recomputed, and its realized $\Delta_i \leq \lambda_i$ is added to E. The margin cannot decrease at this update. This construction gives a simple prefix invariant, but it implicitly assigns an independent adversarial label to every prediction term.

4.3 Shared targets break itemwise additivity

Rolling forecasts often contain several terms for the same target u. Consider two NMAE excess functions that share one scalar label y:

$$\Delta_1(y) = |1 - y| - |0 - y|, \qquad \Delta_2(y) = -\Delta_1(y). \quad (7)$$

Each term has worst-case excess 1, so $\Lambda_{\mathrm{ind}} = 2$. Their sum, however, is zero for every shared label. Exact itemwise bounds therefore need not sum to an exact bound under the true protocol. The converse matters as well. If two genuinely independent targets are merged, this example incorrectly gives a zero bound, whereas a legal product completion can attain total excess 2.

Target coherence is label coherence in the completion problem, not reconciliation of forecast values. Forecasts issued from different origins for the same absolute target may remain numerically different; TC-FSRE only requires their excess terms to share the single realized label in a legal completion.

4.4 Target-coherent completion value and nonnegative risk amount

We now impose the constraint missing from itemwise evaluation: unresolved terms for the same absolute target must share one label. Fix an online history and a fixed issued pending set P. Let U be the set of distinct target blocks, let $u(i)$ denote the block used by term i, and let $\mathcal{Y}_u$ be its feasible label set. Thus the *target partition* groups terms with the same $u(i)$. For $G_u = \{i \in P : u(i) = u\}$, define the group excess $F_u(y) = \sum_{i \in G_u} \Delta_i(y)$ and

$$\Lambda_{\mathrm{TC}}(P) = \sum_{u \in U} \sup_{y \in \mathcal{Y}_u} F_u(y). \quad (8)$$

The corresponding online quantity is $\mathcal{R}_{\mathrm{TC}}(P) = [\Lambda_{\mathrm{TC}}(P)]_+$. If known hard constraints couple different times or channels, the target block must be enlarged accordingly; those coordinates cannot be factorized. Our implementation uses scalar blocks indexed by absolute time and channel.

Theorem 1 (Target-coherent sharp completion value)**.** *Let P be finite and irreversible. Suppose the legal completion set factorizes over distinct target blocks as $\prod_{u \in U} \mathcal{Y}_u$, and every group supremum is finite. Under a prefix protocol that permits issuance to stop and completes every unobserved label in P,*

$$\sup_{(y_u) \in \prod_u \mathcal{Y}_u} \sum_{i \in P} \Delta_i(y_{u(i)}) = \Lambda_{\mathrm{TC}}(P). \quad (9)$$

Thus $\Lambda_{\mathrm{TC}}(P)$ is the signed sharp completion value for this fixed pending set, and $\mathcal{R}_{\mathrm{TC}}(P)$ is the corresponding smallest nonnegative risk amount.

Proof sketch. Rearrange total excess as $\sum_u F_u(y_u)$. Every completion is bounded by the sum of the group suprema. In the reverse direction, choose a label arbitrarily close to the supremum in each finite group. The product completion set allows all such choices simultaneously, so the upper bound can be approached arbitrarily closely. Any value strictly below $\Lambda_{\mathrm{TC}}(P)$ is violated by the corresponding witness, while Eq. (8) bounds every legal completion. Appendix A gives the full proof.

Corollary 1 (Smallest nonnegative amount for deterministic pathwise validity)**.** *Fix an issued pending set P and the stop-and-complete protocol of Theorem 1. Any method that uses a nonnegative amount $K(P)$ to cover every legal completion of this fixed set and thereby asserts deterministic pathwise validity must satisfy*

$$K(P) \geq \mathcal{R}_{\mathrm{TC}}(P). \tag{10}$$

Every smaller nonnegative value is insufficient, whereas $K(P) = \mathcal{R}_{\mathrm{TC}}(P)$ covers all legal completions. Hence $\mathcal{R}_{\mathrm{TC}}(P)$ is the smallest nonnegative risk amount for this fixed set under the stated protocol.

Argument. If $K < \Lambda_{\mathrm{TC}}$, a near-maximizing completion from Theorem 1 violates the proposed bound. Nonnegativity also requires $K \geq 0$. Sufficiency follows from $\Lambda_{\mathrm{TC}} \leq \mathcal{R}_{\mathrm{TC}}$.

In short, grouping by target removes impossible combinations of labels without discarding any completion allowed by the model.

4.5 Exact NMAE coherent–itemwise gap

The ordering alone does not explain when target grouping helps. For NMAE, the gap is governed by whether the terms in a target group can attain their separate worst cases at one common label. Consider one scalar target group G with $q_i > 0$, base prediction b_i, and issued displacement h_i. Let

$$\Lambda_{\mathrm{ind},G} = \sum_{i \in G} q_i |h_i|, \qquad \Lambda_{\mathrm{TC},G} = \sup_{y \in \mathbb{R}} \sum_{i \in G} q_i \big(|b_i + h_i - y| - |b_i - y| \big), \tag{11}$$

and define the conservatism gap $\Gamma_G = \Lambda_{\mathrm{ind},G} - \Lambda_{\mathrm{TC},G}$. Write $[x]_0^r = \min\{\max\{x, 0\}, r\}$ for $r \geq 0$, and let $I_+ = \{i : h_i > 0\}$ and $I_- = \{i : h_i < 0\}$.

Theorem 2 (Exact directional-conflict decomposition)**.** *For* NMAE *with scalar label space* $\mathbb{R}$*,*

$$\Gamma_G = 2 \min_{y \in \mathbb{R}} C_G(y), \qquad C_G(y) = \sum_{i \in I_+} q_i [y - b_i]_0^{h_i} + \sum_{i \in I_-} q_i [b_i - y]_0^{-h_i}. \tag{12}$$

Thus itemwise conservatism is exactly twice the smallest aggregate penalty incurred when one shared label must replace the mutually independent labels used by the itemwise worst cases.

Corollary 2 (Zero-gap condition and strict conflict bound)**.** *Under the conditions of Theorem 2,* $\Gamma_G = 0$ *if and only if*

$$\max_{i \in I_-} b_i \leq \min_{i \in I_+} b_i, \tag{13}$$

where the condition holds automatically if either index set is empty. If it fails, choose $p \in \arg\min_{i \in I_+} b_i$ and $n \in \arg\max_{i \in I_-} b_i$, and set $D_G = b_n - b_p > 0$. Then

$$\Gamma_G \geq 2 \min\{q_p, q_n\} \min\{D_G, h_p, -h_n\} > 0. \tag{14}$$

Consequently, mixed displacement signs are necessary but not sufficient for a positive gap: strict improvement occurs exactly when the independent worst-case label regions have empty intersection. Appendix A gives the proof.

Thus overlap creates an opportunity for a tighter bound, but directional conflict determines whether that opportunity produces a strict gain.

4.6 Partition refinement and dependency-blind validity–sharpness incompatibility

Let π be a target partition of a fixed term set P. Terms in one block $G \in \pi$ share a label, while distinct blocks complete independently within the stated completion class. Define

$$\Lambda_\pi(P) = \sum_{G \in \pi} \sup_{y_G \in \mathcal{Y}_G} \sum_{i \in G} \Delta_i(y_G). \tag{15}$$

Proposition 2 (Monotonicity under target-partition refinement)**.** *Let π' refine π by splitting one or more shared-target blocks into independently completed sub-blocks with copies of the same feasible set. Then*

$$\Lambda_\pi(P) \leq \Lambda_{\pi'}(P) \leq \Lambda_{\mathrm{ind}}(P). \tag{16}$$

Both inequalities can be strict. Equality on the right holds for the singleton partition.

Proof sketch. For each split block, apply $\sup_y \sum_j F_j(y) \le \sum_j \sup_{y_j} F_j(y_j)$ and then sum over blocks. The two-term construction above has value 0 under the shared partition and 2 under the singleton partition, giving a strict gap. Appendix A contains the proof.

Call an upper-bound rule Ψ *dependency blind* if it observes the marginal excess functions, feasible label sets, and delay metadata, but not the target partition π. It must therefore return the same value for two protocols with identical marginal terms and feedback times but different target identities. We use Ψ to distinguish such a rule from the online nonnegative quantity $\mathcal{R}$.

Theorem 3 (Dependency-blind validity–sharpness incompatibility)**.** *No dependency-blind rule Ψ can simultaneously satisfy both conditions:* (i) *pathwise validity for every fixed issued pending set under product completion when nonzero adaptation is disabled and current labels are completed; and* (ii) *signed sharpness for every marginally identical fixed set under diagonal, shared-target completion.*

Proof. Use the two marginal functions above, with identical feasible sets and feedback times. Under product completion, the two labels can vary independently and the signed sharp value is 2, so pathwise validity requires an output of at least 2. Under diagonal completion, the terms share one label and sum to zero everywhere, so the sharp value is 0. A dependency-blind rule returns the same value in both cases. A value below 2 is invalid for the product protocol, while a value at least 2 is not sharp for the diagonal protocol.

Theorem 3 is a fixed-set information result, not a global adaptive-policy optimum: with identical forecasts, marginal excess functions, feasible label sets, and delay metadata, changing only the target partition changes the legal label-completion set. Hence target identity is protocol information needed by a dependency-blind rule to distinguish product-validity from shared-target sharpness, not an aggregation heuristic.

4.7 Exact joint solvers and safe action projection

The certificate is practical only if each target-group supremum and the largest feasible action can be computed exactly or conservatively. For a scalar target group, let term i have base prediction b_i, issued displacement h_i, and positive weight q_i. Then

$$F(y) = \sum_i q_i\big(|b_i + h_i - y| - |b_i - y|\big) \tag{17}$$

is continuous and piecewise affine. Its breakpoints are in $\{b_i, b_i + h_i\}$ and both tails are constant. Evaluating the breakpoints and tails gives the exact group supremum; sorting slope events reduces the group complexity to $O(m \log m)$.

For clipped-nMSE, the group objective is

$$F_\tau(y) = \sum_i q_i \left[\phi_\tau\left(\frac{b_i + h_i - y}{s_i}\right) - \phi_\tau\left(\frac{b_i - y}{s_i}\right)\right]. \tag{18}$$

This objective is piecewise quadratic. Sorting and scanning its $4m$ coefficient-change events yields the exact group supremum in $O(m \log m)$ time. Appendix A.8 gives the piecewise form, candidate extrema, and solver validation.

For a new scalar action $a \in [0, \widetilde{a}_t]$, the NMAE excess is convex in a under every legal label completion. Finite sums, suprema, and sums over target groups preserve convexity. It follows that

$$\mathcal{A}_t = \{a \in [0, \widetilde{a}_t] : \mathcal{R}_{\mathrm{TC}}(P_t \cup \mathrm{new}(a)) \le B - E_t\} \tag{19}$$

is a closed interval containing zero. The ideal rule selects its largest element; the implementation uses 12-step conservative bisection and returns a feasible approximation from below.

4.8 Issuance and asynchronous updates

At each event, the system first processes newly observed targets. It removes the corresponding target groups from P_t, adds their realized group excesses to E_t, and recomputes $\Lambda_{\mathrm{TC}}(P_t)$ and $\mathcal{R}_{\mathrm{TC}}(P_t) = [\Lambda_{\mathrm{TC}}(P_t)]_+$. It then receives $b_t, d_t, \widetilde{a}_t$, finds the largest action satisfying $\mathcal{R}_{\mathrm{TC}}(P_t \cup \mathrm{new}(a)) \le$

$B - E_t$, and registers the new terms by absolute time and channel. Negative realized excess enters E_t only after its label is observed; the online constraint always uses nonnegative $\mathcal{R}_{\mathrm{TC}}$. Figure 1 summarizes the interface.

When target u arrives, $F_u(y_u) \le \sup_y F_u(y)$. For both losses used here with $\mathcal{Y}_u = \mathbb{R}$, every group supremum is nonnegative: the two NMAE tail values have opposite signs, and both clipped-nMSE tails are zero. Removing a resolved group therefore preserves Eq. (3). It also gives $\Lambda_{\mathrm{TC}}(P_t) \ge 0$ and $\mathcal{R}_{\mathrm{TC}}(P_t) = \Lambda_{\mathrm{TC}}(P_t)$ in the implementation. The code still applies the positive part to remove small negative round-off errors. For another loss with a potentially negative group supremum, $\mathcal{R}_{\mathrm{TC}} = [\Lambda_{\mathrm{TC}}]_+$ remains the smallest nonnegative amount for a fixed set under stop-and-complete, but asynchronous groupwise validity requires an additional condition. Leadwise updates can enlarge later feasible sets sooner, but need not give originwise dominance between two adaptive paths.

Theorem 4 (Sequential prefix validity)**.** *Let (E_q, P_q) be the realized excess and pending set after event q. Assume that (i) $E_0 + \mathcal{R}_{\mathrm{TC}}(P_0) \le B$; (ii) every issue event selects an action a satisfying*

$$E_q + \mathcal{R}_{\mathrm{TC}}(P_q \cup \mathrm{new}(a)) \le B; \tag{20}$$

and (iii) every feedback event settles an entire target group, while every group supremum in every reachable pending state is nonnegative. Then, for any finite ordering of issue and feedback events,

$$\boxed{E_q + \mathcal{R}_{\mathrm{TC}}(P_q) \le B \quad \textit{for every event prefix } q.} \tag{21}$$

Corollary 3 (Anytime fallback and finite-stream completion)**.** *At any event prefix, subsequent actions may be set to zero while the frozen forecaster continues operating; every legal completion of all previously issued adapted forecasts then has cumulative excess at most B. If a finite deployment stream ends and every pending target group is eventually settled, its final realized excess satisfies $E_{\mathrm{final}} \le B$.*

Theorem 4 is an event-level induction: issuance preserves feasibility by construction, and settlement replaces a nonnegative group supremum by a realized value no larger than that supremum. The proof is in Appendix A.7. We call Corollary 3 an *anytime fallback certificate*: for example, a drift alarm or operator can disable further adaptation without stopping the base forecasting service, while pending corrections continue to mature within the feasible margin.

5 Experiments

5.1 Setup and testable questions

Experiments test whether target coherence tightens identical-action bounds, permits more adaptation, and whether partition interventions isolate shared-target dependence.

The main panel contains six FreTS-H192 datasets (three checkpoints each) and five PatchTST-H720 datasets, using every rolling origin in the test part of a 70/10/20 split. We set $s_t = \max\{\sqrt{\mathrm{mean}(x_t^2)}, 10^{-6}\}$ and weight all $H \times C$ coordinates equally. The task-invariant rate $B/H = 0.0025$ was fixed before aggregate evaluation, not tuned per task for accuracy. Causal FTL uses only origins matured at least H steps earlier; projection uses 12 bisections. Async-FSRE (the itemwise baseline) and TC-FSRE differ only in completion geometry. Appendix B gives full protocols.

The statistical unit is a panel–dataset pair: FreTS checkpoints are averaged within dataset, yielding 11 clusters with PatchTST; origins and recorded terms are not independent samples. The main candidate is deterministic Hankel. A retrospective matched-interface panel uses three seeds of TAFAS, DSOF, and COSA on ETTh2, ETTm1, Exchange, and Weather with $H = 720$; it is summarized separately and is not added to the 11-cluster inference.

For trajectory j, let P_{jt}^{TC} be the post-issue pending set generated by the TC-FSRE action path. We define the same-action bound reduction as

$$\mathrm{Red}_j = \frac{\sum_t \left[\mathcal{R}_{\mathrm{ind}}(P_{jt}^{\mathrm{TC}}) - \mathcal{R}_{\mathrm{TC}}(P_{jt}^{\mathrm{TC}})\right]}{\sum_t \mathcal{R}_{\mathrm{ind}}(P_{jt}^{\mathrm{TC}})}, \tag{22}$$

Both bounds evaluate identical issued forecasts. We use the trajectory-sum ratio Red_j and $\Delta a_j = \mathrm{mean}_t(a_{jt}^{\mathrm{TC}} - a_{jt}^{\mathrm{ind}})$; MSE/MAE are percentage gains over the base, with Δ in percentage points. Appendix B gives the aggregation details.

Table 1: **Main results and contemporary interfaces.** Block A averages checkpoints within FreTS datasets and then 11 dataset–backbone units equally; Block B averages three seeds within task and four tasks equally. ΔMSE gain is issued $\text{MSEGain}_{\text{TC}} - \text{MSEGain}_{\text{ind}}$ in pp. Reduction is same-action; Δa and ΔMSE compare separately evolving action paths.

Setting	Interface	Year	Units	ΔMSE gain (pp) ↑	Reduction ↑	Δa ↑	Valid
A. Main completion-geometry panel							
FreTS-H192	TC vs. itemwise	–	6	–	71.57%	+0.1365	18/18
PatchTST-H720	TC vs. itemwise	–	5	–	84.01%	+0.2363	5/5
Overall	**TC vs. itemwise**	–	**11**	–	**77.23%**	**+0.18184**	**23/23**
B. Contemporary candidate interfaces under a matched protocol							
PatchTST-H720	TAFAS	2025	4	+0.0214	4.24%	+0.01421	24/24
PatchTST-H720	DSOF†	2025	4	+0.1247	21.97%	+0.00825	24/24
PatchTST-H720	COSA	2026	4	+0.0038	8.18%	+0.00736	24/24

5.2 Sharper completion permits more adaptation

Across 11 units, target-coherent completion reduced the same-action bound by 77.23% and increased mean issued action by 0.18184 under the same threshold.

5.3 Controlled interventions isolate target dependence

Without overlap, target-coherent and itemwise evaluation agree. As overlap increases, legal target merging reduces the bound, whereas splitting true groups into a safe refinement is more conservative (Appendix Figure 5). With forecasts, labels, proposals, and thresholds fixed, this isolates the target partition rather than rescaling. Theorem 3 likewise shows that a target-blind rule cannot be both product-valid and sharp for shared targets.

The gap also follows the predicted directional-conflict mechanism (Figure 3). Across 23 paths and 1,829,085 exposed scalar groups, one-direction and mixed-compatible groups had zero gap, while mixed-conflicting reduction rose from 47.3% to 80.0% with overlap; intervals resample the 11 units. Appendix Table 10 summarizes the additional candidate, loss, and backbone checks.

The bound reduction persisted for TAFAS, DSOF, and COSA, with all 72 certificate paths valid. To test whether this extra freedom can improve forecasting utility, we hold the PatchTST base, saved TAFAS candidates, proposal, and threshold fixed and change only completion geometry. TC-FSRE increased mean action by 0.01421 and the equal-task mean MSE gain by 0.0214 percentage points; all 24 paths remained valid. The effect was dataset-dependent, with most of the aggregate improvement coming from ETTm1 (Appendix C.1).

6 Discussion and Limitations

Delay determines when feedback arrives; target identity determines which worst-case labels may coexist. Forecast-error gains depend on the candidate and sequential state, as the TAFAS task variation shows. The guarantee assumes irreversible issuance and target-factorized completion; cross-target constraints require larger blocks. Maximal projection is proved for scalar NMAE, while clipped-nMSE has an exact fixed-set solver (Appendix D).

7 Conclusion

Repeated forecasts of one target cannot be completed with different labels. Encoding this fact yields a sharp pending-risk certificate, explains itemwise conservatism for NMAE, and permits substantially more adaptation in our experiments while preserving the stated cumulative-excess guarantee.

AI USE STATEMENT

Generative AI tools were used to provide feedback on research framing and experiment design; assist with implementation and testing; support literature discovery and metadata checking; and edit manuscript text and LaTeX. All reported numerical results were produced by executable code rather than generated by these tools. The authors reviewed the AI-assisted text, code, and analyses and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The appendices provide complete theorem assumptions and proofs, event ordering, loss definitions, dataset splits, model and adapter settings, feedback protocols, aggregation rules, computational costs, source hashes, and negative-result boundaries. The accompanying code reconstructs the evaluation streams, runs both completion rules on identical issued actions, reproduces all tables and figures, and tests the core invariants.

REFERENCES

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1616–1630, 2024a. URL https://proceedings.mlr.press/v235/angelopoulos24a.html.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=33XGfHLtZg.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Felipe Areces, Christopher Mohri, Tatsunori Hashimoto, and John Duchi. Online conformal prediction via online optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 1604–1649, 2025. URL https://proceedings.mlr.press/v267/areces25a.html.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. doi: 10.1137/080734510.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2337–2363, 2023. URL https://proceedings.mlr.press/v202/bhatnagar23a.html.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Spencer Hutchinson and Mahnoosh Alizadeh. Safe online convex optimization with multi-point feedback. In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 168–180, 2024. URL https://proceedings.mlr.press/v242/hutchinson24a.html.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L7Z5wBMPrW.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1453–1461, 2013. URL https://proceedings.mlr.press/v28/joulani13.html.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/bdc4626aa1d1df8e14d80d345b2a442d-Abstract.html.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):17868–17876, 2025. doi: 10.1609/aaai.v39i17.33965. URL https://ojs.aaai.org/index.php/AAAI/article/view/33965.

Ying-Yee Ava Lau, Zhiwen Shao, and Dit-Yan Yeung. Fast and slow streams for online time series forecasting without information leakage. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I0n3EyogMi.

Thomas L. Lee, William Toner, Rajkarn Singh, Artjom Joosen, and Martin Asenov. Lightweight online adaption for time series foundation model forecasts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33736–33764, 2025. URL https://proceedings.mlr.press/v267/lee25ag.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Chutian Ma, Grigorii Pomazkin, Giacinto Paolo Saggese, and Paul Smith. Beyond accuracy: A stability-aware metric for multi-horizon forecasting. *arXiv preprint arXiv:2601.10863*, 2026. URL https://arxiv.org/abs/2601.10863.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1270–1278, 2015.

Alexander Ryabchenko, Idan Attias, and Daniel M. Roy. Capacity-constrained online learning with delays: Scheduling frameworks and regret trade-offs. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4973–5014, 2025. URL https://proceedings.mlr.press/v291/ryabchenko25a.html.

Mona Schirmer, Metod Jazbec, Christian Andersson Naesseth, and Eric Nalisnick. Monitoring risks in test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2705. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/746960ad49ddb47248970a0e1404230c-Abstract-Conference.html.

Alexander Timans, Rajeev Verma, Eric Nalisnick, and Christian A. Naesseth. On continuous monitoring of risk violations under unknown shift. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4204–4226, 2025. URL https://proceedings.mlr.press/v286/timans25a.html.

Shanika L. Wickramasuriya, George Athanasopoulos, and Rob J. Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019. doi: 10.1080/01621459.2018.1448825.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014. doi: 10.1287/opre.2014.1314.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1254–1262, 2016. URL https://proceedings.mlr.press/v48/wu16.html.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain MLPs are more effective learners in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76656–76679, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f1d16af76939f476b5f040fd1398c0a3-Abstract-Conference.html.

Yifan Zhang, Qingsong Wen, Xue Wang, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69949–69980, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dd6a47bc0aad6f34aa5e77706d90cdc4-Abstract-Conference.html.

Table 2: Theorem-level comparison with neighboring paradigms. "Deterministic" means pathwise over every legal completion, rather than in expectation or with high probability.

Paradigm	Controlled object	Assumptions and guarantee	Feedback semantics	Shared-target dependence
Conservative/safe learning	Cumulative expected reward or per-round feasibility	High-probability safety and regret under parameter confidence sets	Reward or constraint feedback follows each action	Does not model legal completions of fixed overlapping point forecasts
Delayed online learning	Regret and expert weights	Asymptotic bounds under adversarial or stochastic delays	Round-level losses may be delayed	Encodes losses and delays, but not a target partition
Conformal risk/coverage	Prediction-set risk, coverage, or quantiles	Exchangeable expected risk or online coverage bounds	Calibration-then-test or label-triggered updates	Does not control prefix excess of point forecasts relative to a frozen base
Forecast reconciliation	Forecast-value vector	Projection or adjustment under aggregation or coherence constraints	Forecast values are reconciled at issue or aggregation time	Constrains forecast values; TC-FSRE shared-target dependence instead lies in the latent-label completion set of a cumulative-excess certificate
TC-FSRE (ours)	Remaining cumulative excess of a fixed issued pending set	Signed sharp value under stop-and-complete; its positive part is the smallest nonnegative risk amount	Multi-horizon labels arrive asynchronously by absolute target	Uses the target partition; a dependency-blind rule cannot be both valid and sharp

A PROOFS

A.1 EXACT VALUE UNDER ITEMWISE INDEPENDENT COMPLETION

Consider a finite fixed issued pending set P after nonzero adaptation is disabled. Suppose every term i has an independent label $y_i \in \mathcal{Y}_i$, and let $\lambda_i = \sup_{y_i} \Delta_i(y_i)$. Every completion satisfies $\sum_i \Delta_i(y_i) \le \sum_i \lambda_i$. Conversely, for any $\epsilon > 0$, choose y_i^ϵ such that $\Delta_i(y_i^\epsilon) > \lambda_i - \epsilon/|P|$. The product completion set permits all these choices simultaneously, giving total excess greater than $\sum_i \lambda_i - \epsilon$. Taking $\epsilon \downarrow 0$ yields

$$\sup_{(y_i)\in\prod_i \mathcal{Y}_i} \sum_i \Delta_i(y_i) = \sum_i \lambda_i. \tag{23}$$

Both sides are zero for an empty pending set. The item-separable bound is therefore sharp for the *enlarged itemwise product completion*, but not necessarily for a shared-target protocol.

A.2 PROOF OF THE TARGET-COHERENT VALUE

If $U = \varnothing$, Theorem 1 follows from the empty-sum convention. Otherwise, rearrange total excess as $\sum_{u\in U} F_u(y_u)$. The sum of group suprema gives the upper bound. For any $\epsilon > 0$, choose y_u^ϵ for each finite target group such that $F_u(y_u^\epsilon) > \sup_y F_u(y) - \epsilon/|U|$. The product set over distinct targets permits these near-maximizers simultaneously, producing total excess greater than $\Lambda_{\mathrm{TC}}(P) - \epsilon$. Equation (9) follows.

Let E be the current realized excess and $C = B - E$ the feasible margin. If $C < \Lambda_{\mathrm{TC}}(P)$, a sufficiently accurate near-maximizing completion makes the final cumulative excess exceed B; if $C \ge \Lambda_{\mathrm{TC}}(P)$, every legal completion is bounded. The exact signed stop-and-complete condition is therefore $C \ge \Lambda_{\mathrm{TC}}(P)$. Online decisions use the nonnegative counterpart of this bound. Within this class, the smallest value covering $\Lambda_{\mathrm{TC}}(P)$ is $[\Lambda_{\mathrm{TC}}(P)]_+ = \mathcal{R}_{\mathrm{TC}}(P)$, and $C \ge \mathcal{R}_{\mathrm{TC}}(P)$ is sufficient. This also proves Corollary 1. The positive-part convention does not use an overall negative completion value in an issue decision; negative realized excess enters E only after observation. For a randomized method, the same conclusion applies seed by seed only if pathwise validity is required after conditioning on its internal randomness.

A.3 Proof of the exact NMAE gap decomposition

For a single term, subtract its shared-label excess from its itemwise maximum. Direct evaluation of the three regions separated by b_i and $b_i + h_i$ gives

$$q_i|h_i| - q_i\big(|b_i + h_i - y| - |b_i - y|\big) = \begin{cases} 2q_i[y - b_i]_0^{h_i}, & h_i > 0, \\ 2q_i[b_i - y]_0^{-h_i}, & h_i < 0, \\ 0, & h_i = 0. \end{cases} \tag{24}$$

Summing Eq. (24) over G and using that the itemwise value is constant in y yields

$$\Gamma_G = \Lambda_{\mathrm{ind},G} - \sup_y F_G(y) = \inf_y\{\Lambda_{\mathrm{ind},G} - F_G(y)\} = 2\inf_y C_G(y). \tag{25}$$

The function C_G is continuous and piecewise affine with constant tails, so the infimum is attained. This proves Theorem 2.

Every summand of C_G is nonnegative. A positive-displacement term has zero penalty exactly when $y \le b_i$, whereas a negative-displacement term has zero penalty exactly when $y \ge b_i$. A common zero therefore exists if and only if $[\max_{i\in I_-} b_i, \infty) \cap (-\infty, \min_{i\in I_+} b_i]$ is nonempty, which is Eq. (13).

If the intersection is empty, take the extremal pair (p, n) from Corollary 2. For every y,

$$C_G(y) \ge q_p[y - b_p]_0^{h_p} + q_n[b_n - y]_0^{-h_n} \tag{26}$$

$$\ge \min\{q_p, q_n\}\left([y - b_p]_0^{h_p} + [b_n - y]_0^{-h_n}\right) \tag{27}$$

$$\ge \min\{q_p, q_n\}\min\{D_G, h_p, -h_n\}. \tag{28}$$

The last inequality follows by considering $y \le b_p$, $b_p \le y \le b_n$, and $y \ge b_n$; on the middle interval the two clipped distances sum to at least the minimum of the interval length and the two clipping radii. Minimizing over y and applying Theorem 2 proves Eq. (14) and the strictness claim.

A.4 Monotonicity under partition refinement

Suppose π' splits a block $G \in \pi$ into $G_1, \dots, G_r$, and define $F_j(y) = \sum_{i\in G_j} \Delta_i(y)$. Then

$$\sup_y \sum_{j=1}^{r} F_j(y) \le \sum_{j=1}^{r} \sup_{y_j} F_j(y_j), \tag{29}$$

because the right-hand side allows each sub-block to choose an independent and potentially inconsistent maximizing label. Applying this inequality to every split block gives $\Lambda_\pi(P) \le \Lambda_{\pi'}(P)$. Refining further to the singleton partition gives $\Lambda_{\pi'}(P) \le \Lambda_{\mathrm{ind}}(P)$. In the construction $\Delta_2 = -\Delta_1$, the shared partition has value zero and the singleton partition has value two, so the gap can be strict. This proves Proposition 2.

A.5 Quantifiers in the dependency-blind incompatibility

Write a dependency-blind rule as

$$\Psi(\{\Delta_i, \mathcal{Y}_i, m_i\}_{i\in P}), \tag{30}$$

where m_i contains the visible feedback-time and delay metadata for term i, but the input omits the target partition. Use two terms with identical feedback times and feasible sets. Product and diagonal completion present the same input above, so Ψ must return the same value. Under product completion, a legal label pair makes both terms equal to $+1$, and pathwise validity requires $\Psi \ge 2$. Under diagonal completion, $\Delta_1(y) + \Delta_2(y) = 0$ for every shared label, so the signed sharp completion value is zero. Thus $\Psi \ge 2$ is not sharp for the diagonal protocol, whereas $\Psi < 2$ is invalid for the product protocol. This proves Theorem 3. Its quantifiers cover fixed issued pending sets and label completion after nonzero adaptation is disabled. They do not rule out methods supplied with a richer dependency graph or rules involving future actions.

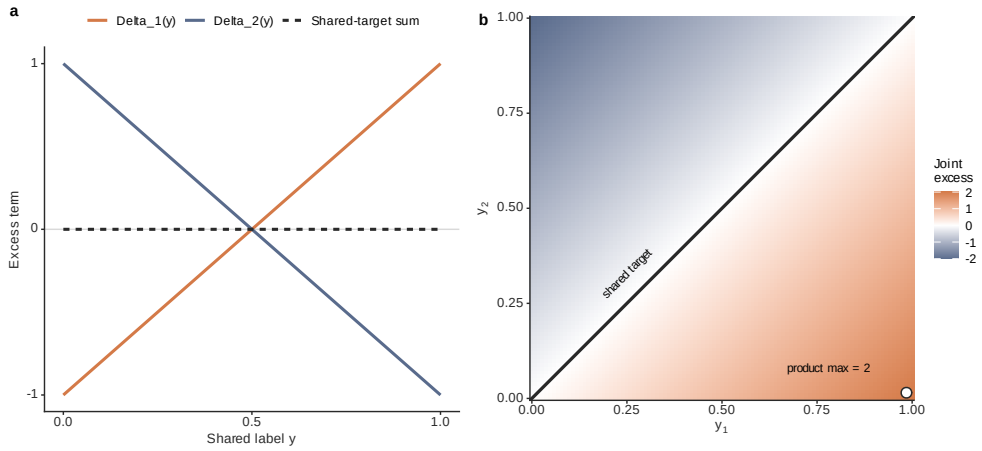


Figure 2: Analytic witness for dependency-blind validity–sharpness incompatibility. **a,** The two visible terms have opposite excess functions, so their sum is identically zero when they share one label. **b,** Under product completion, the labels may vary independently and the joint excess reaches two at the marked point; the black diagonal is the legal shared-target completion and has value zero throughout. The marginal terms and feedback metadata are unchanged, while the target partition changes the exact completion value.

A.6 Convexity of the NMAE action-feasible set

Fix the current pending set. For any legal label completion, the excess of a newly issued NMAE forecast is a sum of absolute-value affine functions of a and is therefore convex in a. Taking a supremum over labels, summing across target groups, and taking the maximum with zero preserve convexity, so the updated $\mathcal{R}_{\mathrm{TC}}(a)$ is convex. At $a = 0$, the new forecast equals the base and does not change excess; the pre-issue state is feasible. The sublevel set within $[0, \widetilde{a}_t]$ is therefore a closed interval containing zero. Conservative bisection returns an inner approximation to its largest feasible action.

A.7 Proof of sequential prefix validity

We prove Theorem 4 by induction over events. The initial state is feasible by assumption. At an issue event, Eq. (20) is exactly the claimed invariant for the post-issue pending set.

Now consider a feedback event that settles target group G_u. Write

$$\sigma_u = \sup_{y \in \mathcal{Y}_u} F_u(y) \geq 0. \tag{31}$$

Because every group supremum is nonnegative, the positive part is inactive and target-group factorization gives

$$\mathcal{R}_{\mathrm{TC}}(P_q) = \Lambda_{\mathrm{TC}}(P_q) = \sigma_u + \mathcal{R}_{\mathrm{TC}}(P_q \setminus G_u). \tag{32}$$

For the observed label y_u, the update is $E_{q+1} = E_q + F_u(y_u)$ and $P_{q+1} = P_q \setminus G_u$. Since $F_u(y_u) \leq \sigma_u$,

$$\begin{aligned} E_{q+1} + \mathcal{R}_{\mathrm{TC}}(P_{q+1}) &= E_q + F_u(y_u) + \mathcal{R}_{\mathrm{TC}}(P_q \setminus G_u) && (33)\\ &\leq E_q + \sigma_u + \mathcal{R}_{\mathrm{TC}}(P_q \setminus G_u) && (34)\\ &= E_q + \mathcal{R}_{\mathrm{TC}}(P_q) \leq B. && (35) \end{aligned}$$

Several simultaneously observed target groups may be settled in any order, so the argument covers batched and asynchronous feedback. Induction proves Eq. (21). If all later actions are zero, newly issued base forecasts have identically zero excess and do not change the invariant; Proposition 1 then yields the anytime-fallback statement. For a finite stream whose groups all settle, $P_{\mathrm{final}} = \varnothing$ and $\mathcal{R}_{\mathrm{TC}}(\varnothing) = 0$, hence $E_{\mathrm{final}} \leq B$.

A.8 CLIPPED-NMSE SINGLE-TERM ENVELOPE AND JOINT SOLVER

Let $\phi_\tau(x) = \min(x^2, \tau^2)$ and $u = |h|$. The worst-case excess of one displaced term is

$$\sup_x\{\phi_\tau(x+h) - \phi_\tau(x)\} = g_\tau(u), \qquad g_\tau(u) = \begin{cases} 2\tau u - u^2, & u \le \tau, \\ \tau^2, & u > \tau. \end{cases} \tag{36}$$

For $u \le \tau$, the maximum occurs where the shifted term reaches the clipped plateau while the base term remains in its quadratic region. For $u > \tau$, one can place the shifted term on the plateau while the base term is zero, attaining τ^2. The itemwise worst-case clipped-nMSE is therefore the coordinate average of $g_\tau(a|d/s|)$. This bound is attainable but applies only to the normalized clipped loss at the chosen τ.

For center c, scale $s > 0$, and weight $q \ge 0$, the function $q\phi_\tau((c-y)/s)$ is quadratic on $[c-\tau s, c+\tau s]$ and constant at $q\tau^2$ outside. Each term contributes an adapted center $p_i = b_i + h_i$ with sign $+1$ and a base center b_i with sign -1, giving four events:

$$p_i - \tau s_i, \quad p_i + \tau s_i, \quad b_i - \tau s_i, \quad b_i + \tau s_i.$$

Between consecutive events, the group excess has the form $Ay^2 + By + C$. Starting from $(A, B, C) = (0, 0, 0)$ on the left tail, the solver adds or removes the relevant quadratic coefficients while scanning the sorted events. The maximum on each interval occurs at an endpoint or, when $A < 0$ and $-B/(2A)$ lies in the interval, at that vertex. Both infinite tails equal zero. This finite candidate set contains a global maximizer and therefore returns the exact supremum. Sorting costs $O(m \log m)$ and the scan costs $O(m)$. Table 11 reports errors against an independent enumerator on 400 random and degenerate cases.

B PROTOCOL AND REPRODUCIBILITY DETAILS

Main Hankel candidate. The candidate transforms only the frozen forecast b_t and does not read future labels. For each channel and each window of length $W = \min\{96, H\}$, we subtract the window mean and construct an $r \times (W - r + 1)$ Hankel matrix, where $r = \lfloor W/2 \rfloor$. For singular values σ_j, define

$$\beta = \frac{r}{W - r + 1}, \qquad \theta = (0.56\beta^3 - 0.95\beta^2 + 1.82\beta + 1.43)\,\mathrm{median}_j\,\sigma_j,$$

and apply the hard threshold $\sigma'_j = \sigma_j \mathbf{1}\{\sigma_j \ge \theta\}$. After SVD reconstruction, anti-diagonal averaging recovers the sequence and the mean is restored. If $H > W$, windows overlap at stride $W/2$, the last window is aligned to the endpoint, and overlapping positions are averaged uniformly. The resulting forecast is c_t and has no trainable parameters. ELF uses a causal, protocol-matched local implementation. COSA uses its output-adapter architecture with batch size 48, 3 steps, context 10, PAAS disabled, and training delayed until the full H-step label is available. These methods are interface cases; we do not claim bitwise reproduction of the original implementations.

Frozen foundation model and learned adapter. The retrospective interface evaluation uses frozen `google/timesfm-2.5-200m-pytorch` forecasts on PEMS03/04/07/08, with $H = 96$, context 512, eight held-out sensors, origin stride one, 112 calibration origins, and 200 scored origins per dataset. The learned candidate is the pinned COSA `SimpleOutputAdapter` (upstream commit `43a8c8d`) with batch size 4, three gradient steps, context 10, and PAAS disabled. An origin enters training only after its complete 96-step target is observable. Three seeds are averaged within each dataset; neither origins nor seeds are independent statistical units. TimesFM weights remain frozen. This protocol-matched adapter is not an unmodified official COSA benchmark result, and the panel was designed after the main study. Accordingly, it is used only as a retrospective interface stress test under severe negative transfer, not as evidence that the adapter improves foundation-model forecasts.

Scale, proposal policy, and test stream. Data follow the 70/10/20 temporal split used by TAFAS, and the main text uses every rolling origin in the test stream. The scale is the context RMS, with

Table 3: Test-stream sizes for the main panel. FreTS has three checkpoints per dataset; PatchTST has one.

Backbone	Data	Origins	H	C	B	Backbone	Data	Origins	H	C	B
FreTS	ETTh1	3293	192	7	0.48	PatchTST	ETTh1	2765	720	7	1.80
FreTS	ETTh2	3293	192	7	0.48	PatchTST	ETTh2	2765	720	7	1.80
FreTS	ETTm1	13745	192	7	0.48	PatchTST	ETTm1	13217	720	7	1.80
FreTS	ETTm2	13745	192	7	0.48	PatchTST	ETTm2	13217	720	7	1.80
FreTS	Weather	10348	192	21	0.48	PatchTST	Weather	9820	720	21	1.80
FreTS	Exchange	1326	192	8	0.48						

coordinate weight $1/(HCs_t)$. At time t, causal FTL includes only complete forecasts from $t-H$ or earlier and returns

$$\widetilde{a}_t = \mathrm{clip}_{[0,1]}\left(\frac{\sum_j z_j}{\sum_j q_j}\right),$$

where q_j is the mean squared normalized candidate displacement and z_j is the normalized inner product between displacement and base residual. Main experiments fix $B = 0.0025H$, use 12 bisection steps, and process leads as they naturally arrive. The clipped-nMSE audit uses $\tau = 1$; the descriptive clipped metric in the main results uses $\tau = 3$. Table 3 gives the task dimensions.

Statistical protocol. All thresholds and loss parameters were fixed before the main results were aggregated. The three FreTS checkpoints are first aggregated within each dataset; the cluster bootstrap resamples the 11 panel–dataset units 100,000 times. Same-action sums use nonempty post-issue states, and zero denominators receive reduction zero. PatchTST is a retrospective cross-backbone confirmation, not a blind test maintained throughout the project. At every event, we log E_t, $\mathcal{R}_{\mathrm{TC}}(P_t)$, the remaining margin, and any violation. Unobserved labels may be completed offline at the end of the test stream, but cannot change issued actions. Code, configurations, source hashes, and event-level trajectories will accompany the submission in an anonymous repository.

Matched contemporary-interface protocol. The retrospective panel in Table 1 fixes the same released PatchTST-H720 base forecasts, dataset-native temporal splits (60/20/20 for ETTh2 and ETTm1; 70/10/20 for Exchange and Weather), complete-origin causal feedback, $B/H = 0.0025$, NMAE, 12 projection bisections, and four common tasks for TAFAS (Kim et al., 2025), DSOF (Lau et al., 2025), and COSA (Im & Kwon, 2026). Three fixed seeds are averaged within task before the four task means are averaged. Each generator supplies c_t; itemwise and target-coherent completion rules then process the same saved candidate tensor and proposal policy. TAFAS and COSA use the common released PatchTST both as b_t and as their frozen reference. DSOF cannot be reduced to that interface without changing its method: its c_t comes from the native DSOF PatchTST teacher plus MLP student, while the common released PatchTST still defines b_t. This structural exception is not an official like-for-like DSOF benchmark. All generators use only feedback legal under the complete-origin protocol. Candidate files, source revisions, and reproducibility metadata are provided with the artifact.

The 11/11 direction of same-action reduction corresponds to a two-sided exact sign value of 0.00098. Because these are selected benchmark units rather than a random population sample, we report this calculation only as a finite-panel sensitivity summary and do not use it to support population-level significance or any secondary outcome.

C Additional Results and Validation

C.1 Matched contemporary candidate interfaces

Exchange produces no nonzero TC adaptation for TAFAS or COSA under the fixed proposal policy. Across the full panel, all 72 certificate paths satisfy the prefix constraint.

Table 5 gives the task-level TAFAS results. At the primary threshold, TC-FSRE increased mean action by 0.01421 and the equal-task mean MSE gain by 0.0214 percentage points. Most of this improvement came from ETTm1, while all 24 paths satisfied the cumulative-excess constraint. Weather

Table 4: Dataset-level candidate, geometry, and action results. Each row averages three fixed seeds. Raw MSE gain is the unprojected full candidate relative to the common frozen PatchTST base. Reduction is a same-action geometry comparison, whereas the action columns come from independently evolving itemwise and target-coherent paths. Valid counts both completion rules; Exposed counts seed-level paths with a nonzero same-action denominator. The DSOF structural exception is defined in Appendix B.

Candidate	Dataset	Raw MSE gain	Reduction ↑	a_{ind}	a_{TC}	Δa	Valid	Exposed
TAFAS (2025)	ETTh2	+4.478%	0.887%	0.09312	0.09320	+0.00008	6/6	3/3
	ETTm1	+3.284%	5.996%	0.43194	0.49138	+0.05944	6/6	3/3
	Exchange	+0.067%	0.000%	0.00000	0.00000	+0.00000	6/6	0/3
	Weather	+5.417%	10.074%	0.06756	0.06488	−0.00268	6/6	3/3
DSOF (2025)†	ETTh2	−97.065%	28.736%	0.01544	0.02520	+0.00976	6/6	3/3
	ETTm1	−25.128%	25.475%	0.13984	0.15926	+0.01941	6/6	3/3
	Exchange	−208.537%	0.180%	0.00317	0.00318	+0.00001	6/6	3/3
	Weather	−37.342%	33.506%	0.01704	0.02084	+0.00381	6/6	3/3
COSA (2026)	ETTh2	−1.117%	4.067%	0.03644	0.03778	+0.00135	6/6	3/3
	ETTm1	−1.709%	11.517%	0.20752	0.23358	+0.02605	6/6	3/3
	Exchange	+0.001%	0.000%	0.00000	0.00000	+0.00000	6/6	0/3
	Weather	−0.918%	17.137%	0.06483	0.06689	+0.00206	6/6	3/3

Table 5: Dataset-level descriptive issued-path accuracy. MSE and MAE entries are percentage gains relative to the frozen base; each Δ is target-coherent minus itemwise in percentage points. Harm is the percentage of origins whose issued raw MSE exceeds the base. Each row averages three fixed seeds.

Candidate	Dataset	$\mathrm{MSE}_{\mathrm{ind}}$	$\mathrm{MSE}_{\mathrm{TC}}$	ΔMSE	$\mathrm{MAE}_{\mathrm{ind}}$	$\mathrm{MAE}_{\mathrm{TC}}$	ΔMAE	$\mathrm{Harm}_{\mathrm{ind}}$	$\mathrm{Harm}_{\mathrm{TC}}$
TAFAS (2025)	ETTh2	0.1149	0.1138	−0.0011	0.0892	0.0886	−0.0006	4.59%	4.59%
	ETTm1	1.4520	1.5978	+0.1458	0.8288	0.8303	+0.0015	26.95%	28.04%
	Exchange	0.0000	0.0000	+0.0000	0.0000	0.0000	+0.0000	0.00%	0.00%
	Weather	0.3886	0.3296	−0.0590	0.1055	0.0682	−0.0373	25.31%	25.14%
DSOF (2025)†	ETTh2	0.1739	0.2652	+0.0912	0.0528	0.0808	+0.0280	31.90%	32.31%
	ETTm1	3.4297	3.7739	+0.3442	2.0349	2.1922	+0.1573	19.76%	20.66%
	Exchange	0.4776	0.4788	+0.0012	0.3755	0.3765	+0.0011	0.00%	0.00%
	Weather	0.3887	0.4511	+0.0624	0.1508	0.1493	−0.0016	23.70%	23.83%
COSA (2026)	ETTh2	0.0730	0.0829	+0.0099	0.0880	0.0954	+0.0074	11.79%	11.79%
	ETTm1	0.6029	0.6479	+0.0450	0.2964	0.3038	+0.0074	29.44%	30.13%
	Exchange	0.0000	0.0000	+0.0000	0.0000	0.0000	+0.0000	0.00%	0.00%
	Weather	0.2564	0.2168	−0.0396	0.0842	0.0617	−0.0226	23.20%	23.33%

also illustrates that a tighter same-action bound need not produce a larger average action after the two methods evolve through different sequential states.

Table 6 reports the predefined threshold grid. The equal-task mean MSE difference remained positive at every nonzero threshold, although dataset-level effects remained mixed. The fixed $B/H = 0.0025$ result remains the primary comparison.

Table 7 reports the originwise comparison. Among origins where TC-FSRE issued a larger action, lower raw MSE occurred more often than not on each task with a nonempty conditioning set; these overlapping origins are used only as descriptive diagnostics.

C.2 Full cluster effects

Figure 4 expands the aggregate row of Table 9 into the prespecified statistical clusters and separates checkpoint variation from cluster-level uncertainty.

Table 6: Forecasting utility with TAFAS candidates across the predefined threshold grid. Each row averages three fixed seeds within dataset and then weights the four datasets equally. MSE and MAE are percentage gains over the frozen PatchTST base; differences are target-coherent minus itemwise in percentage points. Harm is the fraction of origins whose issued raw MSE exceeds the base.

B/H	Red.	a_{ind}	a_{TC}	Δa	$\mathrm{MSE}_{\mathrm{ind}}$	$\mathrm{MSE}_{\mathrm{TC}}$	ΔMSE	$\mathrm{MAE}_{\mathrm{ind}}$	$\mathrm{MAE}_{\mathrm{TC}}$	ΔMAE	$\mathrm{Harm}_{\mathrm{ind}}$	$\mathrm{Harm}_{\mathrm{TC}}$	Valid
0	0.000%	0.00000	0.00000	0.00000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.00%	0.00%	24/24
0.00025	4.438%	0.04156	0.05243	+0.01087	0.1778	0.2100	+0.0321	0.1325	0.1496	+0.0172	12.00%	12.21%	24/24
0.00050	4.349%	0.06733	0.08151	+0.01418	0.2621	0.2891	+0.0270	0.1791	0.1849	+0.0058	12.44%	12.78%	24/24
0.00100	4.326%	0.10207	0.11697	+0.01490	0.3576	0.3768	+0.0192	0.2176	0.2133	−0.0043	13.15%	13.46%	24/24
0.00250	4.239%	0.14816	0.16237	+0.01421	0.4889	0.5103	+0.0214	0.2559	0.2468	−0.0091	14.21%	14.44%	24/24
0.00500	4.180%	0.19714	0.20705	+0.00991	0.6900	0.7701	+0.0802	0.3104	0.3492	+0.0388	15.06%	15.17%	24/24
0.01000	4.147%	0.25130	0.25584	+0.00453	1.0661	1.0842	+0.0181	0.4639	0.4800	+0.0162	15.91%	15.97%	24/24

Table 7: Descriptive originwise diagnostics for TAFAS at $B/H = 0.0025$. Increased is the fraction of origins with $a_{\mathrm{TC}} > a_{\mathrm{ind}}$; Win conditions on that event and reports the fraction with lower TC raw MSE. Counts sum the three fixed-seed paths. The fixed action-grid diagnostic is retrospective and does not affect issuance.

Dataset	Increased	Win	Mean $\mathrm{MSE}_{\mathrm{TC-ind}}$	Median $\mathrm{MSE}_{\mathrm{TC-ind}}$	Wins/ties/losses	Grid opt. $> a_{\mathrm{ind}}$	Grid opt. $> a_{\mathrm{TC}}$	Corr.
ETTh2	8.35%	84.85%	-3.083×10^{-5}	-2.210×10^{-5}	810/6444/1041	59.57%	59.57%	−0.0855
ETTm1	62.99%	66.98%	-1.387×10^{-3}	-3.619×10^{-4}	17349/11739/10563	57.73%	54.55%	+0.0032
Exchange	0.00%	–	–	–	0/2394/0	60.03%	60.03%	–
Weather	22.12%	64.55%	-3.877×10^{-5}	-7.707×10^{-5}	9327/4275/15858	69.55%	69.99%	−0.1598

Table 8: Dataset-level issued-path constraint diagnostics. Bind is the percentage of origins whose issued action is below the causal proposal. $U_{\max}$ and U_{mean} are the maximum and mean issue-prefix budget utilization. Negative mean utilization can occur because settled cumulative excess is signed. Valid counts both evaluation paths; Exposed counts seed-level paths with a nonzero same-action denominator.

Candidate	Dataset	$\mathrm{Bind}_{\mathrm{ind}}$	$\mathrm{Bind}_{\mathrm{TC}}$	$U_{\max,\mathrm{ind}}$	$U_{\max,\mathrm{TC}}$	$U_{\mathrm{mean},\mathrm{ind}}$	$U_{\mathrm{mean},\mathrm{TC}}$	Valid	Exposed
TAFAS (2025)	ETTh2	22.31%	22.31%	1.000	1.000	0.375	0.373	6/6	3/3
	ETTm1	70.38%	57.03%	1.000	1.000	0.711	0.330	6/6	3/3
	Exchange	0.00%	0.00%	0.000	0.000	0.000	0.000	6/6	0/3
	Weather	85.49%	84.51%	1.000	1.000	0.865	0.863	6/6	3/3
DSOF (2025)†	ETTh2	73.48%	72.67%	1.000	1.000	0.737	0.737	6/6	3/3
	ETTm1	53.16%	42.60%	1.000	1.000	−6.372	−10.495	6/6	3/3
	Exchange	8.81%	8.77%	1.000	1.000	0.093	0.093	6/6	3/3
	Weather	92.38%	92.35%	1.000	1.000	0.925	0.925	6/6	3/3
COSA (2026)	ETTh2	25.15%	25.00%	1.000	1.000	0.321	0.320	6/6	3/3
	ETTm1	56.59%	52.28%	1.000	1.000	0.404	0.137	6/6	3/3
	Exchange	0.00%	0.00%	0.000	0.000	0.000	0.000	6/6	0/3
	Weather	74.19%	73.83%	1.000	1.000	0.799	0.802	6/6	3/3

Table 9: Effect of target-coherent evaluation on the worst-case bound and issued action. FreTS rows average three checkpoints within each dataset; every PatchTST row is one trajectory. "Reduction" compares target-coherent and itemwise values on the actions issued by TC-FSRE. MSE and MAE are descriptive improvements over the frozen base, and each Δ column is target-coherent minus itemwise.

Backbone	Dataset	Red. (%)	a_{ind}	a_{TC}	Δa	$\mathrm{MSE}_{\mathrm{ind}}$(%)	$\mathrm{MSE}_{\mathrm{TC}}$(%)	ΔMSE (pp)	ΔMAE (pp)
FreTS-H192	ETTh1	55.41	0.3142	0.4784	+0.1642	0.3854	0.4748	+0.0894	+0.0722
	ETTh2	61.33	0.4496	0.6771	+0.2275	0.7172	0.9538	+0.2367	+0.1097
	ETTm1	77.13	0.7486	0.8067	+0.0580	0.4956	0.5239	+0.0283	+0.0200
	ETTm2	67.73	0.7278	0.7669	+0.0391	0.4138	0.4449	+0.0312	+0.0127
	Weather	86.51	0.8605	0.9528	+0.0923	0.2767	0.2961	+0.0194	+0.0145
	Exchange	81.32	0.6173	0.8552	+0.2378	0.3482	0.4353	+0.0871	+0.0502
PatchTST-H720	ETTh1	83.67	0.2730	0.6136	+0.3406	0.0294	0.0457	+0.0163	+0.0093
	ETTh2	70.37	0.1475	0.4642	+0.3167	0.0987	0.1804	+0.0818	+0.0448
	ETTm1	85.23	0.4287	0.7579	+0.3292	0.0864	0.1133	+0.0268	+0.0157
	ETTm2	89.05	0.8136	0.8275	+0.0139	0.0169	0.0172	+0.0003	+0.0002
	Weather	91.74	0.7434	0.9243	+0.1809	0.0567	0.0663	+0.0096	+0.0103
Mean over 11 clusters		77.23	0.5568	0.7386	+0.18184	0.2659	0.3229	+0.05699	+0.03268
95% cluster bootstrap CI		[70.25,83.62]	–	–	[0.11513,0.24851]	–	–	[0.02507,0.09979]	[0.01570,0.05309]

Table 10: Summary of mechanism checks. Same-action reduction always compares completion geometries on a fixed issued forecast set. "Valid" means that every reported path satisfied Eq. (3).

Test	Controlled setting	Same-action bound	Issued action/protocol result	Supported claim
No-overlap control	ETTh1 and Weather; maximum overlap 1	Reduction 0.00% in both cases	Identical TC and itemwise actions	The gain comes from shared targets, not fixed rescaling
Partition intervention	Two tasks; maximum overlap 2–48	True partition never exceeds a safe refinement	At overlap 48: $0.0728 \rightarrow 0.1162$ and $0.5519 \rightarrow 1.0000$	Partition-refinement monotonicity
Heterogeneous candidates	Two COSA paths and one ELF path	7.01%–21.41%	Action change $+0.00245$–$+0.0368$	The risk layer uses only the forecast-space interface
Foundation × learned adapter	Frozen TimesFM 2.5 + causal COSA; four PEMS units	Mean 1.11%; range 0–3.32%	24/24 certificate paths valid; action rose in 2/4 units	Compatibility persists under negative adapter transfer
Second loss	clipped-nMSE; 12 fixed pending states	Mean 62.49%, minimum 46.27%	Exact solver valid in 12/12 states	A fixed-set joint reduction also occurs under clipped-nMSE
Asynchronous maturity	Group size refined from H to 1	Resolved terms leave the pending set earlier	FreTS/PatchTST actions rose to $0.6197/0.4813$; 23/23 valid	Target-level updates enlarge later feasible sets
Feedback protocol	Additional/random delays, missingness, reversed batches	22.47%–37.20%	All 6 settings pathwise valid	Event order changes the state, not the invariant

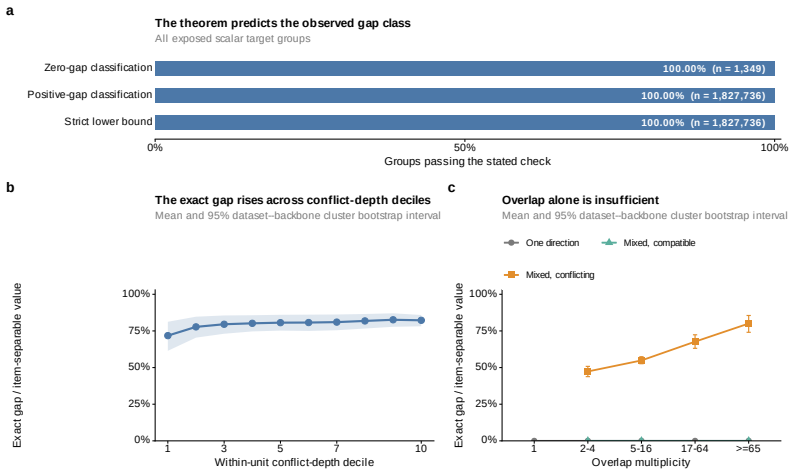


Figure 3: Conflict, not overlap alone, explains the coherent–itemwise gap. **a**, Exact theorem verification; **b**, conflict-depth deciles; **c**, overlap and direction class. Intervals resample 11 dataset–backbone units.

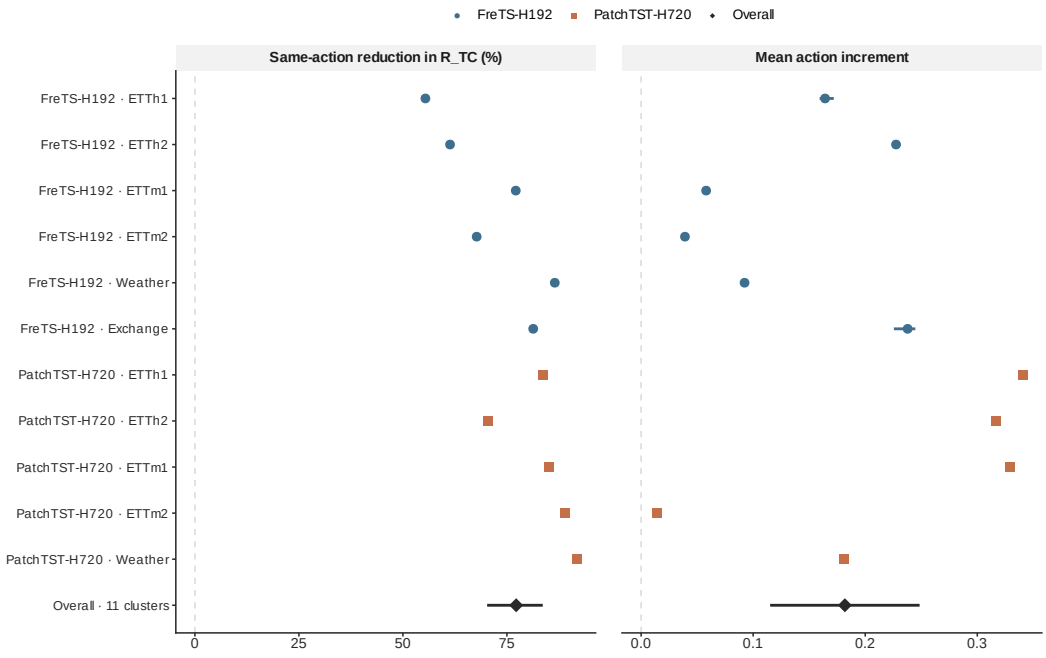


Figure 4: Clusterwise effect of target-coherent completion. The left panel reports the same-action reduction in the nonnegative risk amount, and the right panel reports the increase in mean issued action. Each row is one panel–dataset cluster. Horizontal segments for FreTS show the minimum and maximum over three checkpoints and are not confidence intervals; PatchTST rows contain one retained trajectory. The overall diamonds and intervals are the mean and 95% cluster-bootstrap interval over the 11 statistical clusters. Origins and forecast terms are not resampled as independent observations.

C.3 SOLVER VALIDATION AND TARGET-DEPENDENCE INTERVENTIONS

Table 11: Exact-solver and necessity checks for fixed issued pending sets under stop-and-complete. Errors are measured against independent NumPy enumerators.

Object	Cases	Maximum error	Outcome
NMAE joint solver	400	4.88×10^{-15}	All passed
clipped-nMSE joint solver	400	3.38×10^{-14}	All passed
$\Lambda_{\mathrm{TC}} - \epsilon$ witness	500	2.84×10^{-14}	500/500 violated

Gap-mechanism validation. For each of the 23 main-panel trajectories, we replayed the issued TC-FSRE action path. Immediately before each target settlement, we recorded every fully accumulated scalar group indexed by absolute time and channel, then computed its exact item-separable value, target-coherent value, gap, conflict depth from Corollary 2, overlap multiplicity, and direction class. This produced 1,897,204 rows, of which 1,829,085 had a positive item-separable bound. Observed zero was defined with the scale-aware tolerance $\max(10^{-18}, 10^{-8}\Lambda_{\mathrm{ind},G})$; changing this tolerance cannot affect the exact theorem classification stored separately from the floating-point comparison. Figure 3 uses within-unit deciles because raw conflict depth inherits dataset scale. Its intervals come from 10,000 deterministic cluster-bootstrap resamples of the 11 dataset–backbone units. The editable vector figure and panel-level source tables are generated from these group-level records by the accompanying scripts.

Table 12: Full target-partition and overlap intervention. Forecasts, candidates, labels, proposed actions, and thresholds are fixed; only origin stride or partition changes.

Dataset	Maximum overlap	Partial red. (%)	True red. (%)	$a_{\mathrm{singleton}}$	a_{partial}	a_{true}
ETTh1	1	0.00	0.00	0.9909	0.9909	0.9909
ETTh1	2	0.00	13.41	0.9507	0.9507	0.9752
ETTh1	6	15.80	27.43	0.5862	0.6625	0.7198
ETTh1	12	28.70	36.59	0.2356	0.3254	0.3679
ETTh1	48	26.48	32.70	0.0728	0.1050	0.1162
Weather	1	0.00	0.00	1.0000	1.0000	1.0000
Weather	2	0.00	9.33	1.0000	1.0000	1.0000
Weather	6	16.68	34.72	1.0000	1.0000	1.0000
Weather	12	34.01	51.24	0.9832	1.0000	1.0000
Weather	48	68.78	80.28	0.5519	1.0000	1.0000

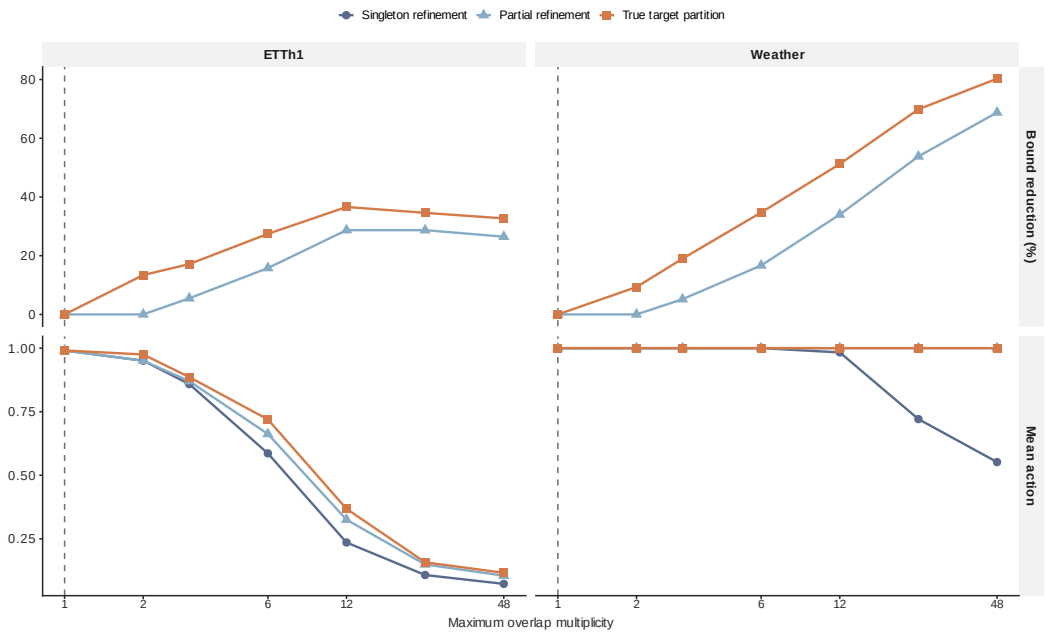


Figure 5: Full target-partition and overlap intervention. The upper row gives the reduction in the same-action nonnegative risk amount relative to the singleton refinement; the lower row gives the issued action under the three partitions. Forecasts, labels, proposed actions, and thresholds are fixed. At overlap multiplicity one the reduction is exactly zero. Legal merging by the true target partition never gives a larger risk amount than either safe refinement.

C.4 Candidate interfaces and a second controlled loss

Table 13: Heterogeneous non-Hankel candidate cases. These results test compatibility with the forecast-space interface, not broad generalization across candidates.

Candidate	Backbone/task	Bound reduction	Action change	Itemwise MSE	TC MSE
COSA	FreTS/ETTh1/H192	7.01%	+0.0368	+0.8244%	+0.9189%
COSA	PatchTST/ETTh2/H720	7.22%	+0.00245	+0.00562%	+0.01401%
ELF	FreTS/Weather/H192	21.41%	+0.0169	+0.5851%	+0.7213%

C.5 Asynchronous feedback and maturity

Table 14: Retrospective frozen-TimesFM 2.5–learned-COSA evaluation; each row averages three seeds. NMAE is the controlled loss. Gains are percentages relative to the frozen base, so negative values denote degradation. Bind is the percentage of test origins on which the issued action is below the causal FTL proposal; Harm is the percentage with higher raw MSE than the base. Valid reports the path certificate. Red is the same-action bound reduction and is shown once per dataset on the TC row. The unprojected full candidate is not assigned a certificate.

Dataset	Method	NMAE	NMAE gain	MSE gain	MAE gain	Mean a	Bind	Harm	Valid / Red
PEMS03	Frozen base	0.098	0.000	0.000	0.000	0.00000	–	0.00	yes / –
	Full candidate	0.100	−1.983	−1.884	−2.051	1.00000	–	61.83	– / –
	Itemwise-FSRE	0.098	0.000	0.000	0.000	0.00000	0.00	0.00	yes / –
	TC-FSRE	0.098	0.000	0.000	0.000	0.00000	0.00	0.00	yes / 0.00%
PEMS04	Frozen base	0.107	0.000	0.000	0.000	0.00000	–	0.00	yes / –
	Full candidate	0.114	−6.523	−11.618	−6.472	1.00000	–	96.83	– / –
	Itemwise-FSRE	0.108	−0.581	−1.011	−0.573	0.06489	34.17	34.83	yes / –
	TC-FSRE	0.108	−0.591	−1.029	−0.583	0.06694	34.17	34.83	yes / 3.32%
PEMS07	Frozen base	0.059	0.000	0.000	0.000	0.00000	–	0.00	yes / –
	Full candidate	0.067	−14.644	−31.438	−14.348	1.00000	–	99.50	– / –
	Itemwise-FSRE	0.059	−0.534	−1.120	−0.540	0.06020	0.00	6.67	yes / –
	TC-FSRE	0.059	−0.534	−1.120	−0.540	0.06020	0.00	6.67	yes / 0.62%
PEMS08	Frozen base	0.066	0.000	0.000	0.000	0.00000	–	0.00	yes / –
	Full candidate	0.075	−12.559	−42.672	−12.481	1.00000	–	83.17	– / –
	Itemwise-FSRE	0.067	−0.953	−3.886	−0.939	0.02768	11.83	8.50	yes / –
	TC-FSRE	0.067	−0.955	−3.896	−0.941	0.02781	11.83	8.83	yes / 0.49%

Table 15: Full maturity-granularity curves. Group size 1 processes individual leads; group size H waits for the full horizon. Every path satisfied the constraint.

Panel	Group size	MSE improvement (%)	Mean action	Mean pre-issue margin	Valid paths
FreTS-H192	1	0.4395	0.6197	2.3358	18/18
	4	0.4380	0.6153	2.3106	18/18
	16	0.4206	0.5932	2.1979	18/18
	48	0.3404	0.5160	1.7268	18/18
	192	0.1900	0.3301	0.4360	18/18
PatchTST-H720	1	0.0576	0.4813	0.4258	5/5
	12	0.0571	0.4768	0.4165	5/5
	48	0.0546	0.4616	0.3904	5/5
	180	0.0412	0.4010	0.3214	5/5
	720	0.0190	0.2308	0.3001	5/5

Table 16: Event-driven feedback stress test on ETTh1 ($H = 24$, 2 channels, 120 origins). Forward and reverse within-batch processing under random delay produced identical states.

Feedback protocol	Mean action	Pending groups	Pending risk	Same-action red. (%)	Minimum margin	Valid
Delay 0	0.1054	48	0.0406	35.19	9.94×10^{-13}	Yes
Delay 5	0.0928	58	0.0678	37.20	2.70×10^{-13}	Yes
Delay 24	0.0371	96	0.0548	28.90	6.61×10^{-13}	Yes
Random delay	0.0568	74	0.0562	34.05	5.62×10^{-13}	Yes
Random delay, reversed batch	0.0568	74	0.0562	34.05	5.62×10^{-13}	Yes
45% deployment feedback missing	0.0206	161	0.0519	22.47	6.59×10^{-13}	Yes

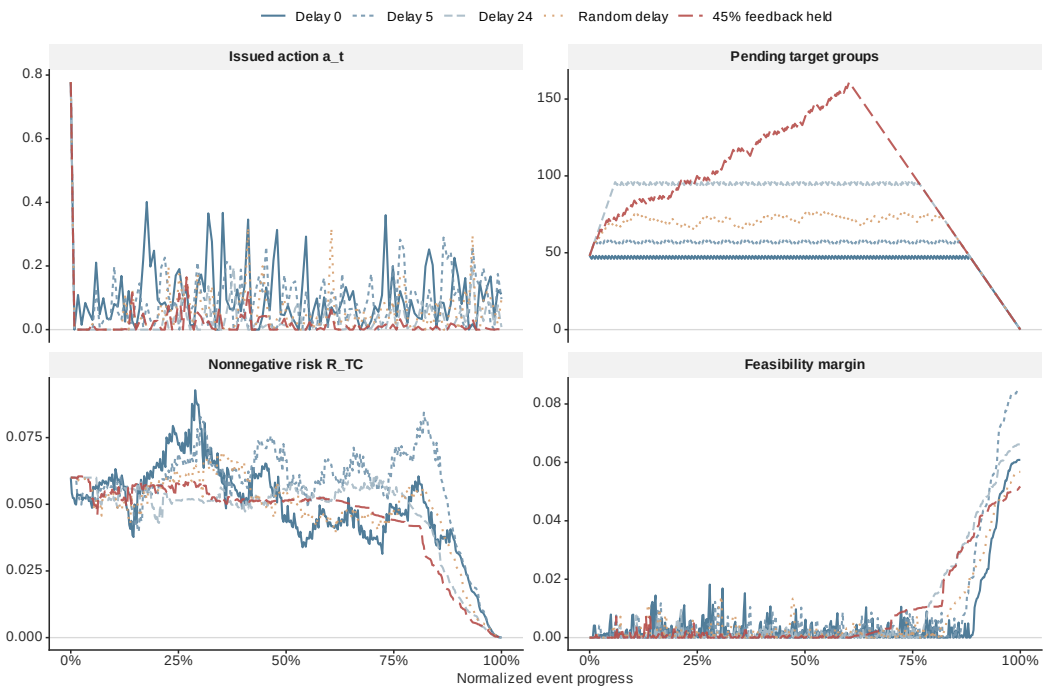


Figure 6: Event-wise trajectories under delayed and incomplete feedback. Actions are shown only at issue events; the other panels follow every issue or settlement event and use normalized event progress to align paths. Larger delays and held feedback retain more target groups and change both R_{TC} and the feasible action path. The feasibility margin remains nonnegative, and all five plotted paths close with valid prefix constraints. The reverse-batch control is omitted because its state and action trajectory is numerically identical to the random-delay path.

C.6 Policy and loss sensitivity

Table 17: Outer policy interface. The standard stream reports two controlled losses. The recurrent-switch stress test reports the raw policy and its clipped-nMSE-wrapped counterpart.

Scenario	Policy	Raw MSE (%)	NMAE wrapper (%)	clipped-nMSE wrapper (%)	Worst clipped-nMSE (%)	Wins/tasks	Valid paths
Standard FreTS	FTL	–	0.4395	0.0742	0.2602	18/18	36/36
	Hedge	–	0.4328	0.0799	0.2106	18/18	36/36
	AdaHedge	–	0.4369	0.0800	0.2673	18/18	36/36
Recurrent switch	FTL	−14.8727	–	0.0467	−1.3776	9/18	18/18
	Hedge	−11.4266	–	0.9227	−0.2066	15/18	18/18
	AdaHedge	−65.6548	–	0.9227	−0.2066	15/18	18/18

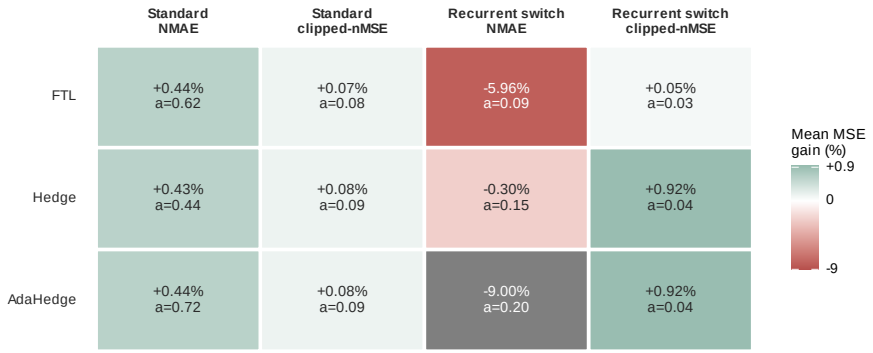


Figure 7: Policy–loss interaction on the standard and recurrent-switch streams. Cell colour and the first printed value give mean raw-MSE improvement; the second value gives mean issued action. All displayed entries use the pathwise wrapper and satisfy 18/18 constraints. The recurrent-switch stress test separates the loss contracts: NMAE remains valid but can have negative raw-MSE utility, whereas the clipped-nMSE wrapper limits the issued action and gives positive mean raw-MSE utility for all three policies.

Table 18: clipped-nMSE joint evaluation and threshold sensitivity. The joint-evaluation row uses true target groups and fixed action 0.5. The remaining rows show the risk–utility sensitivity of itemwise ASYNC-FSRE and were not used to select the main threshold.

Setting	τ	FreTS wins/tasks	FreTS MSE (%)	PatchTST wins/tasks	PatchTST MSE (%)	Valid paths
TC joint evaluation (12 states)	1	–	Mean reduction 62.49%, minimum 46.27%		–	12/12
ASYNC-FSRE sensitivity	0.5	18/18	0.3387	5/5	0.0506	23/23
ASYNC-FSRE sensitivity	1	18/18	0.2443	5/5	0.0298	23/23
ASYNC-FSRE sensitivity	2	18/18	0.1300	5/5	0.0159	23/23
ASYNC-FSRE sensitivity	3	18/18	0.0742	5/5	0.0107	23/23
ASYNC-FSRE sensitivity	5	18/18	0.0352	5/5	0.0063	23/23
ASYNC-FSRE sensitivity	10	18/18	0.0146	5/5	0.0031	23/23

Solver and certificate validation. The independent NMAE and clipped-nMSE solver checks each covered 400 random or degenerate sets. For 500 fixed issued pending sets with $\Lambda_{\mathrm{TC}} > 0$, lowering the sharp value by a prescribed ϵ produced a legal completion that violated the smaller value in every case. All 23 main TC-FSRE paths satisfied the constraint.

D ADDITIONAL ANALYSIS AND COMPUTATIONAL COST

Table 19: Risk–utility boundary at the main and relaxed thresholds. Binding rate is the fraction of origins whose proposed action was constrained.

Task	Threshold rate	Method	Mean action	MSE improvement (%)	Binding rate (%)	Harmful origins (%)
FreTS-ETTh1	0.0025	Independent	0.3181	0.3866	64.93	7.50
FreTS-ETTh1	0.0025	Target-coherent	0.4774	0.4710	1.94	13.09
PatchTST-ETTh2	0.0025	Independent	0.1475	0.0987	71.07	1.66
PatchTST-ETTh2	0.0025	Target-coherent	0.4642	0.1804	8.68	10.96
PatchTST-ETTh2	0.0100	Independent	0.4815	0.1859	–	–
PatchTST-ETTh2	0.0100	Target-coherent	0.5088	0.1810	–	–

Across the main panel, an average of 13.28% of origins under TC-FSRE had raw MSE above the frozen base. At threshold rate 0.01, target-coherent evaluation issued a larger action on PatchTST-

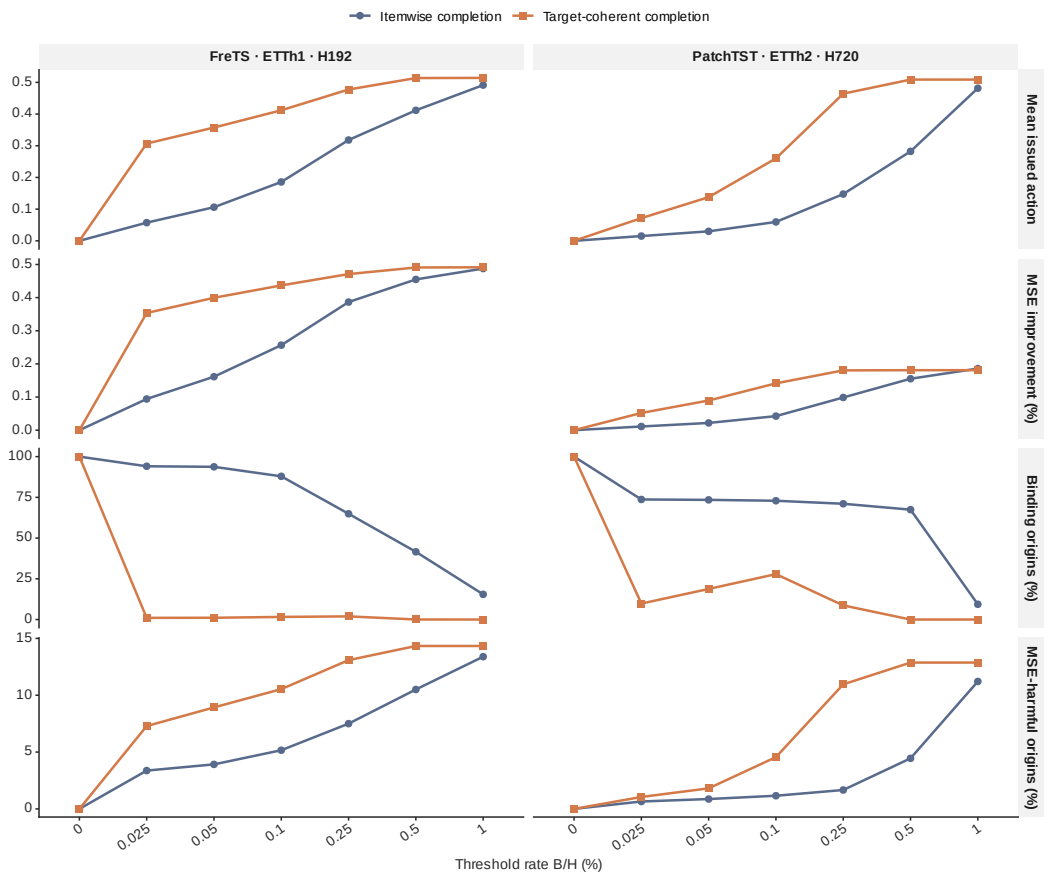


Figure 8: Full threshold–action–accuracy–worst-harm frontier over seven cumulative excess thresholds. A harmful origin has raw MSE above the frozen base. A smaller target-coherent value enlarges the feasible action set but does not imply originwise accuracy dominance.

ETTh2 but improved MSE by 0.00484 percentage points less than the itemwise method. Neither observation violates the cumulative NMAE excess guarantee.

Table 20: Preregistered M4/TimesFM confirmation. The accuracy criterion failed, although both wrapped methods satisfied the path constraint on all 12 paths.

Method	Wins/tasks	Mean improvement (%)	Worst improvement (%)	Valid paths
Directional raw-MSE envelope	5/12	0.0137	−0.0111	12/12
Exact clipped-MSE ($\tau = 1$)	5/12	0.0926	−0.2282	12/12
FTL	5/12	−0.0743	−2.1112	–
Hard Hankel	4/12	−3.2624	−16.5380	–
Identity	0/12	0.0000	0.0000	–

Preregistration provenance. The source revision retained for the M4 confirmation could not be matched to the hash recorded at preregistration because the original snapshot is unavailable. We therefore reran the frozen evaluation as a versioned amendment; the 5/12 accuracy result, 12/12 path validity, and all registered decisions were unchanged, with a maximum rowwise numerical deviation of 4.96×10^{-7}. The source mismatch remains documented, while the reported decisions and constraint checks are unchanged.

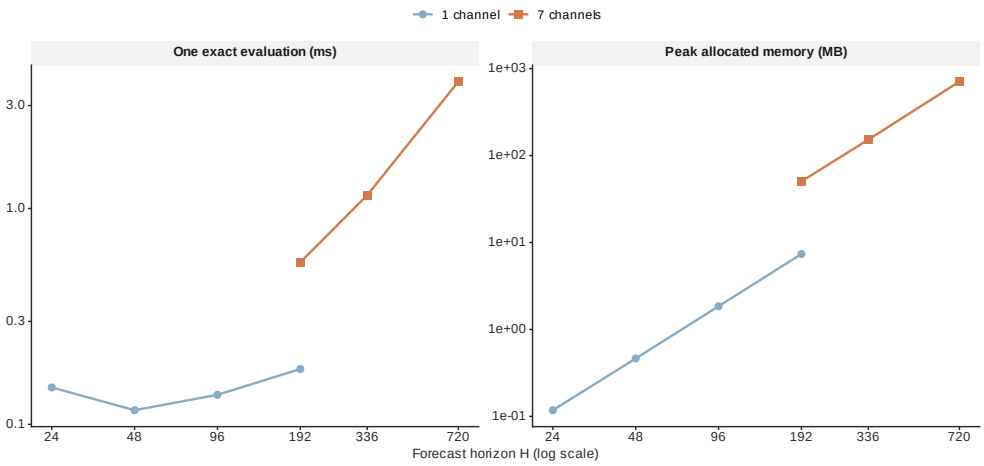


Figure 9: GPU latency and peak memory for one joint evaluation. Each point reports the median of five runs after two warm-up runs on an RTX 3090 Ti in float64. These measurements exclude the repeated bisection calls needed for end-to-end issuance.

Table 21: One joint evaluation by the exact dense reference implementation. Constrained issuance also calls one proposed-action evaluation and up to 12 bisection evaluations.

Shape	Median latency	Peak memory	Resident state
192×7	0.564 ms	50.50 MB	6.49 MB
336×7	1.153 ms	152.61 MB	19.88 MB
720×7	3.872 ms	708.86 MB	92.25 MB

The dense implementation stores an $H \times H \times C$ state and costs $O(CH^2 \log H)$ per evaluation. Increasing the overlap stride from 1 to 8 for H720×7 reduced the active-term fraction from 50.07% to 6.32%, but median latency remained 3.872/3.871 ms and peak memory remained 708.86 MB. The reference implementation therefore does not yet exploit sparse overlap. End-to-end median costs on the main paths were 1.21 ms/origin for FreTS and 6.80 ms/origin for PatchTST. These measurements include the realized constraint-activation pattern and should not be conflated with the single-evaluation results in Table 21.