DO TSFMS SUFFER FROM UNRELIABLE CONTEXTS? RETHINKING CONTEXT USAGE IN FORECASTING WITH TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) have recently improved frozen-backbone forecasting by scaling model capacity, pretraining corpora, and domain coverage. Unlike large-language-model inference, where prompt construction has become an explicit object of optimization, TSFM inference typically passes historical context directly to the forecaster and implicitly treats its observations as equally useful and trustworthy. This assumption can fail when the context contains stale regimes, local anomalies, misleading trends, noise, or irrelevant history. Scaling a forecasting backbone alone does not decide which evidence should be trusted for a particular instance. We therefore study **evidence-aware forecasting**: how should context become a selectable decision variable before a frozen forecaster models the future? The difficulty is that context reliability depends on both the instance and the forecaster, while the future outcome needed to assess an intervention is unavailable at inference. We introduce CARE-TSFM, a frozen-backbone inference-adaptation framework whose external controller is trained offline with counterfactual utility supervision from pre-test data. CARE represents seven evidence patterns, estimates the utility of 24 candidate interventions, and uses a risk gate to select context refinement. Across GIFT-Eval and six frozen TSFMs, CARE achieves consistent multi-metric improvements for all forecasters while revealing structured, forecaster-dependent intervention utility. These results suggest that treating context selection as a decision variable is a promising direction for evidence-aware inference.

1 INTRODUCTION

Recent progress in TSFMs has strengthened frozen-backbone forecasting by scaling model capacity, pretraining corpora, and domain coverage. The Chronos series moves from language-model-style tokenization (Ansari et al., 2024) to universal multivariate forecasting (Ansari et al., 2025), while the TimesFM line progresses from large-scale decoder pretraining (Das et al., 2024) to covariate-aware in-context forecasting (Jain & Sen, 2026). Yet this scaling trajectory leaves a separate inference question untouched: historical context is still passed directly to a forecaster. This interface implicitly assumes that the supplied context reliably describes the current process and that its observations are equally useful and trustworthy.

The assumption becomes fragile when parts of the history no longer describe the process to be forecast. Figures 1(a) and 1(b) show how recent-regime adaptation and trend adjustment, respectively, mitigate Level shift and a misleading trend. Beyond these two cases, Noise, Local spike, Stale history, Trend inconsistency, Seasonal structure, and Endpoint instability can also make parts of a context unreliable for forecasting. These recurring phenomena motivate a context-reliability taxonomy; they may also co-occur within one window. Figure 1(c) moves beyond individual cases by comparing configuration-level MASE distributions on GIFT-Eval (Aksu et al., 2025). Seasonal Naive is a task-difficulty reference; raw-context TimesFM-3 represents a current state-of-the-art frozen TSFM; and oracle-selected context measures headroom within the fixed intervention set. Their median MASE values are 1.27, 0.98, and 0.83. As direct backbone scaling yields increasingly incremental gains, this offline candidate-set analysis raises a constructive question: can context optimization substantially expand the remaining room for improvement?

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.

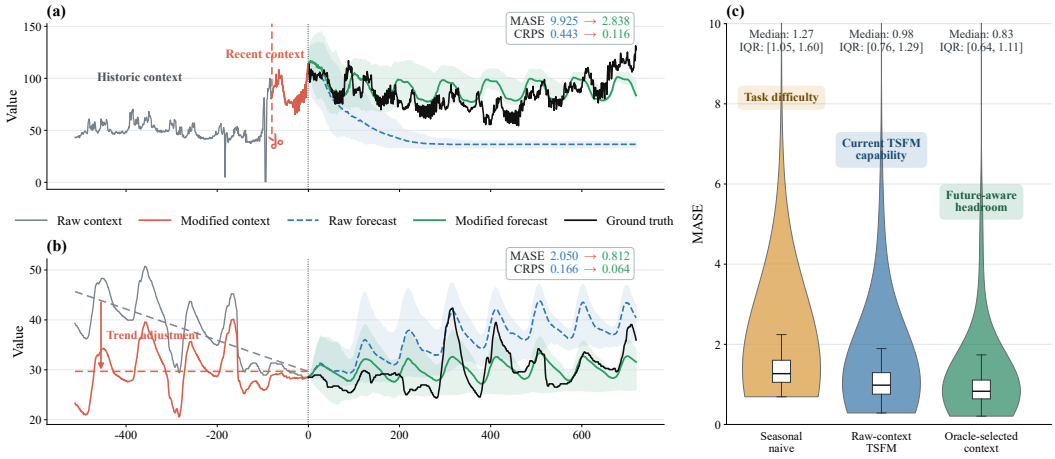


Figure 1: Context interventions reveal forecasting headroom. (a,b) Two examples in which context refinement improves a TSFM forecast. (c) TSFM forecasting substantially improves the MASE distribution over Seasonal Naive, while oracle-selected context exposes additional headroom within the fixed candidate set. The oracle is used for offline candidate-set analysis.

Recent work establishes that context handling can improve forecasting independently of the backbone. Long-context studies show that more history is not invariably useful (Butera et al., 2026); Credal-TTA adapts stale history through reset-and-grow context adaptation (Du et al., 2026); EFE-Time learns dataset-level transformation programs (Taga et al., 2026); and TSRBench evaluates forecasting robustness under context corruptions (Kim et al., 2026). These efforts remain fragmented, however: they address restricted forms of unreliability or apply a common transformation, whereas stale regimes, local corruption, and unstable endpoints call for different responses on different windows. Figure 1(c) quantifies this headroom and motivates the missing inference step: before the future is observed, a controller must infer which evidence is informative for the current window and which candidate intervention is worth applying.

A tempting solution is to detect an apparent context problem and apply its corresponding correction, for example by denoising a rough history or truncating stale observations. This view is too simple. The system must use only the raw context, the frozen forecaster's raw response, and inference-visible metadata, such as horizon, seasonal period, and observable context length, to anticipate whether the unseen future will benefit; a semantically matched intervention can behave differently across windows and TSFMs; and even the most promising intervention can be worse than leaving the context untouched. Context intervention is therefore a selective decision problem under uncertain utility, not merely a transformation problem.

CARE-TSFM turns this selective decision into an evidence-aware context policy. We organize recurring context problems into seven context-failure evidence families that can co-occur. Each family motivates several plausible intervention mechanisms, forming a candidate intervention space. These mechanisms are hypotheses rather than prescribed repairs: an offline controller learns their instance- and forecaster-dependent utility, while a validation-calibrated gate retains the raw context when predicted utility is insufficient. The forecasting backbone remains frozen.

We evaluate the frozen-backbone inference-adaptation policy on GIFT-Eval across six TSFMs. CARE consistently improves MASE and CRPS, while the magnitude and form of the gain vary with the forecaster. A configuration-level audit reveals structured intervention utility and learned ranking beyond random choice. The analyses characterize informative evidence–intervention pairs, isolate the contribution of each component, study forecaster specificity through source–target controller evaluation, and map the coverage–risk behavior induced by the gate. Our contributions are:

1. We uncover an overlooked perspective for TSFM inference: the historical context is not automatically trustworthy, so context reliability itself should be treated as a decision variable before forecasting.

2. We introduce CARE-TSFM, the first evidence-aware selective context-intervention framework for this setting. CARE uses context-failure evidence to motivate a candidate intervention space, learns instance- and forecaster-dependent utility over that space, and gates whether to refine or retain context while keeping the forecasting backbone frozen.
3. We reveal a forecaster-specific evidence–intervention utility landscape in which the value of a context intervention changes across instances and TSFMs. This finding explains why context repair requires learned, selective decisions and provides a basis for studying context usage as a new axis of TSFM improvement.

2 Related Work

2.1 From TSFM Scaling to Evidence-Aware Inference

TSFM research has followed a coherent capability trajectory: architecture-level methods sharpen temporal representation through multi-period modeling, patching, or multivariate tokenization (Wu et al., 2023; Nie et al., 2023; Liu et al., 2024a), building on earlier decomposable architectures such as N-BEATS (Oreshkin et al., 2020); foundation-model work scales this capability across data, model size, tasks, and domains, including generative pretraining for large time-series models such as Timer (Liu et al., 2024b) and universal forecasting across heterogeneous domains (Woo et al., 2024; Goswami et al., 2024); recent systems further enrich the forecasting interface with multivariate and covariate-aware inputs. These advances improve what a forecaster can represent and generalize over, while retaining a direct context-to-forecast interface. Input-side adaptation offers a complementary perspective: context construction can itself become an optimization target. CARE develops this perspective into an evidence-aware decision framework that estimates intervention utility and selects context modifications before frozen TSFM forecasting.

2.2 Context Adaptation: From Modification to Selection

Context selection and input transformation provide the closest methodological setting. Long-context forecasting shows that more history is not invariably useful, while test-time adaptation updates model states or parameters under distribution shift (Wang et al., 2021). Probabilistic forecasting models such as DeepAR also make the forecast distribution, rather than only a point estimate, a first-class output (Salinas et al., 2020). Credal-TTA is especially close to our setting because it detects stale temporal evidence and adapts the usable context, but its primary reset-or-shorten response targets one failure mode and can discard useful long-range information. EFE-Time learns dataset-level transformation programs; such a policy cannot express that different windows from the same dataset may require different interventions or no intervention, and it does not provide an instance-level retain/abstain decision. These limitations motivate a window-level treatment of heterogeneous context failures and intervention utility while keeping the forecasting backbone frozen.

3 Problem Definition and Challenges

3.1 Frozen forecaster and offline-trained controller

Let $x_i \in \mathbb{R}^{L_i}$ denote a context of length L_i, and $y_i \in \mathbb{R}^{H_i}$ its H_i-step future. A pretrained TSFM f_θ, with fixed parameters θ, returns the probabilistic forecast $\widehat{P}_i = f_\theta(x_i)$, represented as point forecasts, quantiles, or samples by a fixed adapter. The forecasting backbone receives no task-specific fine-tuning and cannot access the test target y_i. An external policy π_ϕ is learned offline from pre-test counterfactual utilities, with gate thresholds selected on held-out pre-test windows. Both θ and ϕ are fixed at test time.

3.2 Context interventions and realized utility

The standard interface implicitly assumes that x_i provides appropriate evidence for predicting y_i. This is stronger than assuming correctly measured observations: an error-free history may still contain an obsolete regime, an unrepresentative endpoint, a misleading trend, or a valid but currently irrelevant segment. Let $\mathcal{A} = \{a_0, a_1, \ldots, a_{24}\}$ denote the candidate intervention set, where a_0 is

identity and the remaining elements index interventions. For any intervention indexed by $a \in \mathcal{A}$, $\widehat{P}_i^a$ is its forecast and $L_{i,a,m}$ is its loss under metric m from $\mathcal{M}$, where $\mathcal{L}_m$ denotes the corresponding loss function:

$$\widehat{P}_i^a = f_\theta(a(x_i)), \qquad L_{i,a,m} = \mathcal{L}_m(y_i, \widehat{P}_i^a), \quad m \in \mathcal{M} = \{\text{MASE}, \text{CRPS}\}.$$

Because both losses are minimized, the realized absolute utility relative to identity is

$$\Delta_{i,a,m} = L_{i,a_0,m} - L_{i,a,m}.$$

Positive values indicate improvement, negative values indicate harm, and identity has zero utility by definition. The future-conditioned oracle intervention $a_{i,m}^{\text{oracle}}$ selects the highest-utility candidate for window i and metric m:

$$a_{i,m}^{\text{oracle}} = \arg\max_{a \in \mathcal{A}} \Delta_{i,a,m} = \arg\min_{a \in \mathcal{A}} L_{i,a,m}$$

is used only as an offline upper-bound diagnostic; y_i and $\Delta_{i,a,m}$ are unavailable at inference. We use "realized absolute utility" only for $\Delta_{i,a,m}$.

3.3 Challenges

Before stating the decision challenges, we ask whether a simple context modification can be treated as a universally beneficial preprocessing step. On TimesFM-3 with the GIFT-Eval benchmark, we apply three representative transformations globally to every forecasting window: recent-context truncation, denoising, and local repair. For each window and transformation a, we measure the realized MASE utility $U_{i,a} = \text{MASE}_i^{\text{raw}} - \text{MASE}_i^a$. Positive values indicate improvement and negative values indicate degradation. Figure 2 summarizes the resulting per-window distributions.

The distributions are centered near zero or below it, with substantial negative tails and only a minority of improving windows. Thus, even direct and intuitively motivated modifications can help some instances while harming others. This experiment deliberately uses a small intervention set and no evidence or learned ranking; its role is to establish the difficulty of global context repair. The later evidence-conditioned analysis studies why the same intervention can have different utility across context states.

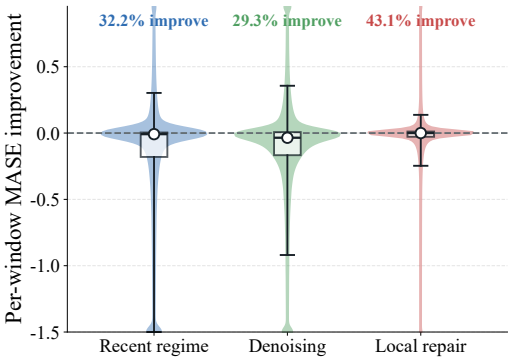


Figure 2: Heterogeneous utility of context modification. Each violin shows per-window MASE improvement relative to raw context; mass at or below zero indicates that no fixed modification reliably helps every window.

The formulation exposes three coupled challenges. **(1) Unobserved future benefit.** Before any refinement decision, the controller has only the raw context, the frozen TSFM's raw forecast, and metadata. These signals can suggest, but cannot reveal, whether the context will help or hurt forecasting; they therefore require a compact target-free context-failure evidence state rather than ground-truth failure labels. **(2) Uncertain intervention utility.** Changing the input of a complex frozen forecaster can help one window yet harm another, and the same intervention can behave differently across TSFM architectures. The controller must therefore estimate intervention utility for the current instance and forecaster instead of routing evidence to interventions by hand. **(3) Selective intervention.** Even the best-ranked intervention may still be worse than doing nothing: ranking indicates relative promise, not positive absolute utility. The controller must therefore decide whether predicted utility is sufficient to justify intervention, using an explicit retain option and validation-calibrated risk gate.

4 CARE-TSFM Method

CARE-TSFM places three input-side decisions before a frozen TSFM: context-failure evidence, intervention ranking, and risk-gated selection. Figure 3 summarizes the complete inference pipeline.

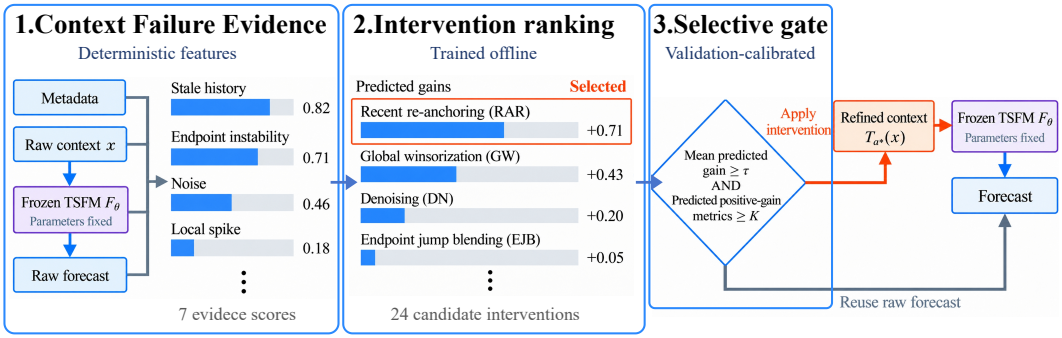


Figure 3: Overview of CARE-TSFM. Seven context-failure evidence families organize a candidate intervention space; the controller estimates candidate utility and a validation-calibrated risk gate selects refinement or retains the raw forecast with the TSFM frozen. Scores and gains are illustrative.

Table 1: Context-failure evidence families and candidate interventions. The columns present the seven evidence families and the eight mechanism groups organizing the 24 interventions. Identity denotes retention; cross-mechanism interventions may span mechanism groups.

Context-failure evidence (7)	Candidate intervention space: mechanism groups (8), interventions (24)
Level shift	Recent-regime adaptation: RC, RAC, RSM, RAR
Endpoint instability	Endpoint correction: EJB
Seasonal structure	Seasonal modification: TSB, SRB
Trend inconsistency	Low-frequency reconstruction: LFR
Stale history	Robustification: GW, IC, LM
Local spike	Local repair: ISI, SAI, AS
Noise	Denoising: DN, RDN
	Cross-mechanism: TEF, SSR, DN→ISI, ISI→DN, DN→AS, AS→DN, RDN→AS, AS→RDN

When the gate retains the context, CARE-TSFM reuses the raw forecast without an additional transformed-context call.

4.1 Context Failure Evidence

CARE-TSFM first evaluates the unmodified input, obtaining the raw forecast $\widehat{P}_i^{\text{raw}} = f_\theta(x_i)$. The deterministic state constructor ψ maps x_i, $\widehat{P}_i^{\text{raw}}$, and metadata to $z_i = \psi(x_i, \widehat{P}_i^{\text{raw}}, \text{metadata})$. The external policy π_ϕ uses the resulting 33-dimensional state, which contains seven context-failure evidence scores, four context-scale features, fourteen raw-forecast distribution features, and eight metadata features. The seven scores summarize Noise, Local spike, Level shift, Stale history, Trend inconsistency, Seasonal structure, and Endpoint instability; these patterns can co-occur. The scores are constructed from 26 lower-level feature identifiers listed in Appendix B. Table 1 presents the evidence taxonomy alongside the candidate intervention space, with each family motivating a set of mechanisms whose utility is learned from counterfactual supervision. Appendix B details state construction and score aggregation, Appendix E lists the intervention identifiers and compositions, and Appendix C specifies the data and leakage protocol.

Intervention abbreviations provide compact presentation labels; Appendix E gives their full identifiers and transformations. TEF and SSR are single registered operators that express two mechanisms, while the other six cross-mechanism interventions are ordered compositions.

4.2 Intervention-Utility Prediction and Ranking

CARE implements a utility-aware decision mechanism: from the observable state, it estimates candidate intervention utility, ranks the candidates, and lets the gate decide whether to execute the selected intervention or retain the raw forecast. The candidate space contains identity and 24 transformations of the observed history. The external policy π_ϕ learns utility prediction from offline counterfactual labels and remains fixed alongside the backbone during test inference. For each frozen TSFM, CARE trains two metric-specific multi-output ExtraTreesRegressor models for MASE and CRPS.

Each model maps the 33-dimensional state z_i to 24 predicted normalized controller targets, giving the controller two tree ensembles and 48 output coordinates. The state z_i is the sole inference input to these regressors; counterfactual forecasts and future targets define the offline training labels. We average predicted utilities into ranking score $s_{i,a}$ and select $a_i^\star$, the highest-scoring non-identity intervention:

$$s_{i,a} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \widehat{u}_{i,a,m}, \qquad a_i^\star = \operatorname*{arg\,max}_{a \in \mathcal{A} \setminus \{a_0\}} s_{i,a},$$

with fixed-order tie breaking. Validation selects and freezes one of three aggregation rules: mean gain, maximin gain, or average rank. Five-fold GroupKFold over `series_key` supplies out-of-fold predictions for ranking diagnostics and gate calibration before final regressors are refit on eligible controller-training windows. For controller fitting, the normalized controller target is $u_{i,a,m} = \Delta_{i,a,m} / \max(|L_{i,a_0,m}|, 10^{-8})$. This normalization preserves the realized absolute utility while stabilizing the learning scale. The two metric-specific targets form the "joint normalized utility" $u_{i,a}^{\text{joint}} = (u_{i,a,\text{MASE}} + u_{i,a,\text{CRPS}})/2$. The aggregation rule is selected on validation data and frozen before test evaluation: Chronos-2, Moirai-2, TimesFM-3, and Kairos use mean gain; Toto-2 uses average rank; and FlowState uses maximin gain.

4.3 Risk-Gated Selection and Context Refinement

Ranking identifies the strongest candidate; the risk gate determines whether its predicted utility supports execution. For the selected intervention $a_i^\star$, let g_i be its mean predicted utility across metrics and C_i its number of positive metric-specific predictions:

$$g_i = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \widehat{u}_{i,a_i^\star,m}, \qquad C_i = \sum_{m \in \mathcal{M}} \mathbf{1}[\widehat{u}_{i,a_i^\star,m} > 0].$$

Given validation-calibrated TSFM-specific thresholds (τ_θ, K_θ), where τ_θ is the mean-predicted-utility threshold and K_θ the minimum positive-prediction count, D_i is the intervention decision and A_i the final intervention. Here $\mathbf{1}[\cdot]$ is the indicator function:

$$D_i = \mathbf{1}[g_i \geq \tau_\theta \ \wedge \ C_i \geq K_\theta], \qquad A_i = \begin{cases} a_i^\star, & D_i = 1, \\ a_0, & D_i = 0. \end{cases}$$

Together, these conditions select instances whose predicted utility clears the validation operating point; coverage measures how often the policy intervenes.

5 Experimental Setup

5.1 Datasets and forecasting tasks

We evaluate on GIFT-Eval using its official temporal boundaries and rolling windows. CARE is fitted and calibrated only on chronologically ordered pre-test windows, with a later validation block held out for outer gate selection; no post-boundary targets enter controller training or gate selection. Further details are given in Appendix C.

5.2 Frozen TSFMs, Baselines, and CARE Configuration

We evaluate six pretrained TSFMs—Chronos-2 Small, Moirai-2 Small, Toto-2 4M, TimesFM 3.0, Kairos 10M, and FlowState Base—using their short names throughout. All checkpoints remain frozen and are paired with independent controllers and validation-calibrated gates. Baselines include raw TSFM, Seasonal Naive for oracle-potential analysis, Credal-TTA with shift detection and adaptive truncation, and fixed/global-intervention policies.

5.3 Metrics and aggregation

We report MASE and the benchmark-normalized nine-quantile CRPS approximation. The main table summarizes raw and CARE losses, configuration-macro RelGain, intervention rate, and MASE Win/Tie/Loss. RelGain is the macro-averaged per-configuration relative loss reduction. Positive gains indicate lower error, and MASE Win/Tie/Loss uses a fixed $\pm 0.1\%$ tie interval. Aggregation details are in Appendix C.

Table 2: CARE results on full GIFT-Eval using MASE and CRPS for controller training and test decisions. "+CARE" denotes the corresponding frozen TSFM with the CARE controller. RelGain is the configuration-macro relative reduction over all test windows.

				MASE		CRPS	
Model transition		Intervention rate	Win/Tie/Loss	Raw↓ → CARE↓	RelGain (%)↑	Raw↓ → CARE↓	RelGain (%)↑
Chronos-2	→ +CARE	14.04%	34 / 43 / 20	1.626 → 1.616	+0.79%	0.223 → 0.222	+0.18%
Moirai-2	→ +CARE	23.27%	45 / 38 / 14	1.718 → 1.694	+1.29%	0.238 → 0.236	+1.06%
Toto-2	→ +CARE	5.66%	25 / 64 / 8	1.721 → 1.719	+0.21%	0.239 → 0.238	+0.09%
TimesFM-3	→ +CARE	6.80%	27 / 59 / 11	1.600 → 1.576	+1.01%	0.223 → 0.220	+0.66%
Kairos	→ +CARE	73.41%	87 / 4 / 6	3.225 → 2.463	+10.05%	0.365 → 0.333	+9.00%
FlowState	→ +CARE	6.68%	27 / 66 / 4	1.615 → 1.607	+0.62%	0.227 → 0.226	+0.66%

6 RESULTS

The analyses follow the three decision questions above. Table 2 summarizes forecasting performance on GIFT-Eval, with baseline details in Table 8. Figure 4 and Table 4 ask whether inference-visible evidence carries useful intervention-utility signals; the component ablation on GIFT-Eval tests whether the evidence and intervention stages improve selection; and Figure 5 exposes the coverage–gain/improvement trade-off induced by selective gating. Appendix E extends these tests with the restricted adaptive-context baseline (Table 8), the five-model ablation table (Table 9), learned-ranking diagnostics (Table 12), and the forecaster-specificity study (Figure 6). These supplementary results are used to assess model dependence and mechanism boundaries, rather than to claim uniform transfer or universal superiority.

6.1 MAIN FORECASTING RESULTS

Table 2 shows consistent MASE and CRPS improvements across all six evaluated TSFMs, despite substantial differences in intervention rate. Gains are largest for Kairos, while other models benefit from more selective changes, showing that useful context refinement depends on the forecaster rather than intervention frequency alone. RelGain is macro-averaged over all 97 configurations. Credal-TTA serves as a representative adaptive-context baseline focused primarily on stale-history adjustment; its reproduction is discussed in Appendix E. The supplementary ablation table further shows that indiscriminate global or single-intervention policies are not consistently beneficial.

6.2 COMPONENT ABLATIONS

Component ablations. The controlled study uses GIFT-Eval with the official 97 configurations and the complete candidate-intervention evaluation. Every variant shares the same pre-test controller-training and validation protocol; random intervention selection averages five fixed seeds; and all gains are configuration-macro RelGain. Table 3 reports TimesFM-3, while the corresponding tables for the remaining forecasters appear in Appendix E (Table 9).

Table 3 separates the roles of evidence, candidate-space breadth, and selective execution. Removing three of the seven context-failure evidence scores (Noise, Local spike, and Level shift) retains positive gains on both metrics, with lower RelGain and higher Ranking Regret than Full CARE; intervention rate rises from 6.80% to 9.95%. This indicates that evidence dimensions shape intervention ranking and coverage, without guaranteeing monotonic accuracy gains as evidence is added. The half- and quarter-sized candidate-space sensitivity variants retain positive gains, while Ranking Regret rises to 0.286 and 0.550 as the candidate set shrinks. These variants filter the test-time candidates while keeping the 24-intervention-trained router and validation gate fixed, so they measure sensitivity to candidate removal rather than the performance of a retrained smaller-library system. Removing the risk gate applies interventions to every window and reverses the CRPS gain, whereas the gated policy remains selective and improves both metrics. Fixed, random, and single-intervention policies further show that broad or semantically chosen modification alone does not reproduce CARE's gains. The w/o cross-mechanism variant retains positive MASE and CRPS gains; Appendix E (Table 10) shows the same two-metric pattern across all eight family-wise removals.

Table 3: TimesFM-3 component ablations on GIFT-Eval. Ranking Regret measures the learned intervention ordering before risk gating and is reported for policies with a learned intervention ranking. Credal-TTA* is the reproduced GIFT-Eval baseline from Table 8.

Configuration	Explanation	Intervention rate	MASE RelGain↑	CRPS RelGain↑	Ranking Regret↓
Baselines					
Raw identity	No intervention	0.000%	+0.000%	+0.000%	–
Credal-TTA*	Truncated-context baseline	2.90%	-0.89%	-1.95%	–
CARE variants					
Full CARE	Evidence-conditioned gated policy	6.80%	+1.01%	+0.66%	0.144
Reduced evidence (remove 3/7)	Remove Noise, Local spike, and Level shift	9.95%	+0.88%	+0.59%	0.175
w/o risk gate	Execute every selected intervention	100.00%	+0.63%	-0.10%	0.144
Candidate-space and execution ablations					
Fixed global	Validation-selected global intervention	100.00%	-0.42%	-0.53%	–
Random	Uniform choice at matched rate	6.80%	-0.96%	-0.83%	–
RAR only	Always apply recent re-anchoring	100.00%	-66.40%	-74.98%	–
DN only	Always apply denoising	100.00%	-8.60%	-11.31%	–
1/2 candidate space	Filter to 12 at test time; full router/gate fixed	2.98%	+1.04%	+0.63%	0.286
1/4 candidate space	Filter to 6 at test time; full router/gate fixed	2.83%	+0.96%	+0.32%	0.550
w/o cross-mechanism	Exclude cross-mechanism interventions	5.79%	+0.91%	+0.50%	0.144

Table 4: Moirai-2 evidence-conditioned intervention profiles. Low/High denote evidence tertiles; Sel. share is the selection share among interventions; all utility values are joint normalized utility, and improvement rate is the fraction with positive joint normalized utility. Profile is a qualitative descriptor of the Low-to-High change.

Evidence	Intervention	Profile	Sel. share (%)	Median joint normalized utility (Low / High)	Improvement rate (Low / High)
Trend inconsistency	LM	Positive (relative)	11.1%	-0.93 / +0.06	31.5% / 50.8%
Stale history	RSM	Positive (relative)	0.2%	-1.22 / -0.10	28.9% / 31.3%
Noise	TEF	Ineffective	0%	-2.97 / -1.36	34.2% / 30.5%
Seasonal structure	RC	Reverse (relative)	13.1%	-1.02 / -6.38	38.4% / 21.6%

6.3 Evaluating Context-Failure Evidence

For Moirai-2, Figure 4(a) should be read as a conditional-utility map rather than an intervention-performance map. Each cell reports the High-minus-Low change in median joint normalized utility for one evidence–intervention pair, where Low and High are the bottom and top evidence tertiles. Large absolute values indicate strong state dependence, while near-zero values indicate weak separation along that evidence dimension; neither directly implies that an intervention is globally beneficial or ineffective.

The resulting landscape is sparse and heterogeneous. Denoising-related interventions show several of the largest state-conditioned shifts, demonstrating that the relative suitability of the same transformation changes across observable windows. Interventions within the same intervention mechanism group also exhibit markedly different patterns. Thus, mechanism groups describe what an intervention does, while learned utility determines when it should be used. Figure 4(b) and Table 4 show that these profiles vary across frozen TSFMs. A useful behavioral clue is that Local repair and endpoint-jump correction remain close to neutral across evidence levels, suggesting that TSFMs often respond more to the broader context distribution than to isolated recent anomalies; their apparent robustness to individual spikes helps explain why local repairs alone rarely yield a reliable gain.

Evidence can favor preservation over repair. Strong Seasonal structure provides a clear preservation signal: RC becomes markedly less suitable as Seasonal structure increases, producing one of the strongest negative contrasts in panel (a). When a context already contains periodic information, modifying it can remove evidence that the frozen forecaster exploits. Evidence-aware inference therefore identifies both what to change and what to preserve.

Finally, applicability crosses nominal mechanism categories. LFR becomes more favorable under Trend inconsistency and Level shift, but moves in the opposite direction under Noise, Local spike, and Endpoint instability. Similar cross-mechanism behavior appears for TEF. The Noise–TEF row is consistent with this interpretation: its ineffective profile is accompanied by a zero selection share, so the controller does not choose TEF for Noise evidence in this diagnostic. Together, these results show why the evidence taxonomy is an interpretable hypothesis space rather than a routing table: candidate mechanisms describe what an intervention does, whereas empirically learned, forecaster-specific utility determines when it should be used.

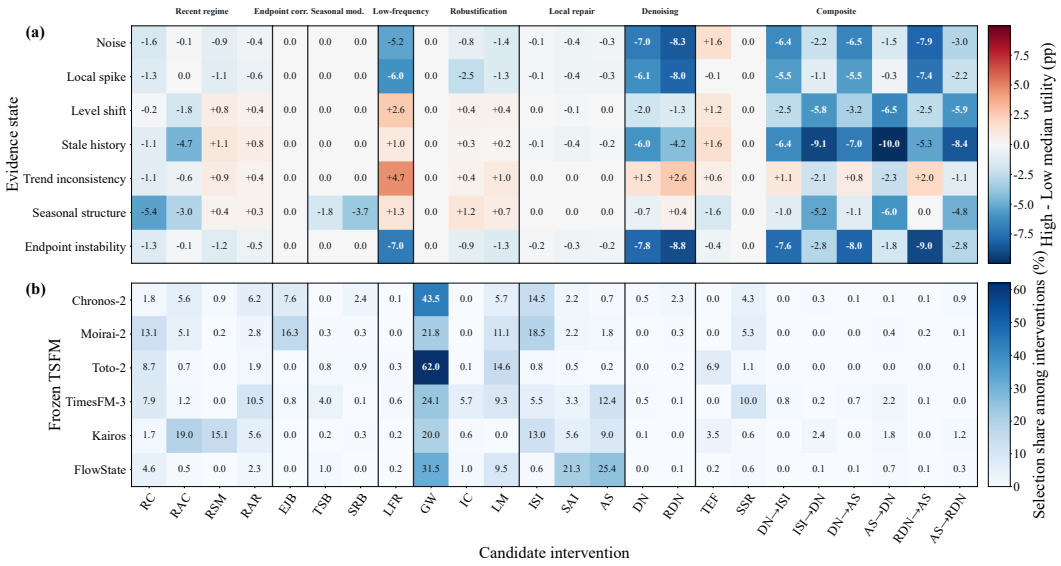


Figure 4: Evidence-conditioned intervention utility and selection. (a) For Moirai-2, each cell reports the High-minus-Low difference in median joint normalized utility for one evidence–intervention pair, where Low/High are the bottom/top evidence tertiles. Positive values mean that the intervention becomes relatively more suitable as the evidence score increases; negative values mean the opposite. Magnitude reflects conditional separation, not absolute intervention quality. (b) Among windows where CARE intervenes, the selection share of each candidate varies across frozen TSFMs. Together, the panels show a sparse, heterogeneous, and forecaster-dependent utility landscape rather than a fixed evidence-to-intervention mapping.

6.4 Selective Gating and Supplementary Diagnostics

Gate threshold sensitivity. Figure 5 for Kairos plots joint relative gain and joint improvement rate against intervention coverage as the mean predicted-gain threshold varies, with intervention ranking and the positive-gain count condition fixed. As coverage increases, relative gain rises and then levels off, while the joint improvement rate generally declines as lower-confidence interventions are admitted. The markers show the current-gate and no-gate operating points, making the gain–improvement trade-off explicit.

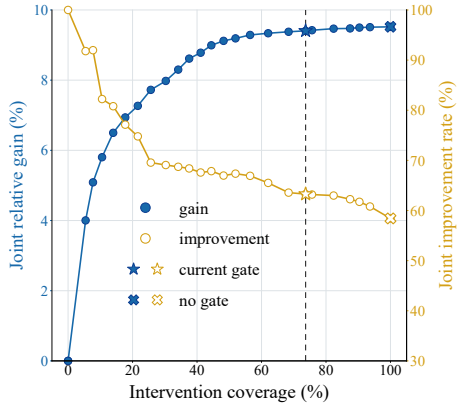


Figure 5: Coverage–risk trade-off for Kairos. Gain is the 97-configuration macro-average of overall policy relative gain; improvement is the macro-average fraction of intervened windows with positive joint utility.

Forecaster-specific utility. The source–target study in Appendix E shows positive same-model gains but mostly negative cross-model gains, supporting forecaster-specific controllers (Appendix E.7, Figure 6). The common GIFT-Eval protocol spans seven domains, providing heterogeneous-domain coverage, while explicit transfer is tested across TSFMs.

7 Conclusion

CARE pairs a frozen TSFM with an offline-trained controller that uses context-failure evidence to rank candidate interventions and gate context refinement. On GIFT-Eval, this selective policy improves MASE and CRPS for all six evaluated TSFMs, while intervention utility varies across instances and forecasters. These findings position context selection as a complementary axis for advancing TSFM inference.

Ethics Statement

The study uses time-series benchmark data and does not make clinical, financial, or safety decisions. Data licenses, privacy considerations, and any domain-specific risks will be documented in Appendix C and Appendix E.

Reproducibility Statement

Appendices B–E specify the complete evidence features and intervention library, data preprocessing and temporal boundaries, metric aggregation, frozen TSFM checkpoints and adapters, controller hyperparameters, validation-only gate selection, and all evaluation protocols. The anonymous code package provides the corresponding preprocessing, oracle construction, controller training, gate calibration, test replay, ablation, and figure/table generation scripts, together with machine-readable configurations and audit manifests. The controller pipeline uses the fixed seed 20260819 wherever randomness enters, while the random-intervention baseline uses five fixed seeds; raw and CARE forecasts for a given TSFM share the same checkpoint, adapter, test origins, target masks, and metric implementation. GIFT-Eval data and pretrained checkpoints are obtained from their public upstream sources and are not redistributed. Test targets are accessed only after the policy decision for evaluation or for analyses explicitly labeled as oracle, as detailed by the leakage protocol in Appendix C.

AI Use Statement

Generative AI tools were used for language editing, structural drafting, and code-assistance suggestions. The authors verified citations, claims, experiments, data handling, and the final manuscript and take responsibility for all submitted content.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. In *International Conference on Learning Representations*, 2025.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jens Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Abdul Fatir Ansari et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Luca Butera, Giovanni De Felice, Andrea Cini, and Cesare Alippi. Long context in time series forecasting. *arXiv preprint arXiv:2606.01999*, 2026.

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.

Yimin Du et al. Breaking context inertia: Adaptive context management for time series foundation models in non-stationary environments. *Neurocomputing*, 2026.

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.

Ayush Jain and Rajat Sen. Timesfm-3: A zero-shot foundation model for multivariate forecasting. Google Research Blog, 2026.

Dongbin Kim, Youngjoo Park, Woojin Jeong, and Jaewook Lee. Local geometry attention for time series forecasting under realistic corruptions. In *International Conference on Learning Representations*, 2026.

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*, 2024a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32369–32399, 2024b.

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.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Ege Onur Taga et al. Evolutionary feature engineering for structured data. *arXiv preprint arXiv:2607.01548*, 2026.

Dequan Wang et al. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

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.

Haixu Wu et al. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Table 5: Complete notation used in Sections 3–5.

Symbol	Definition
i	forecasting window or instance
$x_i \in \mathbb{R}^{L_i}, L_i$	raw context visible at the forecast origin and its observed length (capped at 512)
$y_i \in \mathbb{R}^{H_i}, H_i$	future target and forecast horizon; unavailable at inference
h_i, p_i, d_i	pre-origin history, frequency-dependent seasonal lag, and MASE scaling term
$\mathcal{V}_i$	valid, non-missing target positions in window i
f_θ, θ	frozen TSFM and its fixed parameters
$\widehat{P}, \widehat{P}_i^{\mathrm{raw}}$	model-agnostic point/quantile/sample output and raw forecast $f_\theta(x_i)$
$\mathcal{A}, a_0, a, a_i$	identity-plus-24 intervention set, identity, intervention index, and selected intervention
$\widehat{P}_i^a$	forecast $f_\theta(a(x_i))$ after intervention indexed by a
$\mathcal{M}, m$	$\{\mathrm{MASE}, \mathrm{CRPS}\}$ and one metric in this set
$L_{i,a,m}, \Delta_{i,a,m}$	intervention loss and realized absolute utility against identity
$u_{i,a,m}$	normalized controller target used for controller fitting
z_i, ψ	33-dimensional evidence state and deterministic state constructor
π_ϕ, ϕ	offline-trained intervention policy and its parameters
$\mathcal{C}_k, c(i)$	component-feature set for evidence score k and feature value for instance i
$q_{p,c}, u_c(i), e_{ik}$	identity-row percentile, normalized component value, and k-th evidence score
$\mathcal{L}_m$	forecasting loss associated with metric m
$\mathcal{Q}$	evaluated quantile levels $\{0.1, \ldots, 0.9\}$
$\widehat{y}_{i,t}^{(q)}, \widetilde{y}_{i,t}$	predicted q-quantile and median forecast at horizon position t
$L_{c,m}, \overline{L}_m, \epsilon$	configuration-level loss, unweighted configuration macro loss, and numerical-stability constant
$h_{\theta,m}, \widehat{u}_{i,a,m}, s_{i,a}$	metric-specific utility regressor, predicted normalized controller target, and ranking score
$a_i^\star, g_i, C_i$	top-ranked intervention, mean predicted utility, and positive-utility metric count
$(\tau_\theta, K_\theta), D_i, A_i$	validation-calibrated gate thresholds, intervention decision, and final intervention or identity

Appendix Reader Navigation

Appendix A fixes notation and the inference procedure. Appendix B details CARE's context-failure evidence state, seven evidence families, and the rationale for its candidate intervention space. Appendix C specifies datasets, tasks, metrics, and leakage control. Appendix D records frozen TSFMs, baselines, and experimental configurations. Appendix E collects supplementary experiments and results, including leave-one-intervention-family-out ablations, leave-one-domain-out evaluation, ranking diagnostics, and the complete intervention dictionary. Appendix F discusses limitations and future directions.

A Notation and Algorithms

Table 5 fixes the symbols used throughout the manuscript. The deployable pipeline first computes the raw forecast and z_i, scores all 24 non-identity interventions with the two frozen metric-specific regressors, and then applies the two-condition gate in Section 4.3.

B CARE Details

Interpretation and scope. Evidence scores are deterministic descriptors, not calibrated failure probabilities or ground-truth diagnoses. Mechanism-group labels support interpretation and are not controller inputs. CARE focuses on decision-level evidence selection through predicted utility comparison and selective intervention. Its controller uses benchmark-specific pre-test supervision, so backbone freezing does not make the full system zero-shot. Gate calibration provides empirical risk control rather than a formal probabilistic safety guarantee; coverage is an intervention frequency, not the prevalence of unreliable contexts. A zero High–Low median contrast does not establish equal conditional distributions or an ineffective intervention. Cross-model transfer must be assessed with source–target evaluation rather than inferred from independently fitted controllers. Oracle selection uses future targets solely for offline analysis.

Candidate intervention space construction. We organize recurring context problems into seven human-interpretable context-failure evidence families; multiple patterns may be present in the same window. For each family, we instantiate several deterministic transformations that could plausibly preserve useful evidence, suppress misleading evidence, or align older history with the current regime. This process yields 18 registered single-stage transformations and six ordered two-stage

Table 6: Composition of the 33-dimensional pre-prediction state.

Feature group	Dimensions	Included information
Evidence scores	7	Noise, Local spike, Level shift, Stale history, Trend inconsistency, Seasonal structure, Endpoint instability
Context scale	4	`ctx_base_scale`, `ctx_log_base_scale`, `ctx_value_scale`, `ctx_log_value_scale`
Raw-forecast distribution	14	Availability, interval width/change, dispersion, asymmetry, prefix/suffix width, median offset/slope, and tail ratios
Metadata	8	Horizon, relative context length, seasonal period, observed fraction, and term indicators
Total	**33**	**Prediction-time state dimensions**

compositions, for 24 non-identity candidates in total. TEF and SSR are registered single-stage transformations whose mechanisms span two families; for presentation they are grouped with the six compositions under Cross-mechanism. The mechanisms and parameter values are frozen before the final GIFT-97 test evaluation.

The resulting space is a hypothesis set rather than a hand-coded routing table. CARE never assumes that an intervention motivated by a given evidence family is beneficial for every matching window. Instead, the utility regressors learn instance- and forecaster-specific consequences for all candidates. Identity remains available at every forecast origin, and the risk gate rejects the top-ranked candidate when its predicted utility does not clear the validation-calibrated operating point. Table 1 gives the compact grouping, while Table 13 in Appendix E lists every identifier, mechanism, and composition order.

Evidence-state construction. The 33-dimensional state contains seven aggregated evidence scores, four context-scale features, fourteen raw-forecast distribution features, and eight metadata features, as summarized in Table 6. The seven context-failure evidence scores are deterministic summaries constructed from the 26 lower-level feature identifiers in Table 7; these identifiers support score construction but are not additional dimensions of z_i. For each lower-level feature c, the 5th and 95th percentiles $(q_{0.05,c}, q_{0.95,c})$ are fitted only on controller-training identity rows. After orienting the feature so that larger values indicate stronger evidence, CARE computes

$$u_c(i) = \mathrm{clip}\left(\frac{c(i) - q_{0.05,c}}{q_{0.95,c} - q_{0.05,c} + \epsilon}, 0, 1 \right).$$

For evidence family k, the corresponding state score is the mean of its normalized components:

$$e_{ik} = |\mathcal{C}_k|^{-1} \sum_{c \in \mathcal{C}_k} u_c(i), \qquad k \in \{1, \ldots, 7\}.$$

Thus, the state accounting is $7 + 4 + 14 + 8 = 33$, while the seven evidence scores are linked to 26 lower-level feature identifiers through the family mapping below.

Utility fitting and frozen gate. For each frozen TSFM, CARE fits two metric-specific multi-output `ExtraTreesRegressor` models, one for MASE and one for CRPS. Each model receives the same 33-dimensional state and returns 24 normalized utility predictions, one for each non-identity intervention. Rows with a non-finite target for a metric are excluded only from that metric's regressor. Five-fold `GroupKFold` by `series_key` produces out-of-fold predictions, ensuring that windows from a series are not scored by a regressor trained on that series. These out-of-fold predictions support selector comparison, ranking diagnostics, and gate calibration; the final metric-specific regressors are then refit on all eligible controller-training rows.

Validation freezes both the selector and the gate before official test evaluation. The selector is chosen from mean gain, maximin gain, and average rank. Mean gain averages the two predicted metric utilities, maximin gain favors the candidate with the strongest worst-metric prediction, and average rank averages the two metric-specific candidate ranks. The gate then searches predeclared predicted-utility thresholds and positive-metric-count requirements on the pre-test validation block. Its objective rewards joint validation gain while penalizing harmful operating profiles and enforcing validation coverage requirements. No test target, test oracle utility, or test action outcome participates in this selection. The selected non-identity action is executed only when the frozen gate accepts it; otherwise CARE returns identity and reuses the raw forecast.

Table 7: Lower-level features used to construct the seven context-failure evidence scores; they are not additional dimensions of the 33-dimensional state.

Evidence family	Feature identifier	Meaning
Endpoint instability	endpoint_outlier	Robust outlier score of the terminal observation.
Endpoint instability	endpoint_recent_jump	Magnitude of the latest jump relative to recent variability.
Endpoint instability	endpoint_increment_robust_z	Robust z-score of the terminal increment.
Endpoint instability	endpoint_prediction_gap	Gap between the first forecast value and the endpoint.
Seasonal structure	seasonal_autocorrelation	Strength of the selected seasonal-lag autocorrelation.
Seasonal structure	seasonal_period	Inferred seasonal period relative to the context length.
Seasonal structure	seasonal_tsf_strength	Seasonal-strength summary from the causal prefix-TSF output.
Level shift	level_recent_mean	Mean level over the recent context segment.
Level shift	level_last_minus_mean	Endpoint value minus the full-context mean.
Level shift	level_recent_jump	Recent mean shift relative to the preceding segment.
Trend inconsistency	trend_global_recent_gap	Difference between global and recent history slopes.
Trend inconsistency	trend_prediction_recent_gap	Difference between raw forecast slope and recent history slope.
Trend inconsistency	trend_prediction_global_gap	Difference between raw forecast slope and global history slope.
Stale history	stale_forecast_endpoint_gap	Raw forecast endpoint gap relative to the latest observation.
Stale history	stale_forecast_recent_level_gap	Raw forecast level gap relative to the recent-context mean.
Stale history	stale_distribution_offset	Offset of the raw predictive distribution from the recent level.
Local spike	spike_outlier_ratio	Fraction of globally flagged robust outliers.
Local spike	spike_local_outlier_ratio	Fraction of locally flagged outliers.
Local spike	spike_local_robust_z	Robust local z-score summary for spikes.
Local spike	spike_recent_outlier_ratio	Fraction of recent observations flagged as outliers.
Local spike	spike_anomaly_run	Length summary of consecutive anomalous runs.
Local spike	spike_increment_robust_z	Robust z-score summary of anomalous increments.
Noise	noise_roughness	First-difference roughness of the observed context.
Noise	noise_residual_std	Standard deviation of detrended residuals.
Noise	noise_high_frequency	High-frequency energy ratio in the context.
Noise	noise_tsf_frequency	High-frequency strength from the causal prefix-TSF summary.

C Datasets, Tasks, Metrics, and Leakage Control

C.1 Benchmark and Test Protocol

GIFT-Eval contains 97 logical configurations from 55 physical data sources: 55 short-, 21 medium-, and 21 long-horizon configurations across seven domains and heterogeneous frequencies. A physical source may therefore contribute more than one horizon configuration. Multivariate records are expanded into univariate channels, and configuration identifiers retain source, frequency, and horizon term. We preserve the official temporal test boundary and evaluate every official rolling test origin, yielding 371,330 test windows per TSFM; test origins are neither randomly sampled nor capped.

At each forecast origin, all methods receive at most the most recent 512 observed context values. Longer histories are truncated on the left, while shorter histories preserve their available observations and are left-padded with a separate observation mask. Visible-history missing values are filled by deterministic linear interpolation from observations available before the forecast origin; an entirely missing visible history receives a deterministic linear ramp from 0 to 1. Missing target positions remain missing and are excluded only through the evaluation mask. Interventions operate on observed context values and preserve the padding/observation mask. This common 512-context protocol gives raw CARE, refined CARE, and baselines identical inputs within the paper. It need not reproduce public leaderboard values obtained with model-specific context limits, so such public values are used only as external references rather than exact parity claims.

C.2 Pre-Test Supervision and Leakage Boundary

Controller supervision is constructed strictly before the official test boundary. A 10,000-window training oracle uses early/middle/late temporal stratification and configuration-balanced budgets; a 20,000-window validation oracle uses contiguous pre-test horizons. Each utility controller is fitted on the 10,000 training windows and the 12,926 older validation windows. The latest 7,074 validation windows are isolated for selector and outer-gate selection. The official test windows are accessed only after the controller, action aggregation rule, and gate have been frozen.

This design is benchmark-specific offline supervision rather than a zero-shot controller: pre-test GIFT outcomes train the external policy, although every TSFM checkpoint remains frozen. The oracle executes identity and all 24 non-identity candidates on pre-test windows to construct counterfactual utility labels. On official test windows, future labels are unavailable to the policy and are read only by the evaluator or by explicitly labeled offline oracle analyses.

All inference-visible features are computed only from the observed context, the raw forecast produced by the frozen TSFM, and metadata known at the forecast origin. Percentile normalization and imputation statistics are fitted on controller-training identity rows. Grouped out-of-fold prediction keeps every `series_key` out of the corresponding utility-model fit fold, preventing windows from the same univariate series from entering both sides of an OOF prediction. Temporal ordering additionally separates controller fitting, outer gate selection, and the official test boundary.

Test targets never enter evidence construction, utility prediction, candidate ranking, or gate decisions. They are read only after the decision for metric evaluation and explicitly labeled oracle analyses. Raw and policy branches use the same configurations, origins, target masks, TSFM adapter, checkpoint, and aggregation code. Feature schemas, candidate order, frozen selector/gate manifests, and random seeds are retained for auditability. The main protocol does not claim target-domain holdout because pre-test supervision spans the benchmark configurations; the balanced GIFT-21 leave-one-domain-out experiment in Appendix E separately tests that stronger transfer condition.

C.3 Metrics and Aggregation

For metric computation, let $\mathcal{V}_i$ be the valid target positions for window i, h_i its pre-origin history, p_i its frequency-dependent seasonal lag, and $\widetilde{y}_{i,t}$ the median forecast at horizon position t. The MASE scaling term d_i and per-window MASE M_i are

$$d_i = \mathrm{mean}_{t>p_i} |h_{i,t} - h_{i,t-p_i}|, \qquad M_i = \frac{\mathrm{mean}_{t\in\mathcal{V}_i} |y_{i,t} - \widetilde{y}_{i,t}|}{d_i}.$$

For quantile set $\mathcal{Q} = \{0.1, \ldots, 0.9\}$, $\widehat{y}_{i,t}^{(q)}$ denotes the prediction at quantile q and the doubled pinball loss is $\ell_q(y, \widehat{y}^{(q)}) = 2[q \max(y - \widehat{y}^{(q)}, 0) + (1-q)\max(\widehat{y}^{(q)} - y, 0)]$. For configuration c, the normalized CRPS approximation is

$$\mathrm{CRPS}_c = \frac{|\mathcal{Q}|^{-1} \sum_{q\in\mathcal{Q}} \sum_{i\in c} \sum_{t\in\mathcal{V}_i} \ell_q(y_{i,t}, \widehat{y}_{i,t}^{(q)})}{\sum_{i\in c} \sum_{t\in\mathcal{V}_i} |y_{i,t}|}.$$

For metric m, $L_{c,m}$ is the loss for configuration c, and the primary absolute score is its unweighted macro average $\overline{L}_m = 97^{-1} \sum_{c=1}^{97} L_{c,m}$. RelGain first compares raw and method losses within each configuration and then macro-averages, using ϵ for numerical stability:

$$\mathrm{RelGain}_m = \frac{1}{97} \sum_{c=1}^{97} \frac{L_{c,m}^{\mathrm{raw}} - L_{c,m}^{\mathrm{method}}}{\max(|L_{c,m}^{\mathrm{raw}}|, \epsilon)}.$$

The intervention rate is the fraction of test windows on which the policy selects a non-identity intervention. For RelGain, raw and CARE losses are compared within each configuration over all test windows; the percentage reductions are then macro-averaged over the 97 configurations. Positive RelGain indicates reduced error. Win/Tie/Loss counts use per-configuration MASE RelGain with a fixed $\pm 0.1\%$ tie interval and are reported in win/tie/loss order.

Table 8: Credal-TTA reproduction on GIFT-Eval across frozen TSFMs. Here λ is the validation-selected lower bound on the adaptive reset/truncation ratio threshold; larger values make intervention less frequent.

TSFM	MASE	CRPS	Intervention rate	λ
Chronos-2	1.6302 (-0.69%)	0.2258 (-1.22%)	4.26%	512
Moirai-2	1.7274 (-0.53%)	0.2401 (-0.50%)	2.90%	1024
Toto-2	1.7368 (-0.35%)	0.2407 (-0.79%)	2.90%	1024
TimesFM-3	1.5964 (-0.89%)	0.2266 (-1.95%)	2.90%	1024
Kairos	2.9181 (-5.74%)	0.4004 (-18.04%)	6.25%	256
FlowState	1.9281 (-20.50%)	0.2500 (-18.79%)	2.90%	1024

D Experimental Details

The six evaluated TSFMs remain frozen: Chronos-2, Moirai-2, Toto-2, TimesFM-3, Kairos, and FlowState. Their formal checkpoint variants are defined in Section 5.2. Local checkpoints are loaded through model-specific adapters, but raw and refined branches for a model use the same adapter, weights, quantile conversion, batch policy, context cap, test origins, and target mask. CARE changes only the historical context; no TSFM parameter is fine-tuned or updated online.

Raw TSFM evaluates the identity context. Seasonal Naive appears only in the oracle-potential analysis as a task-difficulty reference. Credal-TTA uses shift detection and adaptive context truncation under a common reproduction protocol. Fixed global selects one intervention on validation and applies it to every test window. Random intervention samples candidates at CARE-matched coverage and averages five fixed seeds. RC-only and DN-only apply one recent-state or denoising intervention globally. The no-gate variant always executes the top-ranked candidate. Candidate-pool and leave-one-family sensitivity variants retain the full trained router and frozen gate, filter the allowed candidates at test time, and rerank only the survivors; they do not retrain a smaller controller or recalibrate its gate.

Controller fitting uses 96-tree `ExtraTreesRegressor` models with maximum depth 16, minimum leaf size 12, feature fraction 0.75, bootstrap sampling, maximum sample fraction 0.85, eight CPU workers, and deterministic model/metric/fold seed offsets derived from seed 20260819. Grouped splitting uses `series_key`. The TimesFM-3 ablation compares Full CARE; a reduced-evidence variant that removes Noise, Local spike, and Level shift while retaining four evidence scores; no-gate execution; fixed, random, and single-intervention controls; and half- and quarter-size candidate-space sensitivity variants.

E Additional Results

This appendix contains the supplementary experiments and result tables: the Credal-TTA reproduction, leave-one-intervention-family-out analysis, leave-one-domain-out controller evaluation, the five-model GIFT-Eval ablation table, learned-ranking diagnostics, coverage–risk diagnostics, the forecaster-specificity figure, and the complete intervention dictionary. Each experiment is discussed around its displayed result rather than treated as an unanalysed data archive.

E.1 Restricted Adaptive-Context Baseline

The first supplementary experiment isolates a narrower intervention paradigm. Credal-TTA primarily changes retained history, so its comparison with CARE asks whether one context-management mechanism can stand in for a heterogeneous, instance-specific candidate intervention space. The table reports the reproduced GIFT-Eval values across all six frozen forecasters.

For Credal-TTA, λ is the validation-selected lower bound on the change-ratio threshold that triggers adaptive context reset/truncation. The implementation compares the current credal-set diameter with the preceding update and triggers when this ratio exceeds the effective threshold, which is the larger of λ and the burn-in mean plus three standard deviations. Larger λ values therefore suppress resets

Table 9: Supplementary component ablations on GIFT-Eval for the frozen TSFMs not shown in the main TimesFM-3 table. RelGain is the configuration-macro relative improvement over raw identity.

Configuration	Explanation	Intervention rate	MASE RelGain↑	CRPS RelGain↑	Ranking Regret↓
Chronos-2					
Baselines					
Raw identity	No intervention	0.00%	+0.000%	+0.000%	–
Credal-TTA*	Adaptive truncation	4.26%	-0.69%	-1.22%	–
CARE variants					
Full CARE	Full pipeline	14.038%	+0.790%	+0.177%	0.137
w/o context-failure evidence	Remove the seven context-failure evidence scores	15.631%	+0.633%	+0.003%	0.143
Fixed global	One global intervention	100.000%	-0.175%	-0.213%	–
Random intervention	Random at CARE coverage	14.038%	-2.082%	-2.480%	–
RC only	Always apply RC	100.000%	-35.292%	-42.619%	–
DN only	Always apply DN	100.000%	-7.834%	-9.271%	–
w/o risk gate	Always apply top-ranked intervention	100.000%	+0.630%	-0.626%	0.137
Moirai-2					
Baselines					
Raw identity	No intervention	0.00%	+0.000%	+0.000%	–
Credal-TTA*	Adaptive truncation	2.90%	-0.53%	-0.50%	–
CARE variants					
Full CARE	Full pipeline	23.266%	+1.294%	+1.055%	0.144
w/o context-failure evidence	Remove the seven context-failure evidence scores	8.185%	+0.805%	+0.536%	0.140
Fixed global	One global intervention	100.000%	-0.441%	-0.632%	–
Random intervention	Random at CARE coverage	23.266%	-2.228%	-3.098%	–
RC only	Always apply RC	100.000%	-32.083%	-32.952%	–
DN only	Always apply DN	100.000%	-5.845%	-8.017%	–
w/o risk gate	Always apply top-ranked intervention	100.000%	+0.844%	+0.616%	0.144
Toto-2					
Baselines					
Raw identity	No intervention	0.00%	+0.000%	+0.000%	–
Credal-TTA*	Adaptive truncation	2.90%	-0.35%	-0.79%	–
CARE variants					
Full CARE	Full pipeline	5.663%	+0.208%	+0.085%	0.172
w/o context-failure evidence	Remove the seven context-failure evidence scores	4.332%	-0.007%	-0.070%	0.177
Fixed global	One global intervention	100.000%	-0.692%	-0.691%	–
Random intervention	Random at CARE coverage	5.663%	-0.538%	-1.013%	–
RC only	Always apply RC	100.000%	-25.567%	-50.874%	–
DN only	Always apply DN	100.000%	-5.199%	-6.640%	–
w/o risk gate	Always apply top-ranked intervention	100.000%	+0.232%	-2.744%	0.172
Kairos					
Baselines					
Raw identity	No intervention	0.00%	+0.000%	+0.000%	–
Credal-TTA*	Adaptive truncation	6.25%	-5.74%	-18.04%	–
CARE variants					
Full CARE	Full pipeline	73.408%	+10.045%	+8.995%	0.150
w/o context-failure evidence	Remove the seven context-failure evidence scores	69.712%	+8.432%	+7.418%	0.155
Fixed global	One global intervention	100.000%	+0.546%	+1.045%	–
Random intervention	Random at CARE coverage	73.408%	+0.289%	+0.552%	–
RC only	Always apply RC	100.000%	-8.301%	-5.834%	–
DN only	Always apply DN	100.000%	-0.006%	-0.190%	–
w/o risk gate	Always apply top-ranked intervention	100.000%	+9.312%	+8.278%	0.150
FlowState					
Baselines					
Raw identity	No intervention	0.00%	+0.000%	+0.000%	–
Credal-TTA*	Adaptive truncation	2.90%	-20.50%	-18.79%	–
CARE variants					
Full CARE	Full pipeline	6.681%	+0.624%	+0.662%	0.128
w/o context-failure evidence	Remove the seven context-failure evidence scores	13.861%	+0.502%	+0.379%	0.133
Fixed global	One global intervention	100.000%	-0.820%	-0.451%	–
Random intervention	Random at CARE coverage	6.681%	-0.441%	-0.520%	–
RC only	Always apply RC	100.000%	-22.450%	-23.837%	–
DN only	Always apply DN	100.000%	-5.890%	-7.930%	–
w/o risk gate	Always apply top-ranked intervention	100.000%	+0.713%	-1.212%	0.128

and reduce intervention frequency. The searched values and selected per-model operating points are fixed before official test evaluation.

Credal-TTA provides a restricted intervention paradigm for comparison: it primarily changes the amount of retained history and does not expose the heterogeneous candidate intervention space used by CARE. Its per-model MASE, CRPS, and coverage vary, especially for Kairos and FlowState, so it provides a mechanism boundary rather than a universal baseline ranking. This spread is useful for assessing whether adaptive context length alone can replace instance-specific utility prediction and gating.

E.2 Supplementary Component Ablations

The second supplementary experiment asks whether the component pattern observed for TimesFM-3 remains visible for the other five frozen forecasters under the same GIFT-Eval protocol. Read the RelGain columns together with intervention rate: a positive gain at very high coverage has a different operating meaning from a smaller gain obtained through selective execution. Ranking Regret is reported only where a learned intervention ordering is available.

Table 10: TimesFM-3 leave-one-intervention-family-out results on GIFT-Eval. RelGain is the configuration-macro relative improvement over raw identity.

Configuration	Explanation	Intervention rate	MASE RelGain↑	CRPS RelGain↑	Ranking Regret↓
w/o recent regime	Exclude recent-regime actions	5.79%	+1.02%	+0.59%	0.145
w/o endpoint	Exclude endpoint-correction actions	6.27%	+0.94%	+0.54%	0.149
w/o seasonal	Exclude seasonal-modification actions	6.09%	+0.94%	+0.55%	0.144
w/o low frequency	Exclude low-frequency-reconstruction actions	6.30%	+0.93%	+0.54%	0.144
w/o robustification	Exclude robustification actions	5.18%	+0.94%	+0.56%	0.143
w/o local repair	Exclude local-repair actions	5.18%	+1.00%	+0.63%	0.142
w/o denoising	Exclude denoising actions	6.31%	+0.94%	+0.54%	0.144
w/o cross-mechanism	Exclude cross-mechanism actions	5.79%	+0.91%	+0.50%	0.144

Across these five forecasters, Full CARE is the only listed policy that achieves positive configuration-macro gain on both metrics for every model. The magnitude and coverage remain strongly model-specific: Full CARE intervenes on 5.66% of Toto-2 windows, 6.68% of FlowState windows, 14.04% of Chronos-2 windows, 23.27% of Moirai-2 windows, and 73.41% of Kairos windows. Fixed global and single-intervention policies do not retain consistent two-metric gains across forecasters. Without the gate, only Moirai-2 and Kairos remain positive on both metrics; Toto-2 retains a small MASE gain but incurs a CRPS loss. These variations reinforce the need for a forecaster-specific, selective policy rather than a universal context repair.

The supplementary ablation therefore answers a narrower question than the main table: it tests whether the component pattern survives a change of frozen forecaster under the same full GIFT-97 protocol. The answer is positive for the full policy's two-metric direction, but coverage and magnitude are not invariant. In particular, the large Kairos gain is accompanied by high coverage, whereas the smaller Toto-2 and FlowState gains are obtained with sparse intervention. This is evidence for model-conditioned utility and selective execution, not evidence that one intervention or one coverage level is optimal for every TSFM.

E.3 Leave-One-Intervention-Family-Out Analysis

To test whether the main result depends on one intervention family, we remove each family from the allowed test-time candidate set in turn, rerank the surviving candidates, and retain the Full CARE router and validation-calibrated gate. These are frozen-policy candidate-sensitivity evaluations; the controller is not retrained and the gate is not recalibrated. The table reports configuration-macro RelGain, intervention rate, and Ranking Regret.

All eight leave-one-family variants retain positive MASE and CRPS gains, with intervention rates between 5.18% and 6.31%. The relatively narrow gain range suggests that the full policy's improvement is not driven by any single intervention family; the cross-mechanism ablation remains positive as well.

E.4 Leave-One-Domain-Out Controller Evaluation

We assess cross-domain generalization on the balanced GIFT-21 subset, which contains three configurations from each of seven domains. Each fold holds out one complete domain (three configurations) for test and uses the remaining six domains (18 configurations) for controller development and validation-based gate selection. This produces seven independently fitted controllers and frozen gates per TSFM. Table 11 compares each held-out-domain result with the standard Full CARE controller evaluated on exactly the same three configurations and test windows. No held-out-domain outcome is used to fit or select the LOO controller or its gate; Full CARE retains its standard pre-test supervision from all 21 configurations.

The comparison shows that removing target-domain controller supervision changes both coverage and utility rather than inducing a uniform degradation. Kairos retains sizable positive transfer on most held-out domains but loses MASE on Climate; Chronos-2 transfers positively to Healthcare while incurring losses on Energy; and several low-coverage models abstain on entire domains. The domain-LOO protocol therefore exposes model- and domain-specific transfer behavior that is hidden by a six-model macro average. It supports transferable utility in selected cases but not a claim of uniformly positive cross-domain generalization.

Table 11: Per-domain comparison of standard Full CARE (F) and leave-one-domain-out CARE (L) on balanced GIFT-21. Each cell reports F/L. Metrics match the component-ablation tables: RelGain is configuration-macro relative improvement on the indicated domain, and Ranking Regret is pre-gate regret under joint normalized MASE/CRPS utility.

Test domain	Intervention rate F/L (%)	MASE RelGain F/L (%)	CRPS RelGain F/L (%)	Ranking Regret F/L
Chronos-2				
Climate	1.374/0.000	+0.3347/+0.0000	+0.0267/+0.0000	0.106/0.112
Economics	0.000/0.500	+0.0000/-0.0020	+0.0000/+0.0313	0.195/0.210
Energy	0.000/41.247	+0.0000/-0.1770	+0.0000/-0.3108	0.166/0.166
Healthcare	0.000/31.336	+0.0000/+0.2538	+0.0000/+0.0433	0.178/0.193
Sales	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.088/0.077
Transport	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.084/0.085
Web/CloudOps	3.646/0.000	-0.1400/+0.0000	-0.5582/+0.0000	0.262/0.280
Moirai-2				
Climate	0.824/3.022	+0.0087/-0.0078	+0.0017/-0.0017	0.142/0.174
Economics	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.206/0.220
Energy	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.192/0.190
Healthcare	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.121/0.126
Sales	4.427/0.025	+0.0485/+0.0000	-0.1689/+0.0000	0.073/0.074
Transport	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.074/0.073
Web/CloudOps	3.394/2.816	+0.0218/-0.0112	+0.0018/+0.0001	0.253/0.254
Toto-2				
Climate	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.121/0.113
Economics	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.184/0.200
Energy	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.195/0.195
Healthcare	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.125/0.130
Sales	0.000/5.414	+0.0000/+0.0432	+0.0000/+0.0201	0.098/0.103
Transport	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.064/0.067
Web/CloudOps	0.036/4.946	-0.0010/-0.0114	-0.0042/-0.0006	0.209/0.222
TimesFM-3				
Climate	0.275/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.169/0.152
Economics	0.040/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.202/0.282
Energy	11.511/0.000	+0.0409/+0.0000	+0.0362/+0.0000	0.177/0.178
Healthcare	0.125/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.116/0.119
Sales	0.051/0.025	+0.0000/+0.0004	+0.0000/+0.0000	0.090/0.087
Transport	0.095/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.051/0.051
Web/CloudOps	1.733/0.000	-0.6195/+0.0000	-0.0425/+0.0000	0.243/1.897
Kairos				
Climate	100.000/95.192	+1.7568/-1.4180	+6.6825/+0.8367	0.188/0.360
Economics	99.460/99.920	+16.5726/+0.1339	+8.7333/+0.6120	0.219/0.309
Energy	95.204/82.494	-0.8008/+0.9921	+1.6497/+1.4152	0.194/0.191
Healthcare	100.000/94.881	+30.1196/+14.2850	+32.8151/+17.0232	0.106/0.247
Sales	99.975/84.113	+3.4382/+2.4607	+2.7132/+2.4385	0.134/0.141
Transport	91.698/35.019	+7.1484/+4.4790	+9.0386/+6.0459	0.092/0.113
Web/CloudOps	98.123/95.054	+11.9786/+11.0187	+10.2599/+8.0682	0.272/0.355
FlowState				
Climate	1.236/2.198	+0.7410/+0.0876	+0.0345/+0.0418	0.129/0.140
Economics	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.195/0.197
Energy	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.124/0.121
Healthcare	3.995/0.250	+0.0012/+0.0000	+0.0009/+0.0000	0.127/0.139
Sales	8.222/0.607	+0.0277/+0.0024	-0.1388/+0.0011	0.111/0.099
Transport	0.000/0.000	+0.0000/+0.0000	+0.0000/+0.0000	0.059/0.065
Web/CloudOps	0.505/0.000	+0.0026/+0.0000	+0.0014/+0.0000	0.223/0.249

E.5 Decision Quality

The third experiment evaluates intervention ranking before risk gating on GIFT-Eval. Beneficial@K, Near-opt., Spearman, and Regret should be read jointly: the first three measure whether the ordering contains useful signal, while regret records the remaining distance from the oracle-selected intervention.

The ranking diagnostics show that the learned ordering contains signal before risk gating: all six forecasters have positive Spearman correlation, and Kairos has the strongest Beneficial@1 and near-optimal accuracy among the listed models. The nonzero regret values show why ranking quality cannot by itself guarantee a safe intervention; the gate remains necessary to abstain when predicted utility is insufficient. These metrics are evaluated on GIFT-Eval and should therefore be read as diagnostic evidence for intervention ranking within the reported benchmark protocol.

Table 12: Pre-gate intervention-ranking quality on GIFT-Eval under joint normalized utility. B@K (Beneficial@K) is the percentage of windows for which at least one of the top-K predicted interventions has positive realized absolute utility; Near-opt. is the percentage whose top-ranked joint normalized utility is within $\delta = 0.01$ of the oracle; Spearman is the rank correlation across the 24 candidate interventions; Regret is oracle joint normalized utility minus top-ranked joint normalized utility.

TSFM	B@1	B@3	Near-opt.	Spearman	Regret
Chronos-2	23.79	59.76	13.75	0.305	0.1584
Moirai-2	23.08	62.44	12.70	0.281	0.1774
Toto-2	30.14	62.48	13.47	0.249	0.1478
TimesFM-3	23.45	58.41	12.73	0.311	0.1512
Kairos	58.34	76.90	18.90	0.239	0.1746
FlowState	28.66	63.78	13.78	0.280	0.1335

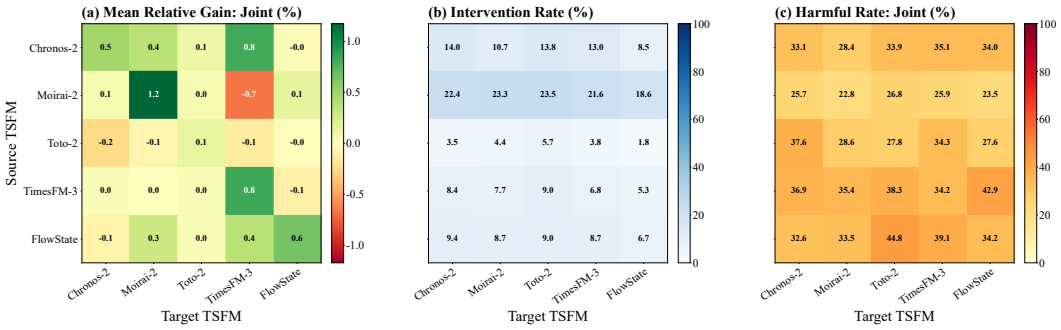


Figure 6: Forecaster specificity across five TSFMs. Rows are source routers and columns are target forecasters. The panels report (a) joint mean relative gain, (b) intervention rate, and (c) joint harmful-intervention rate. Diagonal cells are same-forecaster references; off-diagonal cells apply a source router to a different target TSFM.

E.6 Diagnostic Evaluation

Ranking diagnostics report Beneficial@K, near-optimal accuracy with $\delta = 0.01$, average regret, and Spearman correlation. The positive correlations and nonzero regrets in Table 12 show that utility ranking is informative but imperfect, which motivates the separate risk gate. Coverage–risk curves vary only τ_θ while keeping the selected intervention and other gate conditions fixed; their purpose is to expose the operating trade-off rather than to select a new test-time threshold. Evidence–mechanism plots fit tertile thresholds on training identity rows and retain identity windows in the denominator, so selection association is not interpreted as causal evidence that a mechanism will improve a forecast.

E.7 Forecaster Specificity Study

The final supplementary experiment asks whether an intervention-utility controller can be reused across frozen forecasters. The figure compares five source routers with five target TSFMs; rows identify the source router and columns the target forecaster, while diagonal cells are same-model references. The diagonal is the relevant within-forecaster control, and the off-diagonal cells test source–target dependence.

The diagonal and off-diagonal contrast is the key diagnostic: a controller trained for the same forecaster retains positive utility, whereas reusing that controller across forecasters is unreliable. The result does not identify a single cause for each off-diagonal failure, but it rules out treating the intervention-utility model as forecaster-agnostic. In deployment terms, the controller and gate must remain paired with the frozen backbone whose forecast defines the utility labels. This is a forecaster-specificity study, not evidence of general transfer capability.

E.8 Supplementary interpretation of the evidence–intervention landscape

For Moirai-2, Figure 4(a) contains several useful patterns beyond the main sparse-landscape summary. The strongest contrasts are concentrated in a small subset of evidence–intervention pairs, while many cells remain close to zero. This concentration is consistent with evidence scores acting as selectively informative cues rather than interchangeable measures of global context reliability.

LFR provides a clear cross-semantic example. Its High-minus-Low contrast is favorable under Trend inconsistency and Level shift, but reverses under Noise, Local spike, and Endpoint instability. The same intervention therefore responds to different context mismatches in different directions. Denoising and cross-mechanism denoising interventions also show several of the largest conditional shifts in the map, while TEF exhibits useful behavior across more than one nominal mechanism category. These patterns reinforce the distinction between an intervention's mechanism, which describes what it does, and its learned utility, which determines when it is suitable.

The near-zero Local repair and EJB contrasts require a narrower interpretation: they indicate weak separation between evidence tertiles for those candidates, not zero intervention utility. Consistently, removing the Local repair family leaves the aggregate CARE gain close to the full policy in the leave-one-family-out analysis (Table 10). On this benchmark, isolated-point corrections therefore contribute less distinctive evidence-conditioned structure than broader context transformations, a pattern consistent with frozen TSFMs being relatively insensitive to individual spikes or endpoint anomalies. The heatmap alone does not support a claim that these interventions are universally ineffective.

Intervention dictionary and highlighted pairs. The complete 24-intervention dictionary and all 7×24 values are provided below and in the machine-readable supplement. The order is fixed as five mechanism groups (recent-regime adaptation, trend/endpoint correction, seasonal modification, low-frequency reconstruction, robustification), then Local repair and Denoising, followed by all eight cross-mechanism interventions. Cross-mechanism intervention abbreviations contain no hyphens. These are empirical utility patterns, not hand-coded routing rules.

F Limitations and Future Directions

Evidence-aware context management opens a complementary direction for TSFM improvement. CARE formulates context reliability as a measurable decision problem around a frozen forecaster, with intervention utility and selective execution as its operational objectives. Richer context-failure evidence and candidate intervention spaces can extend this context policy.

F.1 Context-failure evidence and interaction structure

The current state uses seven deterministic evidence summaries and the main analysis studies the conditional effect of one evidence dimension at a time. This choice provides an interpretable first map of the evidence–intervention landscape, but it does not exhaust the interactions among multiple evidence patterns, raw-forecast responses, metadata, and intervention sequences. Future work can learn richer relational or set-based representations, model higher-order evidence interactions, and evaluate whether compositional states improve utility prediction without sacrificing the target-free inference boundary.

F.2 Candidate intervention space and forecaster specificity

The 24 interventions are deliberately constructed from the seven evidence families to provide a broad and auditable candidate space. Their utility is sparse and depends on both the data regime and the forecasting preferences of the frozen backbone, which makes intervention discovery and universal routing difficult. The observed forecaster specificity indicates that a router can retain useful structure yet lose performance when transferred across backbones; a universally effective context optimizer is therefore not assumed by the present framework. Promising directions include data-driven intervention discovery, compositional intervention search, shared utility representations with forecaster-specific calibration, and transfer protocols that separate reusable evidence structure from backbone-dependent preferences.

Table 13: Complete dictionary of the 24 non-identity candidate interventions.

Action ID	Abbrev.	Mechanism	Transformation
recent_context_064	RC	Recent-regime	retain latest 64 observations
recent_amplitude_compress_075	RAC	Recent-regime	compress recent amplitude around endpoint
recent_scale_match_064	RSM	Recent-regime	match older level and scale to latest 64
recent_reanchor_100	RAR	Recent-regime	fully re-anchor older history to recent state
endpoint_jump_blend_050	EJB	Endpoint correction	half-strength terminal-jump pullback
tail_seasonal_blend_050	TSB	Seasonal	seasonal blend on latest 96 points
seasonal_residual_blend_050	SRB	Seasonal	seasonal blend of detrended residual
low_frequency_reconstruct_050	LFR	Low-frequency	half-strength Fourier reconstruction blend
global_winsorize_4	GW	Robustification	robust four-scale global winsorization
increment_clip_4	IC	Robustification	clip anomalous increments and preserve endpoints
local_mean_025	LM	Robustification	mix with local mean at weight 0.25
isolated_spike_interp	ISI	Local repair	interpolate length-one internal anomalies
short_anomaly_interp_3	SAI	Local repair	interpolate anomaly runs of length at most three
adaptive_spike_blend_100	AS	Local repair	replace flagged local spikes at full strength
denoise_100	DN	Denoising	full-strength five-point Gaussian smoothing
residual_denoise_100	RDN	Denoising	fully smooth residuals, preserve linear trend
trend_endpoint_flatten_000	TEF	Trend+endpoint	remove fitted trend and anchor endpoint
seasonal_spike_repair_100	SSR	Seasonal+local	seasonal-lag repair of flagged spikes
denoise100_then_isolated_interp	DN→ISI	Cross-mechanism	apply DN then ISI
isolated_interp_then_denoise100	ISI→DN	Cross-mechanism	apply ISI then DN
denoise100_then_adaptive_spike100	DN→AS	Cross-mechanism	apply DN then AS
adaptive_spike100_then_denoise100	AS→DN	Cross-mechanism	apply AS then DN
residual_denoise100_then_adaptive_spike100	RDN→AS	Cross-mechanism	apply RDN then AS
adaptive_spike100_then_residual_denoise100	AS→RDN	Cross-mechanism	apply AS then RDN

F.3 Richer evidence interfaces

CARE currently learns a context intervention policy through feature-based utility prediction and selective gating. Future work could add explanations grounded in predicted intervention consequences, helping users understand which observations are preserved or modified while retaining quantitative utility and risk checks.

F.4 Beyond time-series forecasting

The same perspective may apply wherever the quality of input evidence directly affects a foundation model's downstream decision. Retrieval-augmented generation could select or discount retrieved evidence before answer generation; scientific discovery systems could assess which observations or sources are reliable before hypothesis scoring; and sequential decision-making systems could gate interventions according to the expected value and risk of their available state information. These settings provide opportunities to test whether evidence-aware inference is a broader design principle for foundation models.