S2G-HAMC: FORECAST-CALIBRATED CONTEXT ADAPTATION FOR FROZEN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) can forecast without task-specific retraining. However, every frozen forecaster still depends on the numerical context it receives. Deployment observations may not align with all coordinates in that context. The system must therefore map timestamped observations to a canonical context. We introduce **S2G-HAMC**, a forecast-calibrated context-adaptation framework for frozen TSFMs. Its **S2G-Adapter** builds a support-aware canonical context. It learns two calibration scalars on synthetic signals by using forecast loss from the frozen forecaster. It does not fit real benchmark targets. **HAMC** is a fixed robustness module. It evaluates nearby S2G context views and combines their forecasts across the prediction horizon. It does not change the backbone or add trainable parameters. S2G-HAMC improves 35/36 backbone–dataset pairs across 12 real-world datasets and three frozen forecasting families. Its mean Weighted Quantile Loss (WQL) gains are 7.62–8.60% over the corresponding validation-selected Linear interface. S2G alone yields 7.39–8.24% mean gains, and the complete system has a larger mean gain than S2G alone on all three backbones. The S2G-HAMC gain also increases as observation retention falls from 0.40 to 0.10. At the sparsest level, the gains reach 9.61–15.15%. With all adaptation choices fixed, the framework also improves quantile loss in 22/24 cells across three previously unused domains. These results show that S2G-HAMC can adapt frozen TSFMs through their numerical context.

1 INTRODUCTION

Time-series foundation models (TSFMs) such as Chronos and TimesFM can forecast without task-specific backbone training (Ansari et al., 2024; Das et al., 2024). However, every frozen forecaster still receives a numerical context. Deployment observations may not align with all coordinates in that context. The system therefore needs a rule that maps observed timestamps to model context. Most pipelines treat this rule as preprocessing. We ask whether this mapping can serve as an adaptation interface while the forecasting model remains frozen.

Recent work adapts time-series systems in several other ways. FlexTSF changes the predictor so that it can model variable temporal regularity directly (Xiao et al., 2024). TATO searches transformation pipelines around a frozen forecasting model (Qiu et al., 2026). Other methods change feature spaces, add covariates, use online feedback, or correct model outputs (Benechehab et al., 2025; Dange & Sarawagi, 2025; Lee et al., 2025; Liang et al., 2026; Im & Kwon, 2026). These methods do not focus on the observation-to-context map. A frozen backbone that expects a canonical numerical context still needs such a map. We study this map as the object of forecast-calibrated adaptation.

S2G-HAMC uses forecast loss to guide context construction. S2G-ADAPTER maps timestamped observations to the canonical grid with a degree-one local projection. It uses local observation support to choose the projection scale. We learn its two calibration scalars on synthetic signals through the pinball loss of the frozen forecaster. We do not fit real benchmark targets. This step produces a forecast-calibrated context without changing the backbone. HAMC is the framework's fixed robustness module. It evaluates nearby S2G scales through the same frozen forecaster and combines their forecasts with horizon-dependent weights. This design reduces dependence on one

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.

calibrated context scale. A fixed local-polynomial context view adds another forecast to the final composition. The full S2G-HAMC method changes the context interface and forecast composition. It does not change the TSFM parameters.

We test S2G-HAMC on 12 public datasets and three frozen forecasting families. The complete system reaches mean WQL gains of 8.58%, 8.60%, and 7.62% on Chronos-2, TimesFM-3, and TiRex-2. These gains are measured against the corresponding validation-selected Linear interface. The system improves 35/36 backbone–dataset pairs. S2G alone yields 8.11%, 8.24%, and 7.39%, and the complete system has a larger mean gain than S2G alone on all three backbones. The S2G-HAMC gain also increases as observation support becomes sparse. At $r = 0.10$, the mean gains reach 14.74%, 9.61%, and 15.15%. In a held-out three-domain test, S2G-HAMC improves quantile loss in 22/24 domain–support-pattern–severity cells. In a broader seven-system comparison, Chronos-2 + S2G-HAMC has the lowest mean rank (2.67) and the best WQL on 6/12 datasets.

Our contributions are:

- **Problem formulation.** We treat the observation-to-context map as the adaptation target for a frozen TSFM. The forecasting backbone stays fixed.
- **Method.** We introduce S2G-HAMC . S2G-ADAPTER builds a support-aware, forecast-calibrated context. HAMC is a fixed robustness module that combines forecasts from nearby context views across the prediction horizon. HAMC adds no trainable parameters.
- **Evidence.** The complete system reaches 7.62–8.60% mean WQL gains over the corresponding validation-selected Linear interface and improves 35/36 backbone–dataset pairs. The gains increase as observation support becomes sparse. The frozen method also transfers to previously unused domains.

2 RELATED WORK

2.1 TSFMS AND TEMPORAL INPUT STRUCTURE

Modern TSFMs are pretrained across many domains and frequencies (Ansari et al., 2024; Das et al., 2024; Liu et al., 2025). Their inference interfaces still define the structure of the context given to the model. Some work changes the forecasting architecture to handle variable temporal regularity directly. FlexTSF, for example, uses continuous-time patching for this purpose (Xiao et al., 2024). Our setting keeps the forecasting architecture fixed. We study the map from timestamped observations to the canonical numerical context used by an existing TSFM.

2.2 ADAPTING FROZEN FORECASTERS

Several methods adapt pretrained forecasters without full backbone retraining. AdaPTS learns feature-space adapters, and TFMAdapter adds covariate-aware refinement (Benechehab et al., 2025; Dange & Sarawagi, 2025). TATO searches context slicing, normalization, and outlier-correction pipelines around a frozen time-series model (Qiu et al., 2026). Online methods use new feedback to update small modules (Lee et al., 2025). Other work adapts predictions directly. δ-Adapter adjusts model inputs and outputs, and COSA applies a context-aware output correction at test time (Liang et al., 2026; Im & Kwon, 2026). These methods show that adaptation can happen outside the backbone. S2G-HAMC studies a different interface. S2G maps observed temporal support to the canonical context used by the frozen model. HAMC then combines forecasts from nearby S2G context views.

2.3 FORECAST-ORIENTED CONTEXT CONSTRUCTION

Methods for irregularly sampled or missing-value time series often learn models that handle incomplete temporal inputs directly (Fang et al., 2024; Peng et al., 2025; Marisca et al., 2024; Hu et al., 2025). Our setting keeps an already-pretrained forecaster fixed. We treat the observation-to-context map as the adaptation target. Prior work also shows that downstream task value can differ from representation fidelity (Wang et al., 2024). This result motivates our forecast-oriented view of context construction. We optimize the context map with the forecaster's own loss instead of a pointwise rep-

resentation objective. The map uses degree-one local-polynomial projection (Fan & Gijbels, 1996). It serves as a fixed context operator, not as a second forecasting model.

3 METHOD

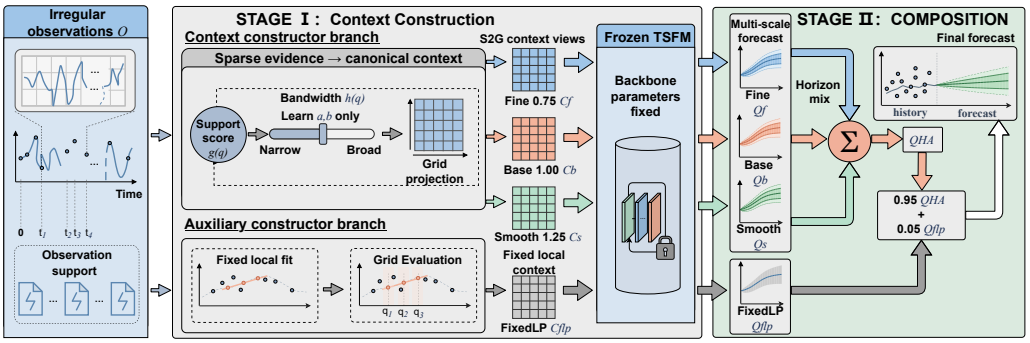

Figure 1: **S2G-HAMC as a context-adaptation chain.** Timestamped observations are mapped to canonical contexts by S2G. The same frozen TSFM evaluates nearby fixed S2G scales. HAMC then combines the resulting forecasts by horizon and adds a complementary forecast from a fixed local-polynomial context map. Only the two S2G calibration scalars are learned; the TSFM remains frozen.

Figure 1 shows the full adaptation path. S2G first maps observations to the model context. It then conditions the map on local observation support. Forecast loss calibrates the two S2G scalars. HAMC finally combines forecasts from nearby context views across the prediction horizon.

3.1 FROM TIMESTAMPED OBSERVATIONS TO A FORECAST-CALIBRATED CONTEXT

Let F_{θ_0} be a pretrained probabilistic forecaster whose parameters remain fixed. Let $\mathcal{O} = \{(t_i, y_i)\}_{i=1}^{N}$ denote the timestamped observations available before the forecast origin. The forecaster consumes a canonical context $\mathbf{C}$, so we introduce a context map G_ϕ such that

$$\widehat{\mathbf{Q}} = F_{\theta_0}(G_\phi(\mathcal{O})), \qquad \theta_0 \text{ fixed.} \tag{1}$$

We choose the map by downstream forecast risk:

$$\phi^\star = \arg\min_{\phi \in \Phi} \mathbb{E}_{(\mathcal{O},Y)}[\ell_{\text{fcst}}(F_{\theta_0}(G_\phi(\mathcal{O})), Y)]. \tag{2}$$

This objective makes the context map the adaptation variable. The backbone stays fixed. S2G uses only two calibration scalars. We fix the projection degree, support-score temperature, clipping limits, ridge form, and fallback rule. Equation 2 defines the forecast objective. The next step makes the map respond to the observation support around each canonical query.

3.2 OBSERVATION-SUPPORT-AWARE CONTEXT MAPPING

Let the canonical context contain C coordinates. We express each observed timestamp relative to the fixed forecast origin,

$$\tau_i = \frac{t_i - \text{context_end}}{\text{canonical_step}},$$

so the model queries are $q \in \{-(C-1), \ldots, 0\}$. For each query, S2G-ADAPTER summarizes nearby observation support with

$$g(q) = -\tau_T \log \sum_{i=1}^{N} \exp\left(-\frac{|\tau_i - q|}{\tau_T}\right), \qquad \tau_T = 0.25. \tag{3}$$

We clip $g(q)$ to $[0, 50]$. The score depends on both the distance to nearby observations and their local multiplicity. Smaller values indicate stronger nearby support.

The adapter converts this support score into a query-specific bandwidth,

$$h(q) = \mathrm{clip}(a + b\, g(q), h_{\min}, h_{\max}), \qquad a = h_{\min} + \mathrm{softplus}(\theta_a), \quad b = \mathrm{softplus}(\theta_b). \tag{4}$$

The nonnegative slope gives the map a simple rule. A query with weak nearby support can use a broader neighborhood. A query with strong nearby support can stay more local.

$$w_i(q) = \exp\left[-\frac{1}{2}\left(\frac{\tau_i - q}{h(q)}\right)^2\right].$$

The fitted local intercept becomes the canonical value at q. Ridge regularization stabilizes the fit, and a smooth fallback toward the kernel mean handles low effective support. Repeating the projection over all canonical queries produces $G_\phi(\mathcal{O})$. For a canonical query outside the observed temporal interval, the map uses endpoint hold:

$$\widetilde{y}(q) = \begin{cases} y_{\text{first}}, & q < \tau_{\min}, \\ \widetilde{y}_{\text{LP}}(q), & \tau_{\min} \le q \le \tau_{\max}, \\ y_{\text{last}}, & q > \tau_{\max}. \end{cases} \tag{5}$$

The main benchmark preserves the first and last context observations, so its canonical queries lie inside the observed temporal interval. The map is now support-aware, but the two scalars a, b still need to be aligned with the behavior of the frozen forecaster.

3.3 Calibrating the context map through the frozen forecaster

We learn only θ_a and θ_b with forecast pinball loss. We generate synthetic signals with trend, seasonal components, a chirp, and heteroskedastic noise. The calibration windows cover several observation densities. Held-out synthetic windows select the two scalars. We do not use real benchmark targets to fit a, b.

The frozen forecaster computes the forecast loss during calibration. Its parameters stay fixed. We calibrate the same two-scalar interface separately for Chronos-2, TimesFM-3, and TiRex-2. Appendix A.3 reports tests of the fixed design choices and the scalar calibration. After calibration, S2G provides the forecast-calibrated context map. HAMC then combines forecasts from nearby context views across the prediction horizon.

Calibration and evaluation separation. The backbone weights remain frozen throughout. Synthetic forecast loss selects only the two S2G scalars. The synthetic training windows use keep rates in $\{1.0, 0.8, 0.6, 0.4\}$. Benchmark validation selects context length and the fixed HAMC constants, and these choices are frozen before final evaluation. The observation-support study later evaluates nominal retention down to $r = 0.10$. Thus, the S2G scalars and HAMC composition remain unchanged at support levels below the lowest keep rate used for synthetic S2G calibration.

3.4 HAMC as a fixed robustness module

HAMC is the fixed robustness module of S2G-HAMC . HAMC evaluates nearby context views to reduce dependence on one calibrated context scale. The backbone remains unchanged. Let $G_{\phi,s}$ denote the frozen S2G map with its query-local bandwidth multiplied by a fixed scale $s \in \{0.75, 1.00, 1.25\}$. The same frozen forecaster evaluates each context:

$$\widehat{\mathbf{Q}}^{(s)} = F_{\theta_0}(G_{\phi,s}(\mathcal{O})), \qquad \theta_0 \text{ fixed}.$$

For horizon index $h = 1, \ldots, H$, HAMC first forms a horizon-dependent forecast,

$$w_h = \exp\left(-\frac{h-1}{0.5H}\right),$$

$$\widehat{\mathbf{Q}}_h^{\text{dir}} = w_h \widehat{\mathbf{Q}}_h^{(0.75)} + (1 - w_h)\widehat{\mathbf{Q}}_h^{(1.25)},$$

and combines it with the calibrated S2G forecast,

$$\widehat{\mathbf{Q}}_h^{\mathrm{HA}} = 0.5\widehat{\mathbf{Q}}_h^{(1.00)} + 0.5\widehat{\mathbf{Q}}_h^{\mathrm{dir}}.$$

A fixed-bandwidth degree-one local-polynomial context map supplies a complementary forecast,

$$\widehat{\mathbf{Q}}^{\mathrm{FLP}} = F_{\theta_0}(G_{\mathrm{FLP}}(\mathcal{O}))\,.$$

The final forecast is

$$\widehat{\mathbf{Q}}^{\text{S2G-HAMC}} = 0.95\,\widehat{\mathbf{Q}}^{\mathrm{HA}} + 0.05\,\widehat{\mathbf{Q}}^{\mathrm{FLP}}. \tag{6}$$

We enforce quantile monotonicity after each linear composition. Benchmark validation selects the HAMC composition constants, including the context-scale set and mixture weights. We freeze these choices before the final holdout. Appendix A.2 reports the final-holdout result. HAMC adds no trainable parameters or learned auxiliary forecaster. In S2G-HAMC, S2G builds the forecast-calibrated context. HAMC combines forecasts from nearby context views across the prediction horizon.

4 Experiments

The experiments test S2G-HAMC as a forecast-calibrated context-adaptation framework for frozen TSFMs. HAMC acts as a fixed robustness module. We first test the complete system across three forecasting families. We then test the S2G support-conditioning mechanism. Next, we vary observation support, forecast origin, and domain. We end with a broader comparison against pretrained forecasting systems. The appendix reports the fixed-design checks, final-holdout result, implementation details, and reproducibility details.

4.1 Experimental Setup

Datasets and observation-support protocol. We evaluate 12 public datasets spanning electricity, healthcare, smart-meter energy, mobility, retail, traffic, pedestrian activity, bike demand, weather, and large-scale retail demand. Electricity Load Diagrams (ECL) is the electricity dataset referenced by its common short name in the tables. The primary cross-backbone study uses a controlled distributed observation-support process with nominal retention $r = 0.20$. The broader pretrained-system comparison also includes an internal-block support process. Both processes preserve the first and last context observations and keep at least three observed locations. The dataset roster, per-dataset context-length grids, and evaluation protocol are fixed before evaluation on the final H-step target; validation never selects datasets.

Validation and evaluation. For a series of length n and horizon H, validation uses $[n-2H, n-H)$ and evaluation uses $[n-H, n)$. Each method selects context length on the validation target with one fixed support realization and then freezes that choice; method-specific interface constants are frozen separately as described in the appendix. The primary cross-backbone study evaluates 12 predefined distributed-support realizations without further tuning. The broader pretrained-system comparison uses three predefined realizations for each of the two support processes.

Uncertainty. The primary cross-backbone study reports 95% confidence intervals (CIs) from 5,000 paired hierarchical bootstrap resamples over dataset, support realization, and series. All relative gains are computed from matched targets.

Forecasting systems and context interfaces. The primary study uses frozen Chronos-2 (Ansari et al., 2025), TimesFM-3 (Jain & Sen, 2026), and TiRex-2 (Podest et al., 2026). For each backbone, the Linear interface provides the validation-selected reference context. S2G is the forecast-calibrated support-aware context component of the proposed framework, and HAMC supplies a fixed robustness-oriented composition over nearby context-induced forecasts. S2G-HAMC is the complete proposed method; the backbone remains unchanged. The broader system-level comparison additionally includes Chronos-Bolt-small (Ansari et al., 2024), Sundial-Base-128M (Liu et al., 2025), and Toto-2.0-313M (Khwaja et al., 2026).

Metric. We evaluate quantiles $\mathcal{Q} = \{0.1, \ldots, 0.9\}$. For pooled target points $\mathcal{S}$, Global Weighted Quantile Loss is

$$\mathrm{WQL} = \frac{2 \sum_{(i,t)\in\mathcal{S}} \sum_{\tau\in\mathcal{Q}} \rho_\tau(y_{it} - \widehat{q}_{it,\tau})}{|\mathcal{Q}| \sum_{(i,t)\in\mathcal{S}} |y_{it}|}, \qquad \rho_\tau(e) = \max\{\tau e, (\tau - 1)e\}.$$

Lower WQL is better.

4.2 S2G-HAMC Transfers across Frozen Forecasting Families

Table 1 is the primary cross-backbone test of the complete S2G-HAMC framework. The mean WQL gains are 8.58%, 8.60%, and 7.62% on Chronos-2, TimesFM-3, and TiRex-2. These gains use the corresponding validation-selected Linear interface as the reference. S2G-HAMC improves 35/36 backbone–dataset pairs, and all three 95% CIs exclude zero. S2G alone yields 8.11%, 8.24%, and 7.39%. The complete system has a larger mean gain than S2G alone on all three backbones.

Table 1: **Primary cross-backbone context-interface result.** WQL is averaged over 12 fixed distributed-support realizations at nominal retention $r = 0.20$. S2G is the context-adaptation component; "+HAMC" completes S2G-HAMC with the fixed robustness-oriented composition module. Mean gains are equal-dataset means relative to the corresponding validation-selected Linear context interface; all methods are evaluated on the same targets and support realizations. Bold marks the lowest WQL within each backbone family for each dataset. Short labels are used for compactness; Appendix Table 9 gives the full dataset names.

	Chronos-2 family			TimesFM-3 family			TiRex-2 family		
Dataset	Linear	S2G	**+HAMC**	Linear	S2G	**+HAMC**	Linear	S2G	**+HAMC**
Dominick	0.6334	0.5640	**0.5618**	0.6306	0.5648	**0.5634**	0.5999	0.5421	**0.5414**
Electricity / ECL	0.1352	**0.1273**	0.1276	0.1432	0.1214	**0.1212**	0.1374	0.1229	**0.1226**
Hospital	0.0726	**0.0705**	0.0728	0.0802	0.0663	**0.0656**	0.0740	0.0654	**0.0653**
London	0.4078	0.3775	**0.3759**	0.4078	0.3903	**0.3898**	0.4200	0.3975	**0.3967**
Loop Seattle	0.0829	0.0670	**0.0660**	0.0802	0.0649	**0.0643**	0.0785	0.0678	**0.0671**
M5	0.6791	0.6162	**0.6104**	0.6251	0.5882	**0.5849**	0.6328	0.5872	**0.5849**
Mexico Bikes	0.7825	0.6515	**0.6409**	0.7717	0.7600	**0.7562**	0.8002	0.6901	**0.6871**
Pedestrian	0.3420	0.3424	**0.3365**	0.3420	0.3311	**0.3297**	0.3336	0.3205	**0.3199**
SZ Taxi	0.1565	0.1524	**0.1516**	0.1541	0.1516	**0.1512**	0.1577	0.1530	**0.1526**
Taxi 30m	0.3469	0.3235	**0.3233**	0.3386	0.3294	**0.3291**	0.3543	0.3440	**0.3437**
Temp./Rain	0.8276	0.7537	**0.7416**	1.0382	0.8743	**0.8725**	0.9423	0.9145	**0.9121**
Vehicle Trips	0.3554	0.3305	**0.3267**	0.3372	0.3311	**0.3301**	0.3473	**0.3331**	0.3332
Mean gain over Linear ↑	–	+8.11%	**+8.58%**	–	+8.24%	**+8.60%**	–	+7.39%	**+7.62%**
Improved datasets ↑	–	11/12	**11/12**	–	12/12	**12/12**	–	12/12	**12/12**

The transfer of S2G-HAMC across three independently pretrained forecasting families shows that the complete framework is not tied to one backbone architecture. The 95% CIs for its mean gains over Linear are [5.25%, 12.13%], [4.77%, 12.66%], and [5.31%, 10.22%] for Chronos-2, TimesFM-3, and TiRex-2, respectively.

4.3 Support Conditioning Provides Additional Gains beyond Forecast Calibration

We hold each dataset–backbone context length fixed to test the support-aware step. We compare the full (a, b) bandwidth law with a forecast-calibrated global scale that sets $b = 0$. Support conditioning yields additional mean WQL gains of **2.42%**, **4.28%**, and **5.53%** for Chronos-2, TimesFM-3, and TiRex-2. All three paired hierarchical bootstrap 95% CIs exclude zero. Appendix Table 6 reports the full calibration ladder and the fixed-scale control.

Table 2: **Support conditioning provides additional gains beyond forecast calibration.** All variants use the same S2G-selected context length C. Values are the mean WQL gain of the full support-conditioned (a, b) calibration over a forecast-calibrated global scale ($b = 0$); brackets are 95% paired hierarchical bootstrap CIs. No real benchmark target selects these scalars.

Backbone	Gain of full (a, b) over calibrated global scale
Chronos-2	$+2.42\%$ **[0.33, 4.76]**
TimesFM-3	$+4.28\%$ **[0.27, 8.54]**
TiRex-2	$+5.53\%$ **[1.13, 10.20]**

4.4 Gains Increase as Observation Support Becomes Sparse

We test how the S2G-HAMC gain changes as observation support becomes sparse. We evaluate $r \in \{0.10, 0.20, 0.30, 0.40\}$. We use the same validation-only context-selection rule at each support level. We keep the S2G calibration, HAMC composition, and backbone parameters fixed. As retention falls from $r = 0.40$ to $r = 0.10$, the Chronos-2 gain rises from 4.10% to 14.74%. The TimesFM-3 gain rises from 2.78% to 9.61%. The TiRex-2 gain rises from 3.28% to 15.15%. All three forecasting families show the same monotone pattern. This pattern is consistent with the support-aware S2G design, while HAMC remains fixed. Appendix Table 8 reports the values and confidence intervals.

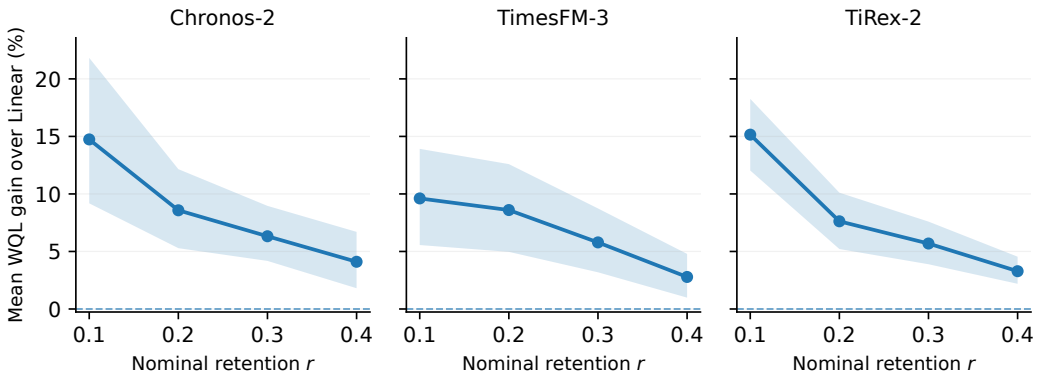

Figure 2: **Context-interface gains increase as observation support becomes sparse.** The panels show the mean WQL gain of S2G-HAMC over the corresponding validation-selected Linear context interface as nominal observation retention r varies. Each panel shows one frozen forecasting family; shaded regions are 95% paired hierarchical bootstrap confidence intervals. Context length follows the same validation-only selection rule at each support level; S2G calibration, HAMC composition, and backbone parameters remain frozen.

4.5 Temporal Transfer across Forecast Origins

We keep the method, calibration, and observation-support protocol fixed across four horizon-aligned forecast origins. Under this test, **43/48 dataset–origin cells improve over Linear**. The overall equal-cell mean relative WQL gain is **+9.10%**. Appendix A.6 gives the full forecast-origin result.

4.6 Generalization to Previously Unused Domains

We test three previously unused UCI domains without retuning S2G or HAMC. Each domain uses four support patterns and two retention levels. This gives a $3 \times 4 \times 2$ grid with 24 *domain* $\times$ *support pattern* $\times$ *severity* cells. The table reports the number of improved cells and the equal-cell mean gain. Appendix A.7 gives the full test protocol.

Frozen-protocol transfer across domain × support pattern × severity cells.

Metric	Evaluation grid	Improved cells	Equal-cell mean gain
Quantile loss	$3 \times 4 \times 2 = 24$	**22/24**	**+8.26%**
Median-forecast MAE	$3 \times 4 \times 2 = 24$	**19/24**	**+6.89%**

The frozen system improves quantile loss in **22/24** cells and median-forecast MAE in **19/24** cells. The equal-cell mean gains are **8.26%** and **6.89%**. The test uses the same S2G calibration and HAMC composition in every cell. The result therefore covers new domains, support structures, and observation severities without held-out-domain tuning.

4.7 Broader System Comparison

We next compare seven frozen/pretrained forecasting systems, including the complete S2G-HAMC framework. This broader test includes both distributed and internal-block observation support.

Table 3: **Broader frozen/pretrained comparison across two observation-support processes.** The table reports Global WQL (↓) at nominal retention $r = 0.20$, averaged over three realizations for distributed and internal-block support. The first six columns are frozen/pretrained comparison systems; the final column is frozen Chronos-2 with S2G-HAMC. C2 denotes Chronos-2, TFM-3 denotes TimesFM-3, and TiRex denotes TiRex-2. This table is for system positioning, while Table 1 is the primary cross-backbone interface comparison. Best is bold and second best is underlined.

Dataset	C2	Bolt	TFM-3	TiRex	Sundial	Toto	**C2 + S2G-HAMC**
Dominick	0.5928	0.5879	0.5955	**0.5722**	0.5836	<u>0.5780</u>	0.5798
Electricity / ECL	0.1355	<u>0.1350</u>	0.1485	0.1442	0.1513	0.1607	**0.1187**
Hospital	0.0735	<u>0.0734</u>	0.0759	**0.0718**	0.0757	0.0926	0.0749
London	<u>0.4142</u>	0.4284	0.4247	0.4497	0.4760	0.4297	**0.3836**
Loop Seattle	0.0772	0.0771	0.0789	0.0774	<u>0.0735</u>	0.0800	**0.0635**
M5	0.6942	0.7356	0.6816	**0.6237**	<u>0.6514</u>	0.7252	0.6659
Mexico Bikes	<u>0.7353</u>	0.7382	0.8061	0.8000	0.8248	0.7896	**0.6444**
Pedestrian	0.3839	<u>0.3401</u>	0.3834	0.3752	0.3988	0.4312	**0.3303**
SZ Taxi	0.1623	<u>0.1591</u>	**0.1553**	0.1633	0.1687	0.1630	0.1708
Taxi 30m	0.3471	0.3437	<u>0.3394</u>	0.3731	0.4050	0.3695	**0.3230**
Temp./Rain	**0.8231**	0.8700	0.9077	<u>0.8531</u>	0.9026	0.9969	0.9311
Vehicle Trips	<u>0.3528</u>	0.3632	**0.3300**	0.3602	0.3677	0.3771	0.3564
Mean rank ↓	3.33	3.42	4.17	3.58	5.17	5.67	**2.67**
# best datasets ↑	1	0	2	3	0	0	**6**

Table 3 uses a broader setting than the primary cross-backbone study. It averages results from both distributed and internal-block observation support. Table 1 uses only distributed support with 12 fixed realizations. We use Table 3 for the broader system comparison, not as a dataset-by-dataset repeat of Table 1. In this broader setting, Chronos-2 + S2G-HAMC has the **lowest mean rank (2.67)** and the **best WQL on 6/12 datasets**.

5 Discussion

The results support S2G-HAMC as a context-adaptation framework for frozen TSFMs. S2G builds the support-aware, forecast-calibrated context map. HAMC adds a fixed robustness module over nearby context views. The complete system achieves positive mean WQL gains across all three forecasting families. The gains increase as observation support becomes sparse, and the held-out-domain test extends the effect to previously unused domains. The two S2G calibration scalars are estimated from synthetic signals rather than benchmark targets.

Why local support matters. Global retention does not describe the evidence available at each canonical query. Two queries in the same context can have different distances to nearby observations even when the window has the same overall retention. A single global scale cannot respond to this difference. S2G uses local support to change the projection scale at each query. Table 2 shows that this query-local step improves WQL beyond a forecast-calibrated global scale at the same context length. Figure 2 shows a second pattern: the complete S2G-HAMC gain becomes larger as observation support becomes sparse. These results connect the support-aware mechanism to the setting in which context construction has the largest observed effect.

Context adaptation is not an imputation objective. S2G does not train a model to reconstruct missing history as an end in itself. It constructs the canonical numerical input consumed by an already-frozen forecaster. The optimization target is the forecaster's own pinball loss. The resulting context is therefore a model-facing input rather than a separate forecasting output. This view makes the observation-to-context map an explicit adaptation surface while the pretrained forecasting model remains unchanged.

The two modules act at different stages of the adaptation path. S2G changes the observation-to-context map before the frozen forecaster. HAMC acts on forecasts produced from nearby context views and combines them across the prediction horizon. This separation keeps the forecasting backbone fixed while addressing two parts of the interface: local observation support and sensitivity to context scale. Table 1 evaluates the complete S2G-HAMC system, while Table 2 isolates the support-aware mechanism inside S2G.

Mechanism and system roles. The matched-context ablation tests the support-aware step inside S2G. Query-local support conditioning yields mean WQL gains of 2.42–5.53% over a forecast-calibrated global scale. All three 95% CIs exclude zero. Table 2 tests this mechanism at a fixed context length. Table 1 tests the complete S2G-HAMC system across three frozen forecasting families. The two tables answer different questions and together link the S2G mechanism to the full-system result.

6 Conclusion

Frozen TSFMs can be adapted without changing their parameters. S2G-HAMC does this by changing the context they receive and by combining forecasts from nearby context views. S2G builds a support-aware, forecast-calibrated context. HAMC provides a fixed robustness module for forecast composition. Across 12 datasets and three forecasting families, the complete system improves 35/36 backbone–dataset pairs. Its mean WQL gains are 7.62–8.60% over the corresponding validation-selected Linear interface. The gains increase as observation support becomes sparse. These results show that numerical context can serve as an adaptation interface for frozen forecasters.

AI use statement

Generative AI tools were used to aid or polish writing and for retrieval and discovery (e.g., finding related work).

Ethics statement

This work uses existing time-series datasets and involves no new human-subject data collection or identity inference. Forecast errors can have unequal operational consequences when temporal observations provide uneven local support. Deployment therefore requires domain-specific assessment of calibration and failure modes, together with compliance with the licenses and privacy conditions of the underlying data.

Reproducibility statement

Sections 3–4 specify the method, evaluation protocol, uncertainty analysis, and mechanism ablation. Appendix A provides the frozen calibration values, exact observation-support sensitivity val-

ues, forecast-origin and held-out-domain tests, robustness to fixed design choices, a final-holdout confirmation, the inference profile, and dataset provenance.

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. *arXiv preprint arXiv:2410.10393*, 2024. URL https://arxiv.org/abs/2410.10393.

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. GluonTS: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL https://jmlr.org/papers/v21/19-820.html.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, 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 Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Abdelhakim Benechehab, Vasilii Feofanov, Giuseppe Paolo, Albert Thomas, Maurizio Filippone, and Balázs Kégl. AdaPTS: Adapting univariate foundation models to probabilistic multivariate time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3731–3748, 2025. URL https://proceedings.mlr.press/v267/benechehab25a.html.

Afrin Dange and Sunita Sarawagi. TFMAdapter: Lightweight instance-level adaptation of foundation models for forecasting with covariates. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 498–507, 2025. doi: 10.1145/3746252.3761272. URL https://doi.org/10.1145/3746252.3761272.

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.

Dheeru Dua and Casey Graff. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2019. URL https://archive.ics.uci.edu/.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall, 1996.

Shikai Fang, Qingsong Wen, Yingtao Luo, Shandian Zhe, and Liang Sun. BayOTIDE: Bayesian online multivariate time series imputation with functional decomposition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12993–13009, 2024. URL https://proceedings.mlr.press/v235/fang24d.html.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://forecastingdata.org/.

Zhangyi Hu, Jiemin Wu, Hua Xu, Mingqian Liao, Ninghui Feng, Bo Gao, Songning Lai, and Yutao Yue. IMTS is worth time $\times$ channel patches: Visual masked autoencoders for irregular multivariate time series prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24805–24822, 2025. URL https://proceedings.mlr.press/v267/hu25y.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.

Ayush Jain and Rajat Sen. TimesFM-3: A zero-shot foundation model for multivariate forecasting. Google Research Blog, August 2026. URL https://www.research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/. Official TimesFM-3 release article, 31 August 2026.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026. URL https://arxiv.org/abs/2605.20119.

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.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=syfWdclGE1.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317, 2025. URL https://proceedings.mlr.press/v267/liu25be.html.

Ivan Marisca, Cesare Alippi, and Filippo Maria Bianchi. Graph-based forecasting with missing data through spatiotemporal downsampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34846–34865, 2024. URL https://proceedings.mlr.press/v235/marisca24a.html.

Jing Peng, Meiqi Yang, Qiong Zhang, and Xiaoxiao Li. S4M: S4 for multivariate time series forecasting with missing values. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7b2f0758334389b8ad0665a9bd165463-Abstract-Conference.html.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex-2: Generalizing tirex to multivariate data and streaming. *arXiv preprint arXiv:2607.01204*, 2026. URL https://arxiv.org/abs/2607.01204.

Yunzhong Qiu, Zhiyao Cen, Zhongyi Pei, Chen Wang, and Jianmin Wang. Adapt data to model: Adaptive transformation optimization for domain-shared time series foundation models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uTK1SNgi1N.

Zhixian Wang, Linxiao Yang, Liang Sun, Qingsong Wen, and Yi Wang. Task-oriented time series imputation evaluation via generalized representers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f88264fcc54775ee1706116e90fe351a-Abstract-Conference.html.

Jingge Xiao, Yile Chen, Gao Cong, Wolfgang Nejdl, and Simon Gottschalk. FlexTSF: A universal forecasting model for time series with variable regularities. *arXiv preprint arXiv:2410.23160*, 2024. URL https://arxiv.org/abs/2410.23160.

A Additional Experimental Details

A.1 Synthetic calibration and implementation

The primary adapter uses the deterministic synthetic calibration protocol described in Section 3.3; no real benchmark target is used to fit the interface parameters a, b. For a series of length $L = 7000$ and $t = 0, \ldots, L-1$, the signal is

$$\begin{aligned} x_t = {} & 0.0008t + 0.25\sin(2\pi t/900) \\ & + [1 + 0.25\sin(2\pi t/700)]\sin(2\pi t/48) \\ & + 0.45\cos(2\pi t/13 + 0.3) \\ & + 0.25\sin\big(2\pi\left[t/180 + 5(t/L)^2\right]\big) + \epsilon_t, \end{aligned}$$

where

$$\epsilon_t \sim \mathcal{N}(0, \sigma_t^2), \qquad \sigma_t = 0.08 + 0.06\left[1 + \sin(2\pi t/350)\right].$$

The series is split 70/15/15% into train/validation/test regions. Training windows use a mixture of predefined observation keep rates, while held-out forecast pinball loss selects the two trainable adapter scalars. The forecasting backbone remains fixed throughout.

Table 4: Synthetic calibration and shared local-projection settings for the primary adapter.

Setting	Value
Synthetic series length	7000
Context / horizon / stride	256 / 48 / 12
Training keep rates	{1.0, 0.8, 0.6, 0.4}
Learned calibration scalars	2 (θ_a, θ_b)
Local polynomial degree	1
Soft-support temperature	0.25
Support-score clipping	$[0, 50]$ canonical steps
Bandwidth clipping	$[0.05, 3.0]$
Local-regression ridge	10^{-4}
Optimizer	Adam with decoupled weight decay (AdamW)
Learning rate / weight decay	0.005 / 0.0
Batch size / max epochs / patience	4 / 80 / 15
Gradient clipping	1.0
Frozen backbone precision	bfloat16

The soft-support temperature, clipping ranges, local-polynomial degree, ridge form, and fallback rule are treated as shared implementation settings rather than dataset-specific free parameters. Real benchmark targets are not used to fit a, b; dataset-specific context selection is a separate validation step described in Appendix A.8.

The frozen backbone-specific calibration values used in the experiments are reported in Table 5. All three use the shared clipping interval $h_{\min} = 0.05$, $h_{\max} = 3.0$.

Table 5: Frozen S2G calibration scalars used for each forecasting backbone.

Backbone	a	b
Chronos-2	0.4909	0.1037
TimesFM-3	0.2000	0.3800
TiRex-2	0.2000	0.3800

For each query, normalized offsets are clipped to $[-10, 10]$. Let $m(q) = \sum_i w_i(q)$ denote kernel mass and

$$\mu_K(q) = \frac{\sum_i w_i(q) y_i}{m(q) + 10^{-8}}$$

the kernel mean. Degree-one weighted least squares uses ridge $10^{-4}[1 + m(q)]$. If $\widehat{y}_{\mathrm{LP}}(q)$ is the local-polynomial intercept, the implementation uses the smooth low-support fallback

$$z(q) = \sigma(2[m(q) - 1.5]), \qquad \widetilde{y}(q) = z(q)\widehat{y}_{\mathrm{LP}}(q) + [1 - z(q)]\mu_K(q),$$

where $\sigma(\cdot)$ is the logistic sigmoid. This formula is applied inside observed temporal support; Eq. 5 in the main method defines endpoint behavior outside that interval.

A.2 Final holdout for HAMC

We selected the HAMC constants on validation data and fixed them before the final holdout. In the final holdout, the complete S2G-HAMC system has a cross-backbone mean relative WQL gain of **+0.452%** over frozen S2G.

A.3 Robustness to fixed design choices and forecast calibration

We test two parts of the S2G context map. The first test checks the fixed implementation settings. The second test measures the effects of forecast calibration and support conditioning. For the fixed-design test, we change the bandwidth clipping range, soft-support temperature, and ridge coefficient one at a time. We keep the Chronos-2 calibration scalars a, b fixed. This sensitivity analysis uses Hospital, M5, Pedestrian, and Temperature/Rain under three fixed support realizations.

For the scalar-calibration ablation, all variants use the same S2G-selected context length C for each dataset–backbone pair. This keeps the context budget fixed. We compare three variants. The first uses a fixed global bandwidth law with $a = 0.5, b = 0$. The second uses a forecast-calibrated global scale with $b = 0$. Held-out synthetic forecast pinball loss selects a for this variant. The third uses the full support-conditioned (a, b) calibration. The calibrated global scale uses $a = 0.90$ for Chronos-2 and $a = 0.75$ for TimesFM-3 and TiRex-2. We do not use real benchmark targets to select these scalars. The evaluation uses the same 12 datasets and 12 fixed distributed-support realizations as Table 1. We compute uncertainty with 5,000 paired hierarchical bootstrap resamples over dataset, support realization, and series.

Table 6: **Forecast-calibration ladder at matched context length.** Each column reports the relative WQL gain of the first method named in the header over the second; brackets are 95% paired hierarchical bootstrap CIs.

Backbone	Gain of calibrated global over fixed $a = 0.5$	Gain of full (a, b) over calibrated global
Chronos-2	$+14.12\%$ [3.66, 25.41]	$+2.42\%$ **[0.33, 4.76]**
TimesFM-3	$+11.66\%$ [3.97, 19.93]	$+4.28\%$ **[0.27, 8.54]**
TiRex-2	$+13.00\%$ [4.67, 22.04]	$+5.53\%$ **[1.13, 10.20]**

Table 7: **Robustness to fixed design choices.** Shared implementation settings are perturbed one at a time while the frozen Chronos-2 calibration scalars remain unchanged. Changes are reported relative to the default setting. The matched-context forecast-calibration mechanism ablation is reported in Table 2.

Perturbation	Mean WQL change	Max absolute WQL change
Wider bandwidth clip $[0.025, 4.0]$	$+0.000\%$	0.000%
Narrower bandwidth clip $[0.10, 2.0]$	$+0.001\%$	0.006%
Soft-support temperature $\times 0.5$	$+0.020\%$	0.231%
Soft-support temperature $\times 2$	-0.068%	0.639%
Ridge 10^{-5}	-0.044%	0.530%
Ridge 10^{-3}	$+0.133\%$	1.353%

Across the fixed-design changes, the mean WQL change stays within 0.14%. The largest cell-wise change is 1.35%. Appendix Table 6 reports the full matched-C calibration ladder. Table 2 in the main text tests the support-conditioning step. We keep C at the S2G-selected value so that this test changes only the scalar law.

A.4 OBSERVATION-SUPPORT SENSITIVITY VALUES

Table 8 reports the numerical values from Figure 2. Values are equal-dataset mean WQL gains of S2G-HAMC over the corresponding validation-selected Linear context interface. Brackets show 95% paired hierarchical bootstrap intervals. The support level r is nominal retention because endpoint preservation and the minimum-observation rule can slightly change the realized fraction on short contexts. Context length follows the same validation-only selection rule at each support level. The S2G calibration, HAMC composition, and backbone weights remain frozen. We use 5,000 paired bootstrap resamples for this sweep. The repeated $r = 0.20$ point estimate matches the primary study; small CI differences can arise from independent bootstrap resampling.

Table 8: Observation-support sensitivity underlying Figure 2. The table reports mean relative WQL gain (%) over Linear with 95% paired hierarchical bootstrap CIs.

Nominal retention r	Chronos-2	TimesFM-3	TiRex-2
0.10	+14.74 [9.18, 21.82]	+9.61 [5.56, 13.92]	+15.15 [12.04, 18.26]
0.20	+8.58 [5.29, 12.15]	+8.60 [4.97, 12.59]	+7.62 [5.21, 10.11]
0.30	+6.32 [4.18, 8.96]	+5.79 [3.19, 8.74]	+5.69 [3.89, 7.60]
0.40	+4.10 [1.80, 6.70]	+2.78 [0.99, 4.79]	+3.28 [2.19, 4.55]

A.5 INFERENCE PROFILE

For polynomial degree $d = 1$, the S2G projection requires $O(CN)$ arithmetic for C canonical queries and N observations. Each local linear solve has constant size. S2G-only uses one frozen-backbone forecast call. The full system uses four fixed-context forward passes through the frozen backbone. It adds no trainable parameters. In the sequential implementation, peak GPU memory stays at the single-branch level.

A.6 FORECAST-ORIGIN TEST

We evaluate frozen Chronos-2 + S2G-HAMC at four horizon-aligned origins from $-1H$ to $-4H$. The test uses all 12 datasets with distributed observation support at nominal retention $r = 0.20$. We keep the S2G and HAMC parameters fixed. The mean relative WQL gains at the four origins are +10.89%, +11.92%, +7.91%, and +5.68%. Across all dataset–origin cells, the equal-cell mean gain is **+9.10%**. In total, **43/48** cells improve over Linear. This test measures temporal transfer without origin-specific retuning.

A.7 HELD-OUT UCI-DOMAIN TEST

We also test the complete frozen S2G-HAMC framework on three UCI domains outside the main 12-dataset set: Appliances, Metro Interstate Traffic Volume, and Bike Sharing. We do not retune the method for these domains. The test uses context length 256, prediction length 48, stride 48, a 60/20/20 train/validation/test split, and three fixed observation-support seeds (10101–10103). Each dataset uses four support patterns: distributed thinning, independent random support, internal-block support, and tail support. We test each pattern at nominal retention 0.2 and 0.1. This design gives 24 domain–support-pattern–severity cells. We keep the S2G and HAMC parameters fixed, and we do not use held-out targets for method selection.

Across these 24 cells, Chronos-2 + S2G-HAMC improves quantile loss in **22/24** cells. It improves median-forecast MAE in **19/24** cells. The equal-cell mean gains are **8.26%** for quantile loss and **6.89%** for MAE. These results extend the context-interface effect beyond the main 12-dataset set and across several observation-support patterns.

A.8 DATASET AND EVALUATION PROTOCOL

Table 9: **Dataset and evaluation protocol.** H is the prediction horizon and N is the number of eligible evaluation series. The context column lists the candidate lengths used for validation selection. Mean Absolute Scaled Error (MASE) appears only as a dataset-period descriptor. Electricity Load Diagrams (ECL) and Shenzhen (SZ) Taxi are written with their common abbreviations.

Dataset	Frequency	H	Eval. N	Context candidates	MASE period
Electricity / ECL	hourly	48	192	96, 168, 336	24
Hospital	monthly	12	192	12, 24, 36	12
London Smart Meters	30 min	48	190	96, 336, 672	48
Vehicle Trips	daily	30	54	30, 60, 80	1
Dominick	weekly	8	192	16, 32, 48	1
Taxi 30min	30 min	48	192	96, 336, 672	48
Pedestrian Counts	hourly	48	66	48, 168, 336	24
Mexico City Bikes	hourly	48	192	48, 168, 336	24
Temperature Rain	daily	30	192	60, 120, 365	1
SZ Taxi	hourly	48	156	48, 168, 336	24
Loop Seattle	hourly	48	192	48, 168, 336	24
M5	daily	28	192	56, 112, 168	1

Dataset provenance. We use public benchmark identifiers and access paths distributed through established forecasting benchmark infrastructure. Standard forecasting collections are aligned with GluonTS and the Monash Time Series Forecasting Archive (Alexandrov et al., 2020; Godahewa et al., 2021); the SZ Taxi and Loop Seattle transport benchmarks follow GIFT-Eval (Aksu et al., 2024). Mexico City Bikes follows the public benchmark release used by Chronos-family evaluations, and the held-out Appliances, Metro Interstate Traffic Volume, and Bike Sharing domains are obtained from the UCI Machine Learning Repository (Dua & Graff, 2019). The artifact file `data/DATASET_PROVENANCE.txt` records the exact benchmark identifiers used in our evaluation code.

For every series of length n, the split is

$$\text{TRAIN} = [0, n-2H), \qquad \text{validation} = [n-2H, n-H), \qquad \text{evaluation} = [n-H, n).$$

The primary cross-backbone analysis uses distributed support at nominal retention $r = 0.20$. The broader seven-system comparison additionally uses an internal-block process. Distributed support thinning places retained observations across the full context, whereas the internal-block process restricts support around one contiguous interior interval. The first and last context points are preserved, and every realized context contains at least three observations. Consequently, all canonical queries in this main protocol lie inside the observed temporal interval.

For each dataset–pattern pair, context length is selected on the penultimate-H validation target using one fixed support realization and at most 64 deterministic series. WQL is the selection metric, and ties favor the smaller context. The selected context is then fixed for evaluation. The broader seven-system table uses three predefined support realizations, while the cross-backbone distributed-support analysis extends the same frozen selections to 12 support realizations. All compared methods use the same fixed evaluation-series roster.