TSDATAMODEL: LEARNING FORECASTING UTILITY FOR EXTERNAL CONTEXT RETRIEVAL WITH FROZEN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models can forecast new domains with frozen weights, but short observed histories may leave their predictions ambiguous. An external archive can supply additional histories, yet conventional retrieval asks which history looks similar to a target query rather than which one actually improves its forecast. We propose TSDATAMODEL, which learns this missing quantity—*forecasting utility*. For a labeled fitting query, utility is the reduction in forecasting loss obtained by prepending a source history to the query. We estimate this utility with global, reference-based, and parametric scorers. AUTO-TSDM chooses among them using a chronologically held-out selection set and freezes the choice before testing; test futures are never used for retrieval or model selection, and the forecasting backbone is never updated. We evaluate six target datasets with TimesFM 2.5-200M and TimeMoE-50M, a disjoint archive built from three source datasets, five random seeds, and length-matched context controls. AUTO-TSDM improves over short-history target-only forecasting in 11 of 12 target–backbone settings, reducing mean MAE by 9.3% on average, and outperforms the strongest non-utility retrieval result in mean MAE in 11 of 12 settings. These results position forecasting-utility retrieval as a practical weight-update-free alternative when longer target histories or weight access are unavailable, rather than as a universal replacement for either.

1 INTRODUCTION

Time-series foundation models (TSFMs) are pretrained on heterogeneous corpora and can forecast previously unseen series without training a separate model for every dataset (Das et al., 2023; Ansari et al., 2024; Woo et al., 2024; Shi et al., 2025). This capability is especially attractive when target labels are limited or when the forecaster is exposed only through inference. However, zero-shot use does not remove a basic information constraint: a short observed history can be compatible with several different futures. Fine-tuning can adapt the model, but it requires weight access and may be costly or unstable under limited supervision; parameter-efficient methods such as LoRA reduce, but do not remove, these requirements (Hu et al., 2021).

We study a complementary setting: the forecaster remains frozen, the target history is short, and a separate archive of historical time-series windows is available. Prepending a few archive windows gives the forecaster a larger temporal context without changing its parameters. The archive may be useful even when it comes from other datasets because temporal shapes such as peaks, recoveries, oscillations, and regime changes can recur across domains. This is not a cold-start assumption—the current target history is observed—and it does not imply that external data are preferable to genuine target history. Rather, the question is whether an archive can help when a longer target history is unavailable.

The main difficulty is retrieval. Similarity-based methods rank a source window by how closely its observed values or representation match the target history. Forecasting requires a different answer: *will placing this source window before this query reduce the error of this particular forecaster?* With a short target history, similar observed patterns may still lead to different futures. A source window can therefore resemble the target query but provide misleading context, while a less similar

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.

source may trigger a useful forecasting pattern. Random prefixes and similarity-based retrieval can consequently introduce negative transfer. Adding more context alone is insufficient; the central problem is to identify context that is useful to the frozen forecaster.

We introduce TSDATAMODEL to learn that distinction from forecasting feedback. For a frozen forecaster F, source history s, target history q, and observed future y, we define

$$u_F(s, q) = \ell(F(q), y) - \ell(F([s; q]), y). \tag{1}$$

A positive value means that s improves the forecast of q; a negative value means that it is harmful. Sparse evaluations of Equation 1 form a source–query utility table. TSDataModel turns this table into retrieval scores at three levels: global estimators identify sources that are reliably useful across queries, a reference estimator transfers utilities from similar fitting queries, and parametric estimators predict query-conditioned utility. As Figure 1 shows, AUTO-TSDM fits these candidates on chronologically earlier queries, selects one by forecasting MAE on a separate selection set, and freezes the selected policy before test. Thus, "Auto" is a validation-only selection procedure, not an oracle estimator chosen from test results.

This formulation separates two tasks that similarity retrieval conflates: describing how two histories relate and measuring whether their interaction with a fixed forecaster is useful. It also separates context selection from forecast-model adaptation. TSDataModel uses target futures only to supervise a small retrieval policy; it neither trains a new forecaster nor changes the backbone. The same interface therefore applies to open-weight and black-box forecasters, provided that their predictions can be queried during offline utility construction.

Our evaluation is designed around the resulting claims and their limits. We use six targets from energy, traffic, mobility, and weather, a disjoint three-dataset source archive, two frozen TSFMs, and five seeds. Prefix methods receive three source histories plus the 12-step target history, while Target-48 receives 48 genuine target observations at the same forecast origin. We compare against random retrieval, raw and feature similarity, DTW, MMD, TS2Vec, backbone embeddings, naive RAF, direct predictors, and LoRA. Utility-aware retrieval is consistently stronger than non-utility retrieval, but genuine long history and LoRA remain stronger in many settings. This boundary is important: TSDataModel is intended for restricted-context, frozen-model deployment, not as a universal substitute for information or parameter adaptation.

Our contributions are:

- We formulate external temporal-context retrieval in terms of *forecasting utility*: the observed loss reduction caused by a source history for a target query under a fixed backbone.
- We develop global, reference-based, and parametric utility estimators, together with AUTO-TSDM, a chronological validation-only procedure that selects an estimator without test-label access or backbone updates.
- We provide a controlled study across six targets and two TSFMs with matched context and supervision comparisons, five-seed uncertainty, computational costs, retrieval-budget analysis, and explicit failure cases.

2 RELATED WORK

2.1 TIME-SERIES FORECASTING AND FOUNDATION MODELS

Time-series forecasting has progressed from statistical and task-specific neural models to large pretrained forecasters. Recent architectures such as DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2022), and iTransformer (Liu et al., 2024a) improve forecasting through specialized inductive biases, but they are generally trained separately for each target dataset. Time-series foundation models instead pretrain on heterogeneous corpora to support zero-shot or few-shot generalization. Representative models include Lag-Llama (Rasul et al., 2023), TimesFM (Das et al., 2023), Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), MOMENT (Goswami et al., 2024), Timer (Liu et al., 2024b), Tiny Time Mixers (Ekambaram et al., 2024), Time-MoE (Shi et al., 2025), and Timer-XL (Liu et al., 2025). These models reduce the need for dataset-specific architecture design, but adaptation to a data-scarce target domain remains challenging. In contrast to fine-tuning, our work studies a tuning-free approach that improves a frozen foundation forecaster through external temporal contexts.

2.2 Data Valuation and Datamodeling

Data valuation estimates how training samples affect downstream predictions or model performance. Representative approaches include influence functions (Koh & Liang, 2017), Data Shapley (Ghorbani & Zou, 2019), KNN-Shapley (Jia et al., 2019), TracIn (Pruthi et al., 2020), and reinforcement-learning-based data valuation (Yoon et al., 2020). Datamodeling (Ilyas et al., 2022) learns a surrogate mapping from training-data subsets to model outputs, enabling reusable analysis of data contributions without evaluating every counterfactual training set. However, most existing methods assume sample-level supervision and a fixed target objective. Time-series context augmentation poses a different challenge because the usefulness of a source window can depend on the target query, forecasting horizon, and frozen forecaster. TSDataModel extends the datamodeling principle to this setting by modeling source-query forecasting utility.

2.3 Time-Series Retrieval and Context Augmentation

Time-series retrieval commonly relies on raw-value distances, dynamic time warping (Sakoe & Chiba, 1978), differentiable alignment measures (Cuturi & Blondel, 2017), statistical descriptors (Fulcher et al., 2013; Fulcher & Jones, 2017), distributional discrepancy such as maximum mean discrepancy (Gretton et al., 2012), or learned representations such as TS2Vec (Yue et al., 2022). These methods provide useful non-utility baselines for selecting historical analogies. More recently, retrieval-augmented forecasting methods have been developed for time-series foundation models. RAF (Tire et al., 2024) studies retrieval augmentation for zero-shot forecasting, TimeRAF (Zhang et al., 2025) learns a retriever over task-specific time-series knowledge bases, TS-RAG (Ning et al., 2026) retrieves relevant segments and fuses them with foundation-model representations, and RAFT (Han et al., 2025) uses similar historical candidates and their observed future trajectories to support prediction. Unlike methods that primarily rely on similarity, representation relevance, or end-to-end retrieval objectives, TSDataModel explicitly uses validation forecasting gains to estimate whether a source window is likely to improve a target query.

3 Problem Formulation

3.1 Frozen forecasting under restricted target context

Let $\mathcal{D}_t$ be a target domain. A forecasting example consists of an observed history and its future,

$$q_t = x_{t-L:t-1} \in \mathbb{R}^L, \qquad y_t = x_{t:t+H-1} \in \mathbb{R}^H, \tag{2}$$

where L is the available target-history length and H is the prediction horizon. A frozen foundation forecaster F produces $\widehat{y}_t = F(q_t)$. Its parameters remain fixed throughout utility construction, model selection, and test inference.

We chronologically partition labeled target examples into disjoint fitting, selection, and test sets, denoted by $\mathcal{Q}_{\mathrm{fit}}$, $\mathcal{Q}_{\mathrm{sel}}$, and $\mathcal{Q}_{\mathrm{test}}$. Futures from the first two sets may supervise utility estimation and estimator selection, respectively. Futures in $\mathcal{Q}_{\mathrm{test}}$ are hidden until final evaluation. This split makes the target-label budget explicit and prevents test-set choice among utility estimators.

For a multivariate or wide-table dataset, we treat each channel as a univariate series. This permits retrieval across datasets that do not share variables or node identities, but intentionally excludes cross-channel dependence from the current problem scope.

3.2 External history retrieval

Let $\mathcal{S} = \{s_i\}_{i=1}^M$ be an archive constructed from one or more source domains, with each $s_i \in \mathbb{R}^L$. A candidate contains only an observed source history: its continuation, timestamp, and target-domain label are unavailable to the forecaster. A retrieval policy π maps a target history q to an ordered set of K distinct source histories, $\pi(q) = (s_{\pi_1}, \ldots, s_{\pi_K})$. Prediction with external context is

$$\widehat{y}_q^{\pi} = F([s_{\pi_1}; \cdots; s_{\pi_K}; q]), \tag{3}$$

where $[\cdot;\cdot]$ denotes temporal concatenation. Its input length is $(K+1)L$. When available, a genuine target history of this same length is a length-matched empirical reference. It contains more target information and is not assumed to be attainable in the restricted-context setting.

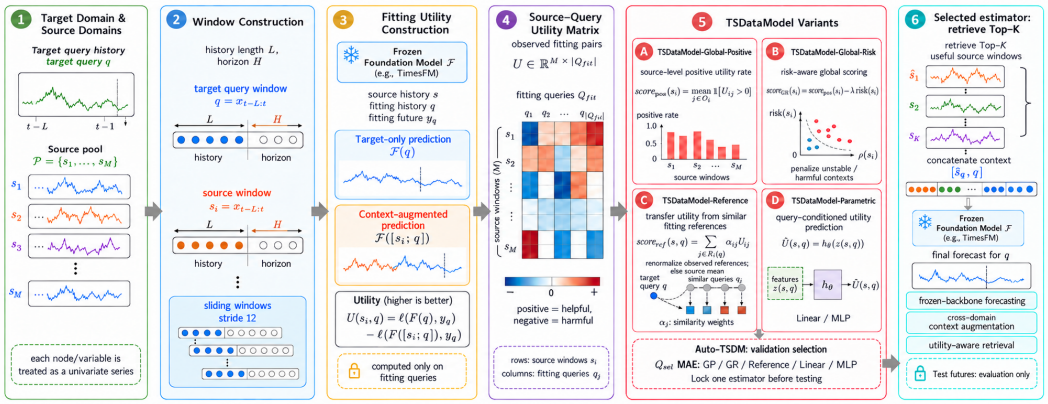


Figure 1: TSDataModel fits utility estimators on fitting queries, selects one by held-out forecasting MAE, and freezes that choice before testing. Source histories augment a frozen forecaster; test futures are used only for evaluation.

3.3 Forecasting utility and the learning objective

For $(q, y) \in \mathcal{Q}_{\text{fit}}$, the single-source forecasting utility is

$$u_F(s, q) = \ell(F(q), y) - \ell(F([s; q]), y). \tag{4}$$

Thus $u_F(s, q) > 0$ means that s reduces loss, whereas $u_F(s, q) < 0$ indicates negative transfer. Evaluating every pair in $\mathcal{S} \times \mathcal{Q}_{\text{fit}}$ may be expensive, so we observe a sparse index set Ω and learn a scorer $g_m(s, q)$ from $\{(s_i, q_j, u_F(s_i, q_j)) : (i, j) \in \Omega\}$. The induced policy ranks archive windows by g_m and retrieves its ordered top K.

Because the best dependence structure is unknown, TSDataModel considers a fixed candidate family $\mathcal{M}$ containing query-independent global, reference-based, and parametric scorers. It chooses the complete retrieval policy using only the selection set:

$$m^* = \arg\min_{m \in \mathcal{M}} \frac{1}{|\mathcal{Q}_{\text{sel}}|} \sum_{(q,y) \in \mathcal{Q}_{\text{sel}}} \ell(F([\pi_m(q); q]), y)\,. \tag{5}$$

After this choice, m^* and its fitted state are frozen. Test retrieval uses only source histories and the observed test history q; test futures are used solely to report performance. The objective is therefore not to predict y with a new model or to prove that external context is better than longer target history. It is to rank history-only contexts that improve a fixed forecaster when target context and weight access are restricted.

4 TSDataModel

TSDataModel learns which external histories help a frozen forecaster F (Figure 1). For source history s_i and fitting query (q_j, y_j), its utility is

$$U_{ij} = \ell_{\text{MAE}}(F(q_j), y_j) - \ell_{\text{MAE}}(F([s_i; q_j]), y_j). \tag{6}$$

We sample source–query pairs and retain unobserved utilities as missing. Let $\mathcal{O}_i$ contain the fitting queries with observed utility for source i; every source has at least one observation. Estimators rank sources from these observations, retrieve K distinct histories, and prepend them in score order to the target history. Labels measure single-source effects, so top-K concatenation is a heuristic, not a model of additive or jointly optimal utility.

4.1 Three Families of Utility Estimators

We distinguish three families by how source utility depends on the current query: global scores, reference-based estimates, and parametric predictions. All use observed fitting utilities with a frozen backbone.

Algorithm 1 Auto-TSDM with a frozen forecasting backbone

Require: Frozen F, source histories $\mathcal{S}$, target query pairs $\mathcal{Q}$, prefix count K
Split $\mathcal{Q}$ chronologically into $\mathcal{Q}_{\text{fit}}, \mathcal{Q}_{\text{sel}}, \mathcal{Q}_{\text{test}}$
Fix $\mathcal{M} = \{\text{GP}, \text{GR}, \text{Reference}, \text{Linear}, \text{MLP}\}$
Sample source–query pairs in $\mathcal{S} \times \mathcal{Q}_{\text{fit}}$, covering every source
Compute observed U_{ij} by equation 6; keep other pairs missing
for each $m \in \mathcal{M}$ **do**
    Fit estimator m using fitting histories and observed utilities only
    Evaluate its ordered top-K prefixes on $\mathcal{Q}_{\text{sel}}$ with F
end for
Select $\widehat{m}$ by equation 8; freeze it without refitting
for each test history q **do**
    Retrieve K distinct histories $\widehat{\pi}_{\widehat{m}}(q)$ from $\mathcal{S}$
    Predict $\widehat{y} = F([\widehat{\pi}_{\widehat{m}}(q); q])$; use test futures only for evaluation
end for

Global estimators (GP and GR). These estimators identify sources that help across fitting queries and produce fixed rankings at inference. GP learns a query-independent positive rate; GR additionally penalizes variability and negative normalized utility:

$$g_i = |\mathcal{O}_i|^{-1} \sum_{j \in \mathcal{O}_i} \mathbb{I}[U_{ij} > 0],$$
$$r_i = g_i - \lambda \left[\text{std}_{j \in \mathcal{O}_i}(\widetilde{U}_{ij}) + |\mathcal{O}_i|^{-1} \sum_{j \in \mathcal{O}_i} \max(-\widetilde{U}_{ij}, 0) \right], \tag{7}$$

where $\widetilde{U}_{ij} = U_{ij}/(\ell_{\text{MAE}}(F(q_j), y_j) + 10^{-6})$ and $\lambda = 0.5$.

Reference-based estimator (Reference). To adapt rankings without a parametric predictor, Reference retrieves five fitting histories most similar to the current history by cosine similarity and averages their utilities with temperature-0.1 softmax weights. Weights are renormalized over observed pairs for each source, falling back to its mean utility if none is observed. Similarity selects reference queries, while forecasting utility determines source scores.

Parametric estimators (Linear and MLP). These models learn linear and nonlinear feature interactions, respectively, to predict utility for the current source–query pair. They regress utility on $z(s, q) = [\phi(s), \phi(q), |\phi(s) - \phi(q)|, \phi(s) \odot \phi(q), d(s, q)]$, where ϕ contains history descriptors and d contains feature cosine similarity and Euclidean distance. Their feature standardization and regularized squared-error fitting use only observed fitting pairs. These query-dependent models need not outperform global scores. Appendix A gives complete features and optimization details.

4.2 Auto-TSDM: Validation-Based Selection

We separate target queries chronologically into fitting, selection, and test sets. On fitting-set utilities, we fit the fixed family $\mathcal{M} = \{\text{GP}, \text{GR}, \text{Reference}, \text{Linear}, \text{MLP}\}$. Writing $\widehat{\pi}_m(q)$ for the ordered top-K histories returned by candidate m, we select

$$\widehat{m} = \arg\min_{m \in \mathcal{M}} \frac{1}{|\mathcal{Q}_{\text{sel}}|} \sum_{(q,y) \in \mathcal{Q}_{\text{sel}}} \ell_{\text{MAE}}\big(F([\widehat{\pi}_m(q); q]), y\big). \tag{8}$$

Auto selects among the three families; it is not a fourth utility estimator. Selection uses actual multi-source forecasts, once per target, backbone, and seed. The chosen policy is frozen without refitting for all test queries and metrics. Its cost includes fitting and selecting all five candidates. Algorithm 1 summarizes fitting, selection, and inference.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate Solar-AL, PEMS04, PEMS-BAY, Bike-DC, Metro-NYC, and Weather using frozen TimesFM 2.5-200M and TimeMoE-50M (Das et al., 2023; Shi et al., 2025). Channels are treated as univariate series at their processed resolutions. A chronological 20/20/60% split supplies 256 fitting, 128 selection, and 512 common test queries per dataset, in total across channels. Per-channel standardization uses statistics computed exclusively from each dataset's first 20% chronological partition. Unless otherwise stated, normalized MAE is computed directly on this standardized target scale. The separate source archive contains 300 historical windows each from ETTh2, Exchange, and METR-LA (900 total), sampled from their first 20%; it excludes the targets and assumes no calendar synchronization.

Target history and horizon are both 12; $K = 3$ source histories of length 12 yield 48 input observations. Target-Only-48 receives genuine longer target history at the same origins. This is a restricted-context experiment, not a 12-observation cold start. Each backbone constructs its own approximately 39.7k utility pairs; candidate definitions and validation selections are fixed before testing. Main baselines appear in Table 1. The Naive RAF† (Zhang et al., 2025) has different access: it averages three archived history–continuation pairs, standardizes and aligns the result, and uses a 36-step input. Implementation details, including TS2Vec (Yue et al., 2022), are in Appendix B.

We report normalized MAE as mean$\pm$sample standard deviation over seeds 41–45. Queries are fixed while source sampling, utility-pair sampling, and initialization vary. Queries are fixed while source sampling, utility pairing, and initialization vary. Prespecified contrasts use joint seed/time-block bootstrap intervals, preserving cross-channel clustering; uncertainty details are in Appendix B.

5.2 MAIN RESULTS

Auto-TSDM achieves lower mean MAE than Target-Only-12 in 11 of 12 target–backbone conditions on the fixed 512-query main cohort. Under the same external-context budget, it outperforms Random-Prefix in all 12 conditions and each individual similarity-based baseline and the RAF-style comparator in at least 11 of 12. These results indicate that the improvement does not arise from adding context alone: forecasting utility provides a more reliable retrieval signal than random selection or similarity between observed histories. GP or GR is selected in 41 of 60 seed-level configurations, further showing that highly query-conditioned estimation is not always necessary. Instead, Auto-TSDM uses held-out forecasting performance to select the appropriate level of utility dependence, retaining transferable source-level utility when it is sufficient and using query-conditioned estimators only when validation supports the added complexity. These counts summarize five-seed mean performance; paired uncertainty results are reported separately.

Genuine longer target history provides an intentionally stronger reference and clarifies the intended deployment regime. Target-Only-48 has lower mean MAE than Auto-TSDM in 9 of 12 conditions, with PEMS-BAY, Bike-DC, and Metro-NYC favoring genuine history on both backbones. This result does not contradict the restricted-context setting: when 48 target observations are available, they should generally be preferred to retrieved external histories. Auto-TSDM instead provides a weight-update-free alternative when only a short target history is accessible. Notably, on Solar, Auto-TSDM reduces MAE relative to Target-Only-48 by 10.62% with TimesFM and 19.16% with TimeMoE, with positive adjusted joint-bootstrap lower bounds at all six block lengths. TimeMoE–Weather remains the main failure case, where augmentation increases MAE by 23.8% relative to Target-Only-12. Target-Only-48 is therefore an information-richer empirical reference rather than a theoretical upper bound.

Comparison with retrieval-augmented forecasting. Auto-TSDM obtains lower mean MAE than the RAF-style continuation comparator in 11 of 12 target–backbone conditions, despite the latter having access to archived source continuations. The native-backbone ARM-style adaptation inspired by TS-RAG is also competitive: its predicted-continuation variant has lower mean MAE than Auto-TSDM in 7 of 12 conditions, while the archived-continuation comparison is evenly split at 6 of 12. These comparisons involve different adaptation and information-access assumptions. ARM trains a forecast-side fusion module, uses a different frozen output head, and, in its archived variant,

Table 1: Normalized MAE over five seeds (mean±SD; lower is better). Bold/underline mark best/second-best unrounded means, excluding Target-Only-48. †RAF uses archived continuations and 36 inputs; other prefix methods use 48 history-only inputs. Formatting is descriptive, not significance.

Method	Solar	PEMS04	BAY	Bike	Metro	Weather
TimesFM 2.5-200M						
Target-12	<u>0.278</u>±0.000	<u>0.278</u>±0.000	0.715±0.000	0.899±0.000	1.126±0.000	<u>0.301</u>±0.000
Repeat-Prefix	0.424±0.000	0.357±0.000	1.061±0.000	1.097±0.000	1.458±0.000	0.439±0.000
Random-Prefix	0.301±0.006	0.311±0.010	0.733±0.006	0.916±0.015	1.086±0.006	0.320±0.012
Raw-Cosine	0.346±0.012	0.311±0.003	0.864±0.009	0.841±0.007	1.250±0.035	0.323±0.006
Raw-Euclidean	0.343±0.005	0.296±0.002	0.864±0.006	0.926±0.006	1.247±0.004	0.321±0.004
Feat.-Cosine	0.323±0.009	0.304±0.005	0.879±0.008	0.938±0.014	1.288±0.005	0.327±0.003
Feat.-Euclidean	0.320±0.004	0.294±0.006	0.868±0.012	0.916±0.018	1.259±0.003	0.319±0.007
DTW	0.353±0.004	0.303±0.003	0.772±0.006	0.861±0.016	1.185±0.011	0.317±0.003
MMD	0.330±0.004	0.296±0.001	<u>0.713</u>±0.021	<u>0.827</u>±0.010	1.086±0.004	0.314±0.008
TS2Vec	0.367±0.003	0.302±0.002	0.809±0.011	0.973±0.030	1.172±0.016	0.324±0.005
Backbone-Embed.	0.300±0.012	0.321±0.011	0.756±0.027	0.870±0.021	1.119±0.027	0.329±0.009
Naive RAF†	0.287±0.011	0.290±0.002	0.750±0.013	1.048±0.020	<u>1.039</u>±0.022	0.309±0.012
Auto-TSDM	**0.250**±0.009	**0.270**±0.024	**0.706**±0.048	**0.812**±0.026	**0.991**±0.063	**0.287**±0.009
Target-48 (ref.)	0.280±0.000	0.259±0.000	0.450±0.000	0.581±0.000	0.311±0.000	0.282±0.000
TimeMoE-50M						
Target-12	<u>0.338</u>±0.000	<u>0.306</u>±0.000	0.829±0.000	1.011±0.000	0.857±0.000	**0.359**±0.000
Repeat-Prefix	0.427±0.000	0.357±0.000	1.076±0.000	1.080±0.000	1.454±0.000	0.438±0.000
Random-Prefix	0.412±0.014	0.449±0.010	<u>0.740</u>±0.011	0.858±0.020	0.869±0.010	0.667±0.011
Raw-Cosine	0.422±0.021	0.338±0.005	0.911±0.020	0.924±0.015	1.106±0.081	0.566±0.009
Raw-Euclidean	0.372±0.002	0.309±0.002	0.868±0.005	0.964±0.004	1.037±0.017	0.400±0.021
Feat.-Cosine	0.361±0.008	0.334±0.007	0.945±0.019	1.004±0.010	1.092±0.029	0.431±0.012
Feat.-Euclidean	0.360±0.006	0.323±0.005	0.909±0.013	0.993±0.012	1.045±0.027	0.448±0.022
DTW	0.382±0.004	0.316±0.002	0.867±0.013	0.863±0.004	1.009±0.018	0.399±0.021
MMD	0.360±0.002	0.308±0.002	0.757±0.006	<u>0.788</u>±0.002	0.957±0.014	0.396±0.020
TS2Vec	0.375±0.003	0.313±0.002	0.871±0.010	1.028±0.031	0.994±0.038	0.406±0.015
Backbone-Embed.	0.415±0.053	0.439±0.005	0.825±0.031	0.865±0.046	0.928±0.030	0.703±0.010
Naive RAF†	0.346±0.012	0.356±0.005	0.872±0.013	1.128±0.046	<u>0.729</u>±0.016	<u>0.381</u>±0.015
Auto-TSDM	**0.266**±0.014	**0.302**±0.009	**0.639**±0.024	**0.746**±0.037	**0.658**±0.036	0.445±0.028
Target-48 (ref.)	0.328±0.000	0.372±0.000	0.490±0.000	0.658±0.000	0.292±0.000	0.360±0.000

accesses observed source continuations. Moreover, within ARM, predicted and archived continuations improve over the no-source control in only 7 of 12 and 2 of 12 conditions, respectively, whereas constant source slots improve it in 8 of 12. Consequently, ARM demonstrates the strength of representation-level adaptation, but its gains cannot be attributed solely to retrieved temporal information. In contrast, Auto-TSDM operates through the backbone's native frozen forecasting path, uses history-only source windows, and requires neither source futures nor a trainable forecast-side fusion module. Complete results and access assumptions are provided in Appendix C.4.

5.3 Matched Supervision and Computational Cost

Using identical target labels, we compare utility estimation, direct Ridge/XGBoost (Chen & Guestrin, 2016)/DLinear (Zeng et al., 2023) prediction from 12-point histories, and LoRA (Hu et al., 2021) adaptation. Pretraining, external-data access, and computation are not matched.

Global offers a competitive accuracy–preparation-cost trade-off under the current implementation and accounting boundaries: it beats LoRA in mean MAE in 5/12 conditions with substantially lower accounted preparation cost and comparable, slightly higher online cost (Table 2). Auto beats each direct predictor in MAE in 8/12 conditions, but LoRA beats Auto in 7/12. Thus, lower preparation cost among backbone methods is a practical benefit, not a claim of uniformly better accuracy.

Table 2: Matched supervision: 256 fit/128 selection/512 test queries. MAE: five-seed means. Costs (backbone rows only): six-target means, seed 41, RTX 3090; offline from standardized inputs, online ms/query at batch 128 including retrieval and prediction. †: measured-stage sums including within-family validation, excluding optional serialization. Global/Parametric select GP/GR or Linear/MLP. Direct predictors report MAE only. Details: Appendix C.8.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather	Off. (s)	On. (ms)
Direct Ridge	0.2783	0.2523	0.9258	0.8385	0.9240	0.3665	–	–
XGBoost	0.2460	0.2719	0.8553	0.8622	0.7813	0.5685	–	–
DLinear	0.2805	0.2511	0.7542	0.8394	0.9070	0.3664	–	–
TimesFM								
Global	0.2534	0.2642	0.6680	0.8123	1.0471	0.2871	36.68†	0.930
Reference	0.2627	0.2898	0.7641	0.8530	0.9637	0.3015	36.58†	1.181
Parametric	0.2435	0.3032	0.7343	0.8559	1.0261	0.2941	67.24†	5.632
Auto	0.2498	0.2705	0.7063	0.8123	0.9912	0.2871	67.75	0.974
LoRA	0.2398	0.2495	0.7349	0.9060	0.9154	0.2802	191.80	0.916
TimeMoE								
Global	0.2704	0.3320	0.6394	0.7430	0.6584	0.5090	79.43†	2.185
Reference	0.3795	0.4042	0.8026	0.8686	0.7121	0.5559	79.16†	2.474
Parametric	0.2825	0.3024	0.6765	0.7969	0.7390	0.4663	112.39†	7.048
Auto	0.2655	0.3016	0.6394	0.7465	0.6584	0.4451	113.47	4.478
LoRA	0.2259	0.2427	0.7597	0.7760	0.7603	0.2693	312.66	2.107

5.4 Utility Estimator Complexity

Global scores sources independently of queries; Reference uses similar fitting queries; Parametric learns source–query scores (Table 2). Auto selects across all five estimators on validation MAE, choosing GP/GR in 41/60 configurations; it fits all candidates offline but runs only the selected estimator online.

Parametric improves on Global in mean MAE in 4/12 conditions. Only TimeMoE–PEMS04 has positive exploratory paired 95% bounds at all six block lengths (0.33198 to 0.30238; Appendix C.12). Global is therefore a strong baseline: additional query-conditioned complexity should be justified by validation gains. For fixed MLP, shuffling labels worsens mean MAE in 12/12 conditions, yet shuffling source order and retraining improves it in 9/12; no condition passes all adjusted mechanism contrasts (Appendix C.4).

5.5 Retrieval Budget and Archive Composition

Number of retrieved windows. We vary $K \in \{1, 3, 5, 7\}$ with fixed utility models and validation-based reselection, using maximum context 96 without truncation. Auto has lower mean MAE than similarity in 42/48 target–backbone–K conditions and random retrieval in all 48. No universal optimum emerges (Figure 2), motivating validation-based choice of K; fixed-estimator controls separate retrieval-count effects from reselection (Appendix C.11).

Single- versus multi-dataset archives. On Solar, METR-LA-only raises MAE by 7.81%/23.75% versus multi-source (TimesFM/TimeMoE), supported across six exploratory interval settings (Appendix C.9). Composition matters at fixed size, but multi-source does not always win (Table 3); other differences lack consistent support. Budgets/availability: Appendices C.5, C.6.

Table 3: Solar archives: Auto MAE, five seeds, $K = 3$, 900 candidates (Multi: 300 per source).

Backbone	Multi	ETTh2 only	Exchange only	METR-LA only
TimesFM	0.2498	0.2539	0.2416	0.2694
TimeMoE	0.2655	0.2602	0.2670	0.3286

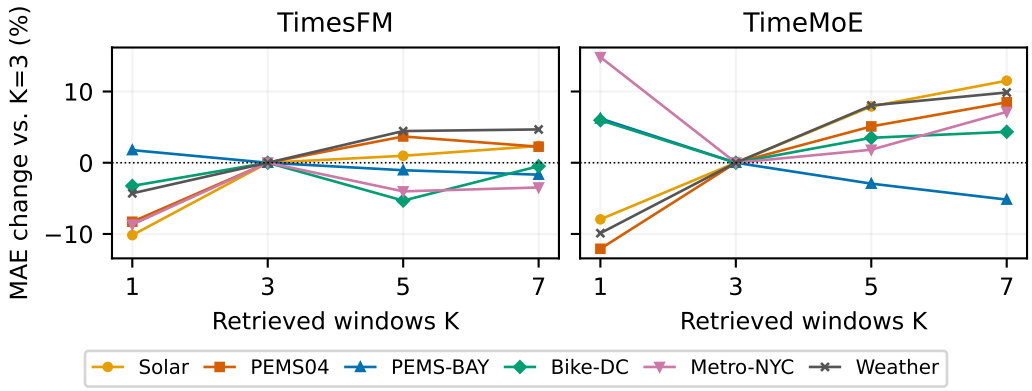


Figure 2: Auto: percentage MAE change from $K = 3$ (lower is better); five-seed means, maximum context 96. Uncertainty and controls: Appendix C.11.

5.6 Discussion

Auto-TSDM outperforms the strongest non-utility retrieval result in mean MAE in 11 of 12 target–backbone settings, indicating that forecasting feedback is more informative for context selection than similarity alone. However, greater estimator complexity is not always necessary: GP or GR is selected in 41 of 60 seed-level configurations, while the parametric family improves on the global family in only 4 of 12 mean comparisons. The main benefit of Auto-TSDM is therefore its ability to select an appropriate level of utility modeling using held-out forecasting performance.

The results also define the intended scope of the method. Target-48 outperforms Auto-TSDM in 9 of 12 settings, showing that retrieved contexts should not replace genuine target history when longer history is available. LoRA performs better in 7 of 12 settings, but requires weight access and substantially higher measured offline preparation cost. TSDataModel instead targets frozen or API-only forecasters with restricted target context and an available external archive.

Several limitations remain. Auto-TSDM cannot currently abstain from augmentation, and single-source utility does not model interactions among multiple retrieved windows. In addition, potential overlap between public benchmarks and TSFM pretraining corpora cannot always be ruled out; absolute performance should therefore be interpreted conditional on the selected frozen backbones, although all retrieval methods within a backbone share the same pretrained model. Future work will address abstention, joint utility modeling, retrieval scale, and multivariate contexts.

6 Conclusion and Future Work

We presented TSDataModel, a framework that retrieves external time-series histories by their measured effect on a frozen forecaster rather than by similarity alone. It builds sparse source–query utility supervision from loss reductions, supports global, reference-based, and parametric estimators, and uses chronological held-out selection to obtain Auto-TSDM without test-label leakage. Across six targets and two TSFMs, this approach is consistently stronger than non-utility retrieval under short target context, while the results also expose clear boundaries: genuine longer target histories and weight-based adaptation often remain preferable. Future work will add an abstention option, learn joint and diversity-aware utility for sets of windows, reduce labeling cost through active pair selection, scale retrieval through candidate generation and reranking, and extend utility modeling to multivariate contexts and broader forecast horizons.

AI Use Statement

The authors used generative AI tools to assist with language editing, manuscript restructuring, code review, and feedback on experimental design, analysis, and presentation. All AI-assisted text, code,

citations, mathematical statements, and experimental descriptions were manually reviewed and verified by the authors. The authors take full responsibility for the final content of the submission.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pp. 894–903. PMLR, 2017.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H Nguyen, Wesley M Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems*, 37:74147–74181, 2024.

Ben D Fulcher and Nick S Jones. hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell systems*, 5(5):527–531, 2017.

Ben D Fulcher, Max A Little, and Nick S Jones. Highly comparative time-series analysis: the empirical structure of time series and their methods. *Journal of the Royal Society Interface*, 10 (83), 2013.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2505.04163*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *International Conference on Machine Learning*, pp. 9525–9587. PMLR, 2022.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

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*, volume 2024, pp. 11116–11140, 2024a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*, 2024b.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 83982–84006, 2025.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. *Advances in Neural Information Processing Systems*, 38:163170–163199, 2026.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, et al. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International conference on learning representations*, volume 2025, pp. 34635–34667, 2025.

Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2411.08249*, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pp. 10842–10851. PMLR, 2020.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8980–8987, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Huanyu Zhang, Chang Xu, Yi-Fan Zhang, Zhang Zhang, Liang Wang, and Jiang Bian. Timeraf: Retrieval-augmented foundation model for zero-shot time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

A Detailed Method and Implementation

A.1 Overview

TSDataModel estimates which source histories improve a target query's forecast under a frozen forecaster. It ranks sources by target-supervised forecasting utility, complementing retrieval based only on historical similarity. As illustrated in Figure 1, its estimators range from query-independent source preferences to query-dependent utility prediction. Auto-TSDM turns this estimator family

into a single deployable procedure by selecting one fitted estimator on held-out target queries before test evaluation. It does not require choosing the best estimator from test results.

Let F denote the frozen forecaster, $\mathcal{P} = \{s_i\}_{i=1}^{M}$ a pool of source histories, and q a target history with future y_q. Labeled target queries are separated chronologically into fitting and selection sets, $\mathcal{Q}_{\text{fit}}$ and $\mathcal{Q}_{\text{sel}}$, followed by the test set $\mathcal{Q}_{\text{test}}$. Only fitting-set labels are used to construct utility observations. For a sampled pair (s_i, q_j), define

$$U_{ij} = U(s_i, q_j) = \ell_{\text{MAE}}(F(q_j), y_{q_j}) - \ell_{\text{MAE}}(F([s_i; q_j]), y_{q_j}), \tag{9}$$

where brackets denote temporal concatenation. Positive utility indicates a reduction in forecasting error. Source continuations are not included in these inputs. We evaluate a sampled set of pairs $\Omega \subseteq \{1, \ldots, M\} \times \{1, \ldots, |\mathcal{Q}_{\text{fit}}|\}$ and write $\mathcal{O}_i = \{j : (i, j) \in \Omega\}$ for the fitting queries with an observed utility for source s_i. Sampling ensures $|\mathcal{O}_i| > 0$ for every candidate source. Unobserved pairs remain missing rather than being assigned zero utility.

The estimators below share these observations but generalize them differently. Global-Positive and Global-Risk aggregate source-level preferences; Reference transfers observed utility through similar fitting histories; Linear and MLP predict utility from source–query features. At deployment, an estimator scores each source and retrieves K distinct histories in descending score order. Utility labels describe single-source augmentation, whereas deployment concatenates K sources; this independent ranking is a heuristic and does not assume that joint forecasting utility is additive.

A.2 TSDataModel-Global

Global estimators assign each source a query-independent score using its observed fitting utilities. Once fitted, they induce the same source ranking for all subsequent queries. Their simplicity makes them useful when the fitting budget is insufficient to estimate query-dependent effects reliably.

A.2.1 Global-Positive

The positive utility rate measures how often a source helps among its observed fitting pairs:

$$\text{score}_{\text{pos}}(s_i) = \frac{1}{|\mathcal{O}_i|} \sum_{j \in \mathcal{O}_i} \mathbb{I}[U_{ij} > \epsilon]. \tag{10}$$

We use $\epsilon = 0$ in the reported experiments. The denominator is the number of observed pairs for that source, not the total number of fitting queries. A source that helps a larger proportion of its observed queries receives a higher score, irrespective of the magnitude of a few unusually large gains.

A.2.2 Global-Risk

Positive coverage alone may overlook unstable or harmful outcomes. We normalize utility by the target-only forecasting error,

$$\widetilde{U}_{ij} = \frac{U_{ij}}{\ell_{\text{MAE}}(F(q_j), y_{q_j}) + \delta}, \tag{11}$$

and define

$$\text{risk}(s_i) = \text{std}_{j \in \mathcal{O}_i}(\widetilde{U}_{ij}) + \frac{1}{|\mathcal{O}_i|} \sum_{j \in \mathcal{O}_i} \max(-\widetilde{U}_{ij}, 0). \tag{12}$$

Here the standard deviation is computed with population normalization over the observed pairs. The two terms penalize utility variability and the mean magnitude of negative normalized utility. The resulting score is

$$\text{score}_{\text{risk}}(s_i) = \text{score}_{\text{pos}}(s_i) - \lambda \, \text{risk}(s_i). \tag{13}$$

We fix $\delta = 10^{-6}$ and $\lambda = 0.5$. At equal positive rates, the score favors sources with smaller estimated risk. Both global estimators require fitting-set utility observations, despite requiring no separate regression model. For reference, the observed source mean $\overline{U}_i = |\mathcal{O}_i|^{-1} \sum_{j \in \mathcal{O}_i} U_{ij}$ is retained as the GlobalMean diagnostic control; it is not an additional Auto-TSDM candidate.

A.3 TSDataModel-Reference

Reference transfers utility through similar fitting histories. For a selection or test history q, it retrieves

$$\mathcal{N}(q) = \mathrm{TopN}_{j:q_j \in \mathcal{Q}_{\mathrm{fit}}} \mathrm{sim}(q, q_j), \tag{14}$$

where similarity is cosine similarity between flattened, dataset-standardized histories, with 10^{-6} added to the product of norms. We use $N = 5$ and temperature $\tau = 0.1$. Because utility observations are sparse, the valid references depend on the source: $\mathcal{R}_i(q) = \mathcal{N}(q) \cap \mathcal{O}_i$. We renormalize weights over those references,

$$\alpha_{ij}(q) = \frac{\exp(\mathrm{sim}(q, q_j)/\tau)}{\sum_{r \in \mathcal{R}_i(q)} \exp(\mathrm{sim}(q, q_r)/\tau)}, \qquad j \in \mathcal{R}_i(q), \tag{15}$$

and score a source by

$$\mathrm{score}_{\mathrm{ref}}(s_i, q) = \begin{cases} \displaystyle\sum_{j \in \mathcal{R}_i(q)} \alpha_{ij}(q) U_{ij}, & \mathcal{R}_i(q) \neq \varnothing, \\ \overline{U}_i, & \mathcal{R}_i(q) = \varnothing. \end{cases} \tag{16}$$

This fallback avoids treating missing utilities as failures or zero gains. Reference uses fitting histories and their utility observations; the independent selection set is not added to its reference bank. No additional utility regressor is trained.

A.4 TSDataModel-Parametric

The parametric estimators predict utility from a history-only pair representation:

$$\widehat{U}_\theta(s, q) = h_\theta(z(s, q)), \tag{17}$$

$$z(s, q) = [\phi(s), \phi(q), |\phi(s) - \phi(q)|, \phi(s) \odot \phi(q), d(s, q)]. \tag{18}$$

In the reported implementation, ϕ combines nine summary statistics, autocorrelations at eight lags, four normalized spectral magnitudes, and the window-normalized history. For 12-step univariate windows this gives 33 features. The two-dimensional $d(s, q)$ contains cosine similarity and Euclidean distance between $\phi(s)$ and $\phi(q)$, yielding 134 pair features. Feature standardization is fitted only on observed fitting pairs. These predictors use numerical history descriptors, rather than pretrained-backbone embeddings or source futures.

We instantiate h_θ as either Ridge regression (Linear) or an MLP. Both learn from the same observed utility targets by minimizing a regularized squared-error objective,

$$\min_\theta \sum_{(i,j) \in \Omega} \big(h_\theta(z(s_i, q_j)) - U_{ij}\big)^2 + \rho \mathcal{R}(\theta), \tag{19}$$

where the regularization and loss scaling follow the corresponding implementation. Linear uses Ridge coefficient 1.0. The MLP has two ReLU hidden layers of width 128 and is trained for 100 epochs with Adam, learning rate 10^{-3}, and batch size 256; its scikit-learn L2 parameter is 10^{-4}. No pair-level early-stopping split or selection-set labels are used for fitting. These two fixed implementations enter Auto-TSDM as separate candidates; gradient-boosted and test-selected "best parametric" models are not used in its main candidate family.

For an unseen history q, candidate sources are scored by $\widehat{U}_\theta(s_i, q)$ and ranked using a stable order, with source index resolving exact ties. Their histories are concatenated in rank order before the target history. This allows query-dependent retrieval without target-future access, but does not guarantee improvement over a fitted global source preference; that distinction is evaluated empirically.

A.5 Auto-TSDM: Validation-Based Estimator Selection

The utility estimators above impose different assumptions: GP and GR learn source-level preferences, whereas Reference, Linear, and MLP allow query-dependent scoring. Their relative effectiveness need not be known for a new target dataset and forecasting backbone. We therefore define

Algorithm 2 Auto-TSDM: fitting, validation selection, and frozen evaluation

Require: Frozen forecaster F; source archive $\mathcal{P}$; labeled fitting and selection sets $\mathcal{Q}_{\mathrm{fit}}$, $\mathcal{Q}_{\mathrm{sel}}$; test histories $\mathcal{X}_{\mathrm{test}}$; retrieval budget K; fixed ordered candidate family $\mathcal{M}$
Ensure: Test predictions and selected estimator $\widehat{m}$
1: Sample source–query pairs from $\mathcal{P} \times \mathcal{Q}_{\mathrm{fit}}$ each fitting query $(\mathbf{x}_j, \mathbf{y}_j)$
2: $e_j \leftarrow \ell_{\mathrm{MAE}}(F(\mathbf{x}_j), \mathbf{y}_j)$ each sampled source s_i for this query
3: $U_{ij} \leftarrow e_j - \ell_{\mathrm{MAE}}(F([s_i; \mathbf{x}_j]), \mathbf{y}_j)$ each estimator $m \in \mathcal{M}$
4: Fit $\widehat{\pi}_m$ using fitting-set utility observations
5: $L_m \leftarrow 0$ each $(\mathbf{x}, \mathbf{y}) \in \mathcal{Q}_{\mathrm{sel}}$
6: Retrieve ordered top-K histories $S \leftarrow \widehat{\pi}_m(\mathbf{x})$
7: $L_m \leftarrow L_m + \ell_{\mathrm{MAE}}(F([S; \mathbf{x}]), \mathbf{y})/|\mathcal{Q}_{\mathrm{sel}}|$
8: $\widehat{m} \leftarrow \arg\min_{m \in \mathcal{M}} L_m$ {Resolve ties by fixed candidate order}
9: Lock $\widehat{m}$ and the fitted policies; do not refit on selection labels each test history $\mathbf{x} \in \mathcal{X}_{\mathrm{test}}$
10: $S \leftarrow \widehat{\pi}_{\widehat{m}}(\mathbf{x})$
11: Output $\widehat{\mathbf{y}}(\mathbf{x}) \leftarrow F([S; \mathbf{x}])$

Auto-TSDM as a validation-based selection procedure over this fixed estimator family, rather than reporting the estimator with the lowest test error.

Let $\mathcal{Q}_{\mathrm{fit}}$, $\mathcal{Q}_{\mathrm{sel}}$, and $\mathcal{Q}_{\mathrm{test}}$ be temporally separated training, selection, and test query sets. For a fixed backbone F and source archive, we first construct utility observations using only the future labels in $\mathcal{Q}_{\mathrm{fit}}$. We fit each candidate policy $\widehat{\pi}_m$, where $m \in \mathcal{M} = \{\mathrm{GP}, \mathrm{GR}, \mathrm{Reference}, \mathrm{Linear}, \mathrm{MLP}\}$. Each policy maps a query history to an ordered set of K source histories; its candidate definition and fitting hyperparameters are fixed before evaluating the selection set. We then select by downstream forecasting MAE:

$$\widehat{m} = \arg\min_{m \in \mathcal{M}} \frac{1}{|\mathcal{Q}_{\mathrm{sel}}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{Q}_{\mathrm{sel}}} \ell_{\mathrm{MAE}}(F([\widehat{\pi}_m(\mathbf{x}); \mathbf{x}]), \mathbf{y}) . \tag{20}$$

Here $[\widehat{\pi}_m(\mathbf{x}); \mathbf{x}]$ denotes the selected source histories concatenated with the target history. In particular, selection evaluates the actual K-source forecasting pipeline, rather than utility-regression fit or test error. Ties follow the fixed candidate order given above.

Selection is performed once per target dataset, backbone, and training seed. The selected policy is retained without refitting on the selection labels and is used for every test query and every reported metric:

$$\widehat{\mathbf{y}}(\mathbf{x}) = F([\widehat{\pi}_{\widehat{m}}(\mathbf{x}); \mathbf{x}]) . \tag{21}$$

The estimator identity $\widehat{m}$ is fixed during testing, although a query-dependent estimator can retrieve different sources for different histories. Test futures are used only to compute the final evaluation metrics: they are unavailable to utility fitting, estimator selection, and source retrieval. Auto-TSDM is therefore neither an ensemble nor a per-query or per-metric test-selected oracle. Its supervision budget includes both the fitting and selection queries, and its offline cost includes fitting and validating all candidate estimators. Results for individual estimators serve as separate diagnostic comparisons and are not substituted for Auto-TSDM after observing test performance.

Algorithm 2 summarizes this procedure. The total target supervision budget includes both fitting and selection queries, while the backbone remains frozen throughout.

B Detailed Setup and Evaluation

We evaluate frozen-backbone forecasting on six targets: Solar-AL, PEMS04, PEMS-BAY, Bike-DC, Metro-NYC, and Weather, spanning energy, traffic, mobility, and meteorology. Each channel is treated as a univariate series at its processed temporal resolution. Each dataset is split chronologically into 20% training, 20% validation, and 60% test data, with 256 fitting, 128 selection, and 512 test queries sampled in total across channels. All methods share the same queries. Per-channel standardization uses only the first 20% of each dataset, replacing standard deviations below 10^{-6} with one. The external archive contains 900 windows, with 300 sampled from the first 20% of each of ETTh2, Exchange, and METR-LA; none of the targets is included. These archives are assumed to be separately available historical resources, without calendar synchronization with the target.

We use TimesFM 2.5-200M[1] and TimeMoE-50M[2] from the TimesFM and Time-MoE families (Das et al., 2023; Shi et al., 2025), retaining their native point-forecasting paths. Target history and prediction horizon are both 12 steps. Prefix methods prepend three distinct 12-step source histories in descending score order, yielding 48 input observations; the sources need not come from different domains. Target-Only-48 receives 48 genuine target observations at the same forecast origin. All queries satisfy this longer-history requirement, so the experiment restricts predictor context rather than assuming only 12 raw observations exist for preprocessing. The original query grid has stride 12; the follow-up below checks all 12 origin phases. Equal step counts need not represent equal elapsed durations across datasets.

Auto-TSDM follows the selection procedure in Section 4.2. Utility labels are generated separately for each backbone using only the 256 fitting queries. For each query, a balanced set of 128 source candidates is combined with 16 nearest and 16 farthest candidates and deduplicated, yielding approximately 39.7k source–query pairs. Labels measure the MAE reduction from adding one source history: labeling uses 24-step inputs, whereas top-three inference uses 48 steps, without modeling joint three-source utility. GP, GR, Reference, Linear, and MLP are fitted on these observations; the MLP uses two width-128 hidden layers and 100 training epochs. For each target, backbone, and seed, validation MAE on the 128 selection queries determines one estimator. The selected policy is retained without refitting and used for all test queries and metrics. Code, models, and selections are locked before test evaluation. Five targets had been examined during earlier development, so these safeguards do not make the study a previously unseen confirmatory benchmark.

The main baselines are Target-Only-12, Repeat-Prefix, Random-Prefix, raw-value and feature-level cosine and Euclidean retrieval, DTW, MMD, TS2Vec (Yue et al., 2022), backbone-embedding retrieval, and a Naive RAF-style adaptation (Tire et al., 2024). Repeat-Prefix repeats target history four times; Random-Prefix samples three sources. MMD uses biased empirical RBF MMD^2 on scalar observations, selecting bandwidth from $\{0.1, 1, 10\}$ by validation MAE. TS2Vec learns 64-dimensional pooled source representations in 500 updates. Backbone-embedding retrieval uses squared Euclidean distance between last hidden representations. The RAF adaptation averages three archived history–continuation pairs of length $12 + 12$, standardizes and aligns the result, and prepends it to the target history. Unlike the history-only prefixes, it accesses archived source continuations and uses 36 total observations; it is not an unmodified official Chronos reproduction. Target-Only-48 is a separate longer-history reference. All main-table backbones remain frozen, although selectors and the TS2Vec encoder are learned. TSDM variants, supervision-matched prediction and adaptation, and learned fusion controls are evaluated separately.

We report normalized MAE as the primary metric and retain MSE as a secondary metric. Results are means and sample standard deviations over seeds 41–45, which vary source sampling, utility-pair sampling, and initialization while keeping queries fixed. Prespecified contrasts additionally use 5,000 paired bootstrap replicates that resample seeds and temporal blocks while preserving cross-channel clustering. Block lengths are 24, 72, 168, 512, 1,024, and 2,048 steps; adjusted 99.5833% intervals account for the 12 target–backbone conditions within each contrast, not all study comparisons. Settings with few nonempty blocks are interpreted cautiously. We use PyTorch on a single NVIDIA RTX 3090 with inference batch size 128. Offline labeling, fitting, and selection costs are reported separately in the computational-cost analysis.

B.1 Forecast-origin robustness

We further check the original stride-12 evaluation using 2,048 queries per target sampled from stride-one candidates covering all 12 origin phases, with the original trained models and validation choices held fixed. Auto improves over Target-Only-12 in 10/12 conditions and over Target-Only-48 in 3/12; its Solar gains over Target-Only-48 remain 8.35% with TimesFM and 20.90% with TimeMoE, with positive adjusted interval lower bounds across all six block lengths. TimeMoE–Weather remains a failure case. This follow-up supports the conditional nature of the findings while addressing evaluation-origin coverage. It uses different queries from the same previously studied test periods and does not revise the original training or validation choices; it is a robustness analysis, not independent new-data confirmation, and is not mixed into Table 1.

[1] https://huggingface.co/google/timesfm-2.5-200m-pytorch

[2] https://huggingface.co/Maple728/TimeMoE-50M

Table 4: Normalized MSE (mean ± sample standard deviation over five seeds; lower is better) on the same 512 test queries per target. Bold and underlined means denote the best and second-best results among the main methods, respectively, based on unrounded values; Target-Only-48 is excluded from both ranks and is shown only as a longer-target-history reference, not a theoretical bound. All main methods receive 12 target observations. History-only prefix methods use 48 total observations. †The Naive RAF-style adaptation instead uses 36 observations and accesses archived source continuations; its inclusion does not imply identical information access. Zero standard deviations reflect deterministic baselines on fixed queries. Formatting does not encode statistical significance.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM 2.5-200M						
Target-Only-12	0.2843 ± 0.0000	0.1571 ± 0.0000	2.3755 ± 0.0000	3.2054 ± 0.0000	2.1704 ± 0.0000	0.5414 ± 0.0000
Repeat-Prefix	0.5169 ± 0.0000	0.2536 ± 0.0000	4.0309 ± 0.0000	5.1688 ± 0.0000	3.3103 ± 0.0000	0.8577 ± 0.0000
Random-Prefix	0.2802 ± 0.0088	0.2017 ± 0.0143	2.3958 ± 0.0193	3.2832 ± 0.1000	2.0022 ± 0.0162	0.5609 ± 0.0337
Raw-Cosine	0.3758 ± 0.0251	0.2000 ± 0.0098	2.7829 ± 0.0319	**2.8799** ± 0.0475	2.3885 ± 0.1358	0.5568 ± 0.0233
Raw-Euclidean	0.3728 ± 0.0097	0.1817 ± 0.0044	2.8008 ± 0.0309	3.4152 ± 0.2002	2.3743 ± 0.0230	0.5674 ± 0.0168
Feature-Cosine	0.3325 ± 0.0146	0.1897 ± 0.0028	2.9171 ± 0.0931	3.3104 ± 0.1334	2.5287 ± 0.0373	0.5569 ± 0.0086
Feature-Euclidean	0.3270 ± 0.0128	0.1786 ± 0.0055	2.8464 ± 0.0660	3.1453 ± 0.1565	2.4403 ± 0.0202	0.5452 ± 0.0106
DTW	0.3878 ± 0.0104	0.1909 ± 0.0041	2.4891 ± 0.0302	3.2167 ± 0.2243	2.2100 ± 0.0175	0.5508 ± 0.0090
MMD	0.3571 ± 0.0078	0.1812 ± 0.0069	**2.3398** ± 0.0453	3.0962 ± 0.1302	1.9959 ± 0.0159	0.5594 ± 0.0206
TS2Vec	0.4135 ± 0.0068	0.1891 ± 0.0066	2.6270 ± 0.0511	3.4167 ± 0.1427	2.1826 ± 0.0523	0.5683 ± 0.0161
Backbone-Embedding Prefix	0.3147 ± 0.0158	0.2187 ± 0.0186	2.4491 ± 0.0588	3.2374 ± 0.2385	2.0880 ± 0.0721	0.5736 ± 0.0223
Naive RAF-style Adaptation†	0.3021 ± 0.0249	0.1705 ± 0.0021	2.4409 ± 0.0382	3.8792 ± 0.1985	1.9661 ± 0.0732	0.5624 ± 0.0476
Auto-TSDM (ours)	**0.2385** ± 0.0276	**0.1477** ± 0.0277	2.3409 ± 0.1203	3.0647 ± 0.0933	**1.7776** ± 0.1380	**0.5108** ± 0.0180
Target-Only-48 (reference)	0.2907 ± 0.0000	0.1401 ± 0.0000	1.7092 ± 0.0000	1.9944 ± 0.0000	0.3360 ± 0.0000	0.5626 ± 0.0000
TimeMoE-50M						
Target-Only-12	0.3849 ± 0.0000	0.1996 ± 0.0000	2.5995 ± 0.0000	3.0262 ± 0.0000	1.3882 ± 0.0000	**0.6404** ± 0.0000
Repeat-Prefix	0.5195 ± 0.0000	0.2538 ± 0.0000	4.1214 ± 0.0000	4.6813 ± 0.0000	3.2989 ± 0.0000	0.8480 ± 0.0000
Random-Prefix	0.3869 ± 0.0334	0.3669 ± 0.0158	2.2334 ± 0.0452	2.7344 ± 0.1394	1.4064 ± 0.0241	1.3296 ± 0.0526
Raw-Cosine	0.4489 ± 0.0246	0.2312 ± 0.0108	2.8285 ± 0.0863	2.7738 ± 0.0661	2.0331 ± 0.2734	0.9737 ± 0.0628
Raw-Euclidean	0.4241 ± 0.0017	0.1954 ± 0.0033	2.7128 ± 0.0501	2.9087 ± 0.0547	1.7685 ± 0.0323	0.6737 ± 0.0489
Feature-Cosine	0.3776 ± 0.0091	0.2201 ± 0.0075	2.9168 ± 0.0986	2.8338 ± 0.0343	1.9040 ± 0.0741	0.6930 ± 0.0349
Feature-Euclidean	0.3772 ± 0.0093	0.2048 ± 0.0040	2.8170 ± 0.0641	2.8355 ± 0.1048	1.7827 ± 0.0827	0.7504 ± 0.0539
DTW	0.4422 ± 0.0091	0.2032 ± 0.0038	2.5546 ± 0.0393	2.7620 ± 0.0417	1.6852 ± 0.0417	0.6691 ± 0.0416
MMD	0.4188 ± 0.0059	0.1983 ± 0.0032	2.3691 ± 0.0541	2.6396 ± 0.0285	1.6496 ± 0.0329	0.6784 ± 0.0352
TS2Vec	0.4322 ± 0.0045	0.1984 ± 0.0035	2.6346 ± 0.0473	2.9499 ± 0.1058	1.6853 ± 0.0954	0.6875 ± 0.0323
Backbone-Embedding Prefix	0.5331 ± 0.1112	0.3576 ± 0.0130	2.5370 ± 0.0756	2.7691 ± 0.1303	1.5148 ± 0.0909	1.5076 ± 0.0770
Naive RAF-style Adaptation†	0.4079 ± 0.0312	0.2632 ± 0.0075	2.9768 ± 0.3436	4.0751 ± 0.4555	1.1736 ± 0.0333	0.7300 ± 0.0658
Auto-TSDM (ours)	**0.2644** ± 0.0316	**0.1820** ± 0.0119	**2.1646** ± 0.0510	**2.5905** ± 0.1442	**0.9883** ± 0.0696	0.7784 ± 0.0872
Target-Only-48 (reference)	0.3459 ± 0.0000	0.2721 ± 0.0000	1.7124 ± 0.0000	1.5812 ± 0.0000	0.3132 ± 0.0000	0.6450 ± 0.0000

B.2 Secondary metric

C Additional Experimental Results

C.1 Study Protocols and Interpretation

The main study, the frozen-model origin-phase follow-up, the supervision-budget study, and the source-availability study use different evaluation cohorts. The main study uses 256 fitting, 128 selection, and 512 test queries per target; the frozen-model follow-up uses 2,048 test queries with the original models and choices unchanged. The budget study uses $N \in \{32, 64, 128, 256\}$ fitting queries, 128 selection queries, and 1,024 common test queries. The availability study uses 128 fitting queries, 128 selection queries, and a separate cohort of 1,024 common test queries. Query counts are totals across channels. These cohorts come from previously studied test periods and are not independent new-data confirmation. Follow-up results are not pooled with the main table.

Unless stated otherwise, table entries are normalized error means and sample standard deviations over seeds 41–45. Queries are shared across seeds. Zero standard deviations can arise for deterministic methods on these fixed queries; seed variability does not quantify uncertainty over test periods. All selections minimize selection-set MAE and remain unchanged when reporting MSE. The primary-study joint bootstrap resamples seeds and temporal blocks while preserving method pairing and cross-channel clustering, using 5,000 replicates and block lengths 24, 72, 168, 512, 1,024, and 2,048. The 99.5833% intervals adjust across 12 target–backbone conditions within each designated contrast, not across all comparisons in the paper. Budget and source-availability analyses

Table 5: Normalized MSE (mean ± sample standard deviation over five seeds; lower is better) on the same 512 test queries per target. Bold and underlined means denote the best and second-best results among the main methods, respectively, based on unrounded values; Target-Only-48 is excluded from both ranks and is shown only as a longer-target-history reference, not a theoretical bound. All main methods receive 12 target observations. History-only prefix methods use 48 total observations. †The Naive RAF-style adaptation instead uses 36 observations and accesses archived source continuations; its inclusion does not imply identical information access. Zero standard deviations reflect deterministic baselines on fixed queries. Formatting does not encode statistical significance.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM 2.5-200M						
Target-Only-12	0.2843 ± 0.0000	<u>0.1571</u> ± 0.0000	2.3755 ± 0.0000	3.2054 ± 0.0000	2.1704 ± 0.0000	<u>0.5414</u> ± 0.0000
Repeat-Prefix	0.5169 ± 0.0000	0.2536 ± 0.0000	4.0309 ± 0.0000	5.1688 ± 0.0000	3.3103 ± 0.0000	0.8577 ± 0.0000
Random-Prefix	<u>0.2802</u> ± 0.0088	0.2017 ± 0.0143	2.3958 ± 0.0193	3.2832 ± 0.1000	2.0022 ± 0.0162	0.5609 ± 0.0337
Raw-Cosine	0.3758 ± 0.0251	0.2000 ± 0.0098	2.7829 ± 0.0319	**2.8799** ± 0.0475	2.3885 ± 0.1358	0.5568 ± 0.0233
Raw-Euclidean	0.3728 ± 0.0097	0.1817 ± 0.0044	2.8008 ± 0.0309	3.4152 ± 0.2002	2.3743 ± 0.0230	0.5674 ± 0.0168
Feature-Cosine	0.3325 ± 0.0146	0.1897 ± 0.0028	2.9171 ± 0.0931	3.3104 ± 0.1334	2.5287 ± 0.0373	0.5569 ± 0.0086
Feature-Euclidean	0.3270 ± 0.0128	0.1786 ± 0.0055	2.8464 ± 0.0660	3.1453 ± 0.1565	2.4403 ± 0.0202	0.5452 ± 0.0106
DTW	0.3878 ± 0.0104	0.1909 ± 0.0041	2.4891 ± 0.0302	3.2167 ± 0.2243	2.2100 ± 0.0175	0.5508 ± 0.0090
MMD	0.3571 ± 0.0078	0.1812 ± 0.0069	**2.3398** ± 0.0453	3.0962 ± 0.1302	1.9959 ± 0.0159	0.5594 ± 0.0206
TS2Vec	0.4135 ± 0.0068	0.1891 ± 0.0066	2.6270 ± 0.0511	3.4167 ± 0.1427	2.1826 ± 0.0523	0.5683 ± 0.0161
Backbone-Embedding Prefix	0.3147 ± 0.0158	0.2187 ± 0.0186	2.4491 ± 0.0588	3.2374 ± 0.2385	2.0880 ± 0.0721	0.5736 ± 0.0223
Naive RAF-style Adaptation†	0.3021 ± 0.0249	0.1705 ± 0.0021	2.4409 ± 0.0382	3.8792 ± 0.1985	<u>1.9661</u> ± 0.0732	0.5624 ± 0.0476
Auto-TSDM (ours)	**0.2385** ± 0.0276	**0.1477** ± 0.0277	<u>2.3409</u> ± 0.1203	<u>3.0647</u> ± 0.0933	**1.7776** ± 0.1380	**0.5108** ± 0.0180
Target-Only-48 (reference)	0.2907 ± 0.0000	0.1401 ± 0.0000	1.7092 ± 0.0000	1.9944 ± 0.0000	0.3360 ± 0.0000	0.5626 ± 0.0000
TimeMoE-50M						
Target-Only-12	0.3849 ± 0.0000	0.1996 ± 0.0000	2.5995 ± 0.0000	3.0262 ± 0.0000	1.3882 ± 0.0000	**0.6404** ± 0.0000
Repeat-Prefix	0.5195 ± 0.0000	0.2538 ± 0.0000	4.1214 ± 0.0000	4.6813 ± 0.0000	3.2989 ± 0.0000	0.8480 ± 0.0000
Random-Prefix	0.3869 ± 0.0334	0.3669 ± 0.0158	<u>2.2334</u> ± 0.0452	2.7344 ± 0.1394	1.4064 ± 0.0241	1.3296 ± 0.0526
Raw-Cosine	0.4489 ± 0.0246	0.2312 ± 0.0108	2.8285 ± 0.0863	2.7738 ± 0.0661	2.0331 ± 0.2734	0.9737 ± 0.0628
Raw-Euclidean	0.4241 ± 0.0017	<u>0.1954</u> ± 0.0033	2.7128 ± 0.0501	2.9087 ± 0.0547	1.7685 ± 0.0323	0.6737 ± 0.0489
Feature-Cosine	0.3776 ± 0.0091	0.2201 ± 0.0075	2.9168 ± 0.0986	2.8338 ± 0.0343	1.9040 ± 0.0741	0.6930 ± 0.0349
Feature-Euclidean	<u>0.3772</u> ± 0.0093	0.2048 ± 0.0040	2.8170 ± 0.0641	2.8355 ± 0.1048	1.7827 ± 0.0827	0.7504 ± 0.0539
DTW	0.4422 ± 0.0091	0.2032 ± 0.0038	2.5546 ± 0.0393	2.7620 ± 0.0417	1.6852 ± 0.0417	<u>0.6691</u> ± 0.0416
MMD	0.4188 ± 0.0059	0.1983 ± 0.0032	2.3691 ± 0.0541	<u>2.6396</u> ± 0.0285	1.6496 ± 0.0329	0.6784 ± 0.0352
TS2Vec	0.4322 ± 0.0045	0.1984 ± 0.0035	2.6346 ± 0.0473	2.9499 ± 0.1058	1.6853 ± 0.0954	0.6875 ± 0.0323
Backbone-Embedding Prefix	0.5331 ± 0.1112	0.3576 ± 0.0130	2.5370 ± 0.0756	2.7691 ± 0.1303	1.5148 ± 0.0909	1.5076 ± 0.0770
Naive RAF-style Adaptation†	0.4079 ± 0.0312	0.2632 ± 0.0075	2.9768 ± 0.3436	4.0751 ± 0.4555	<u>1.1736</u> ± 0.0333	0.7300 ± 0.0658
Auto-TSDM (ours)	**0.2644** ± 0.0316	**0.1820** ± 0.0119	**2.1646** ± 0.0510	**2.5905** ± 0.1442	**0.9883** ± 0.0696	0.7784 ± 0.0872
Target-Only-48 (reference)	0.3459 ± 0.0000	0.2721 ± 0.0000	1.7124 ± 0.0000	1.5812 ± 0.0000	0.3132 ± 0.0000	0.6450 ± 0.0000

use exploratory 95% intervals at block lengths 24, 168, 512, and 2,048, without global multiplicity adjustment.

The original test cohort has 16 nonempty blocks for Solar and Weather, 12 for Bike-DC, six for Metro-NYC and PEMS04, and three for PEMS-BAY at block length 2,048. In the small-block-count settings, bootstrap intervals alone should not support strong claims. Complete interval records and block counts are retained in the accompanying evidence tables. The original fitting and selection grids each cover a single origin-modulo-12 phase; broadening test-origin coverage does not repair this training or selection restriction.

C.2 Secondary Forecasting Metric

Table 5 reports MSE for the frozen-backbone main comparisons using exactly the same models and validation-MAE choices as the primary MAE table. No estimator is reselected for MSE.

C.3 Supervision-Matched Prediction and Adaptation

Ridge, XGBoost, and DLinear predict directly from target history without a forecasting backbone. Direct Ridge selects its regularization coefficient from $\{0.1, 1, 10, 100\}$. XGBoost uses 200 trees, learning rate 0.05, squared-error training, and validation-MAE selection between maximum depths 2 and 4, with a separate regressor for each forecast step. Residual-Ridge learns a correction from target history and the frozen native forecast; its regularization coefficient is selected from $\{0.1, 1, 10, 100\}$. Linear probing fits a readout on frozen native hidden representations. LoRA updates only attention low-rank adapters with rank 8 and scaling parameter 16, using AdamW, MAE loss, gradient clipping

Table 6: Full supervision-matched results: normalized MAE, mean ± sample standard deviation across five seeds. Direct predictors are shared across backbones.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
Ridge	0.2783 ± 0.0000	0.2523 ± 0.0000	0.9258 ± 0.0000	0.8385 ± 0.0000	0.9240 ± 0.0000	0.3665 ± 0.0000
XGBoost	0.2460 ± 0.0000	0.2719 ± 0.0000	0.8553 ± 0.0000	0.8622 ± 0.0000	0.7813 ± 0.0000	0.5685 ± 0.0000
DLinear	0.2805 ± 0.0018	0.2511 ± 0.0008	0.7542 ± 0.0147	0.8394 ± 0.0054	0.9070 ± 0.0172	0.3664 ± 0.0040
TimesFM						
Target-Only-12	0.2780 ± 0.0000	0.2778 ± 0.0000	0.7152 ± 0.0000	0.8991 ± 0.0000	1.1261 ± 0.0000	0.3006 ± 0.0000
Auto-TSDM	0.2498 ± 0.0092	0.2705 ± 0.0236	0.7063 ± 0.0478	0.8123 ± 0.0261	0.9912 ± 0.0632	0.2871 ± 0.0090
MLP	0.2538 ± 0.0272	0.3040 ± 0.0121	0.7469 ± 0.0380	0.8443 ± 0.0274	1.0261 ± 0.0403	0.3155 ± 0.0117
Residual-Ridge	0.2540 ± 0.0000	0.2481 ± 0.0000	0.9299 ± 0.0000	0.8457 ± 0.0000	0.9436 ± 0.0000	0.3242 ± 0.0000
Linear probe	0.9229 ± 0.0000	0.3477 ± 0.0000	2.4687 ± 0.0000	1.3888 ± 0.0000	1.0724 ± 0.0000	1.0059 ± 0.0000
LoRA	0.2398 ± 0.0036	0.2495 ± 0.0019	0.7349 ± 0.0025	0.9060 ± 0.0004	0.9154 ± 0.0025	0.2802 ± 0.0007
TimeMoE						
Target-Only-12	0.3381 ± 0.0000	0.3061 ± 0.0000	0.8295 ± 0.0000	1.0109 ± 0.0000	0.8574 ± 0.0000	0.3595 ± 0.0000
Auto-TSDM	0.2655 ± 0.0135	0.3016 ± 0.0088	0.6394 ± 0.0243	0.7465 ± 0.0365	0.6584 ± 0.0357	0.4451 ± 0.0277
MLP	0.2825 ± 0.0252	0.3024 ± 0.0099	0.7169 ± 0.0208	0.7969 ± 0.0128	0.7390 ± 0.0236	0.4928 ± 0.0264
Residual-Ridge	0.2548 ± 0.0000	0.2451 ± 0.0000	0.9600 ± 0.0000	0.8602 ± 0.0000	0.9124 ± 0.0000	0.3413 ± 0.0000
Linear probe	0.5623 ± 0.0000	0.2878 ± 0.0000	2.6080 ± 0.0000	1.3120 ± 0.0000	1.1924 ± 0.0000	0.8037 ± 0.0000
LoRA	0.2259 ± 0.0067	0.2427 ± 0.0017	0.7597 ± 0.0087	0.7760 ± 0.0094	0.7603 ± 0.0124	0.2693 ± 0.0033

Table 7: Secondary MSE for adaptation and key utility controls on the original 512 test queries. Means ± sample standard deviations use the same validation-MAE selections as the MAE analysis.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM						
Auto-TSDM	0.2385 ± 0.0276	0.1477 ± 0.0277	2.3409 ± 0.1203	3.0647 ± 0.0933	1.7776 ± 0.1380	0.5108 ± 0.0180
GP	0.3321 ± 0.1372	0.1904 ± 0.0789	2.2608 ± 0.0628	3.1167 ± 0.1381	1.9676 ± 0.0565	0.5395 ± 0.0388
GR	0.2547 ± 0.0164	0.1395 ± 0.0098	2.2735 ± 0.0956	3.1697 ± 0.1328	1.8879 ± 0.0481	0.5203 ± 0.0279
MLP	0.2118 ± 0.0218	0.2054 ± 0.0290	2.3878 ± 0.1138	3.0567 ± 0.2310	1.8537 ± 0.1223	0.5538 ± 0.0282
Within-source label shuffle	0.2778 ± 0.0225	0.2164 ± 0.0303	2.4847 ± 0.0833	3.7801 ± 0.5816	2.1177 ± 0.1113	0.5573 ± 0.0238
Residual-Ridge	0.1720 ± 0.0000	0.1217 ± 0.0000	2.9400 ± 0.0000	2.6269 ± 0.0000	1.8029 ± 0.0000	0.5452 ± 0.0000
Linear probe	5.2628 ± 0.0000	0.3576 ± 0.0000	20.7640 ± 0.0000	6.5733 ± 0.0000	2.4327 ± 0.0000	7.9641 ± 0.0000
LoRA	0.2201 ± 0.0074	0.1312 ± 0.0022	2.4167 ± 0.0088	3.2537 ± 0.0022	1.7327 ± 0.0152	0.5127 ± 0.0012
TimeMoE						
Auto-TSDM	0.2644 ± 0.0316	0.1820 ± 0.0119	2.1646 ± 0.0510	2.5905 ± 0.1442	0.9883 ± 0.0696	0.7784 ± 0.0872
GP	0.3232 ± 0.0774	0.2603 ± 0.0215	2.1631 ± 0.0529	2.6159 ± 0.0665	1.0833 ± 0.1479	0.9565 ± 0.2322
GR	0.3575 ± 0.1592	0.1958 ± 0.0174	2.1638 ± 0.0281	2.5693 ± 0.1479	1.0473 ± 0.0690	1.0489 ± 0.2749
MLP	0.2660 ± 0.0278	0.1885 ± 0.0123	2.3099 ± 0.0451	2.8060 ± 0.0770	1.1434 ± 0.0702	0.9385 ± 0.0848
Within-source label shuffle	0.3936 ± 0.0166	0.3432 ± 0.0423	2.3793 ± 0.0647	2.9114 ± 0.1580	1.6752 ± 0.2704	1.2789 ± 0.0695
Residual-Ridge	0.1671 ± 0.0000	0.1197 ± 0.0000	3.2905 ± 0.0000	2.5928 ± 0.0000	1.6918 ± 0.0000	0.5628 ± 0.0000
Linear probe	1.6621 ± 0.0000	0.3354 ± 0.0000	36.4512 ± 0.0000	6.0346 ± 0.0000	3.6869 ± 0.0000	8.3388 ± 0.0000
LoRA	0.2078 ± 0.0092	0.1251 ± 0.0016	2.5571 ± 0.0327	2.7863 ± 0.0588	1.2077 ± 0.0215	0.5105 ± 0.0211

at one, and learning rates $\{3 \times 10^{-5}, 10^{-4}\}$. Validation selects among epochs 10, 25, and 50. DLinear uses moving-average width 25, Adam with learning rate 10^{-3}, MSE training loss, and validation-MAE selection among epochs 25, 50, and 100. Parameter access and model capacity differ even though target supervision is matched.

C.4 Utility Variants and Mechanism Controls

GP, GR, Reference, Linear, and MLP are the five Auto candidates. GlobalMean is an internal control and is not included in Auto's candidate set. The source-only MLP retains the two width-128 hidden layers and sets query features to zero. The additive prior fits source and query fixed effects. Within-source label shuffling preserves the source identities, exposure counts, and label multisets; both MLPs use float32 training features and labels, identical initialization, and 100 training epochs. It disrupts observed source–query label associations while retaining their sparse observation mask. Consequently, this control is informative about label pairing but is not a complete causal removal of query information.

The source-time-shuffle control permutes values within source histories and retrains the estimator. It changes both the source representation and the concatenation boundary, so it does not isolate temporal order from every other numerical effect. Its lower mean MLP error in 9/12 conditions limits a mechanism explanation requiring intact source trajectories. At block length 168, exploratory 95%

Table 8: Utility estimators and controls on the original test queries: normalized MAE, mean ± sample standard deviation. These are fixed variants or internal controls, not test-selected replacements for Auto.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
			TimesFM			
Auto-TSDM	0.2498 ± 0.0092	0.2705 ± 0.0236	0.7063 ± 0.0478	0.8123 ± 0.0261	0.9912 ± 0.0632	0.2871 ± 0.0090
GlobalMean	0.3020 ± 0.0618	0.3170 ± 0.0528	0.7489 ± 0.1024	0.7890 ± 0.0465	1.1099 ± 0.0419	0.3825 ± 0.1220
GP	0.3314 ± 0.1387	0.3136 ± 0.0572	0.6693 ± 0.0351	0.8227 ± 0.0284	1.0756 ± 0.0300	0.3245 ± 0.0664
GR	0.2534 ± 0.0096	0.2658 ± 0.0147	0.6725 ± 0.0369	0.8587 ± 0.0520	1.0463 ± 0.0190	0.2900 ± 0.0088
Reference	0.2627 ± 0.0133	0.2898 ± 0.0067	0.7641 ± 0.0601	0.8530 ± 0.0192	0.9637 ± 0.0518	0.3015 ± 0.0075
Linear	0.2621 ± 0.0226	0.3339 ± 0.0227	0.7532 ± 0.0352	0.8699 ± 0.0253	1.0999 ± 0.0568	0.2941 ± 0.0047
MLP	0.2538 ± 0.0272	0.3040 ± 0.0121	0.7469 ± 0.0380	0.8443 ± 0.0274	1.0261 ± 0.0403	0.3155 ± 0.0117
Within-source label shuffle	0.2881 ± 0.0181	0.3144 ± 0.0093	0.7697 ± 0.0171	0.9484 ± 0.0313	1.1199 ± 0.0379	0.3238 ± 0.0042
Source-only MLP	0.2834 ± 0.0293	0.3025 ± 0.0442	0.7412 ± 0.0839	0.8569 ± 0.1150	1.1288 ± 0.0442	0.3043 ± 0.0199
Additive prior	0.2689 ± 0.0355	0.2792 ± 0.0308	0.7573 ± 0.0798	0.8184 ± 0.0521	1.0912 ± 0.0415	0.2996 ± 0.0171
			TimeMoE			
Auto-TSDM	0.2655 ± 0.0135	0.3016 ± 0.0088	0.6394 ± 0.0243	0.7465 ± 0.0365	0.6584 ± 0.0357	0.4451 ± 0.0277
GlobalMean	0.3938 ± 0.0754	0.3798 ± 0.0510	0.6638 ± 0.0355	0.8162 ± 0.0252	0.8877 ± 0.0625	0.6834 ± 0.1828
GP	0.2953 ± 0.0544	0.3879 ± 0.0136	0.6398 ± 0.0243	0.7623 ± 0.0389	0.7152 ± 0.0648	0.5404 ± 0.0846
GR	0.3078 ± 0.0725	0.3320 ± 0.0163	0.6447 ± 0.0263	0.7811 ± 0.0398	0.6921 ± 0.0606	0.5768 ± 0.0992
Reference	0.3795 ± 0.0267	0.4042 ± 0.0258	0.8026 ± 0.0191	0.8686 ± 0.0427	0.7121 ± 0.0378	0.5559 ± 0.0301
Linear	0.3698 ± 0.0398	0.3854 ± 0.0126	0.6760 ± 0.0231	0.8293 ± 0.0279	0.7504 ± 0.0433	0.4547 ± 0.0209
MLP	0.2825 ± 0.0252	0.3024 ± 0.0099	0.7169 ± 0.0208	0.7969 ± 0.0128	0.7390 ± 0.0236	0.4928 ± 0.0264
Within-source label shuffle	0.3553 ± 0.0172	0.4303 ± 0.0223	0.7416 ± 0.0440	0.8571 ± 0.0312	0.9480 ± 0.0891	0.6421 ± 0.0317
Source-only MLP	0.4545 ± 0.1649	0.3475 ± 0.0212	0.6462 ± 0.0156	0.9013 ± 0.2148	0.9220 ± 0.0591	0.5733 ± 0.0877
Additive prior	0.3169 ± 0.0182	0.3374 ± 0.0286	0.6368 ± 0.0125	0.7805 ± 0.0347	0.8709 ± 0.0721	0.5328 ± 0.0803

Table 9: Fusion and prefix diagnostics on the original test queries: normalized MAE, mean ± sample standard deviation. ARM uses the frozen direct head rather than the native forecasting path.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
			TimesFM			
Frozen direct head	0.2811 ± 0.0000	0.2781 ± 0.0000	0.7127 ± 0.0000	0.9115 ± 0.0000	1.1308 ± 0.0000	0.3037 ± 0.0000
ARM: no source	0.2388 ± 0.0012	0.2536 ± 0.0036	0.8219 ± 0.0121	0.8931 ± 0.0125	0.9592 ± 0.0293	0.2797 ± 0.0028
ARM: constant slots	0.2496 ± 0.0066	0.2583 ± 0.0084	0.8408 ± 0.0188	0.8614 ± 0.0200	0.8905 ± 0.0229	0.2844 ± 0.0126
ARM: predicted continuation	0.2451 ± 0.0075	0.2507 ± 0.0010	0.8554 ± 0.0101	0.8746 ± 0.0107	0.9244 ± 0.0187	0.2760 ± 0.0061
ARM: archived continuation	0.2451 ± 0.0067	0.2672 ± 0.0120	0.8788 ± 0.0185	0.9277 ± 0.0306	0.9501 ± 0.0298	0.3011 ± 0.0128
MLP	0.2538 ± 0.0272	0.3040 ± 0.0121	0.7469 ± 0.0380	0.8443 ± 0.0274	1.0261 ± 0.0403	0.3155 ± 0.0117
Time-shuffled MLP	0.2359 ± 0.0064	0.2625 ± 0.0059	0.7250 ± 0.0216	0.8430 ± 0.0225	1.0217 ± 0.0431	0.3072 ± 0.0151
Time-shuffled GlobalMean	0.2899 ± 0.1021	0.2661 ± 0.0072	0.7181 ± 0.0468	0.8022 ± 0.0212	1.1227 ± 0.0262	0.2937 ± 0.0244
Own-past retrieval	0.3928 ± 0.0000	0.3034 ± 0.0000	0.7710 ± 0.0000	0.8849 ± 0.0003	1.2275 ± 0.0000	0.3311 ± 0.0000
			TimeMoE			
Frozen direct head	0.3386 ± 0.0000	0.3041 ± 0.0000	0.8307 ± 0.0000	0.9828 ± 0.0000	0.8568 ± 0.0000	0.3602 ± 0.0000
ARM: no source	0.2239 ± 0.0041	0.2491 ± 0.0004	0.8371 ± 0.0069	0.8713 ± 0.0047	0.9513 ± 0.0185	0.2761 ± 0.0024
ARM: constant slots	0.2194 ± 0.0011	0.2469 ± 0.0005	0.7967 ± 0.0101	0.8562 ± 0.0051	0.9435 ± 0.0087	0.2729 ± 0.0006
ARM: predicted continuation	0.2222 ± 0.0038	0.2483 ± 0.0013	0.9178 ± 0.0053	0.8824 ± 0.0046	0.9515 ± 0.0178	0.2702 ± 0.0013
ARM: archived continuation	0.2588 ± 0.0389	0.2805 ± 0.0099	0.9233 ± 0.0930	0.8889 ± 0.0166	0.9363 ± 0.0535	0.3073 ± 0.0220
MLP	0.2825 ± 0.0252	0.3024 ± 0.0099	0.7169 ± 0.0208	0.7969 ± 0.0128	0.7390 ± 0.0236	0.4928 ± 0.0264
Time-shuffled MLP	0.2549 ± 0.0156	0.3063 ± 0.0103	0.7402 ± 0.0289	0.7850 ± 0.0173	0.7401 ± 0.0381	0.4474 ± 0.0159
Time-shuffled GlobalMean	0.2964 ± 0.0416	0.3269 ± 0.0283	0.7021 ± 0.0452	0.8551 ± 0.1227	0.8929 ± 0.0800	0.5021 ± 0.0307
Own-past retrieval	0.4077 ± 0.0000	0.3424 ± 0.0001	0.8285 ± 0.0000	0.9080 ± 0.0012	0.9599 ± 0.0000	0.5062 ± 0.0002

joint intervals favor the shuffled MLP on TimesFM–PEMS04, TimeMoE–Solar, and TimeMoE–Weather; these comparisons are not globally corrected.

Our adaptation of the TS-RAG Adaptive Retrieval Mixer (ARM) (Ning et al., 2026) uses frozen native hidden representations and a frozen direct output head with a trainable fusion module. We compare no retrieved source, three constant source slots, three sources with model-predicted continuations, and three sources with archived continuations. The last variant has access to additional source information. The constant-slot control retains the slot count and trainable structure, while the no-source control tests target-only calibration. We use AdamW with learning rate 3×10^{-5} and MAE loss, selecting epochs 25, 50, or 100 by validation MAE. This is a native-backbone adaptation of retrieval fusion, not an unmodified official pretrained retrieval model. Own-past retrieval separately selects three history windows from the target channel's training archive by cosine similarity.

A descriptive Solar stratification further tests whether prefix gains arise mainly on constant target histories. In the original 512-query cohort, 252 histories satisfy $\max(x) = \min(x)$. On this subset, Target-Only-12/MLP MAE is 0.07516/0.11025 for TimesFM and 0.07979/0.10516 for TimeMoE. On the nonconstant subset, it is 0.47456/0.39291 and 0.58840/0.45433, respectively. Thus the fixed

Table 10: Budget study at $N = 32$ fitting queries plus 128 selection queries: normalized MAE, mean $\pm$ sample standard deviation. All budgets share the same 1,024 test queries.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM						
Target-Only-12	0.2831 ± 0.0000	0.2880 ± 0.0000	0.7096 ± 0.0000	0.8509 ± 0.0000	0.9583 ± 0.0000	0.3390 ± 0.0000
Target-Only-48	0.2787 ± 0.0000	0.2729 ± 0.0000	0.4696 ± 0.0000	0.6491 ± 0.0000	0.2786 ± 0.0000	0.3147 ± 0.0000
Auto-TSDM	0.2518 ± 0.0100	0.2843 ± 0.0119	0.7047 ± 0.0382	0.8269 ± 0.0101	0.8964 ± 0.0380	0.3482 ± 0.0121
LoRA	0.2538 ± 0.0003	0.2837 ± 0.0004	0.7056 ± 0.0005	0.8502 ± 0.0004	0.8924 ± 0.0020	0.3388 ± 0.0004
Residual-Ridge	0.3115 ± 0.0000	0.2768 ± 0.0000	0.8824 ± 0.0000	0.8863 ± 0.0000	0.8090 ± 0.0000	0.3774 ± 0.0000
GlobalMean	0.2753 ± 0.0222	0.2918 ± 0.0191	0.6859 ± 0.0340	0.8341 ± 0.0493	0.9852 ± 0.0298	0.3560 ± 0.0134
GP	0.2659 ± 0.0058	0.3118 ± 0.0260	0.7245 ± 0.0175	0.8330 ± 0.0311	0.9079 ± 0.0402	0.3581 ± 0.0126
GR	0.3106 ± 0.1021	0.3313 ± 0.0737	0.6994 ± 0.0322	0.8618 ± 0.0314	0.9398 ± 0.0143	0.3669 ± 0.0351
Reference	0.2747 ± 0.0392	0.3109 ± 0.0199	0.7378 ± 0.0362	0.8408 ± 0.0131	0.9160 ± 0.0304	0.3531 ± 0.0182
Linear	0.2555 ± 0.0141	0.3246 ± 0.0301	0.7470 ± 0.0243	0.9515 ± 0.0108	0.9982 ± 0.0198	0.3865 ± 0.0132
MLP	0.2460 ± 0.0070	0.3264 ± 0.0073	0.7327 ± 0.0152	0.8827 ± 0.0362	0.9345 ± 0.0206	0.3744 ± 0.0144
TimeMoE						
Target-Only-12	0.3559 ± 0.0000	0.3145 ± 0.0000	0.7571 ± 0.0000	0.9639 ± 0.0000	0.8677 ± 0.0000	0.4044 ± 0.0000
Target-Only-48	0.3359 ± 0.0000	0.3712 ± 0.0000	0.5048 ± 0.0000	0.7695 ± 0.0000	0.2641 ± 0.0000	0.4053 ± 0.0000
Auto-TSDM	0.3267 ± 0.0461	0.3165 ± 0.0154	0.6570 ± 0.0203	0.8153 ± 0.0239	0.7247 ± 0.0216	0.5439 ± 0.0465
LoRA	0.2579 ± 0.0022	0.2928 ± 0.0036	0.7096 ± 0.0018	0.8705 ± 0.0072	0.8328 ± 0.0014	0.3564 ± 0.0025
Residual-Ridge	0.2980 ± 0.0000	0.2743 ± 0.0000	0.8948 ± 0.0000	0.8940 ± 0.0000	0.8150 ± 0.0000	0.3689 ± 0.0000
GlobalMean	0.3354 ± 0.0924	0.4310 ± 0.1710	0.6926 ± 0.0439	0.8210 ± 0.0309	0.7714 ± 0.0248	0.6831 ± 0.1209
GP	0.3915 ± 0.0457	0.3491 ± 0.0364	0.7433 ± 0.0757	0.8096 ± 0.0182	0.8802 ± 0.0764	0.5702 ± 0.1052
GR	0.4037 ± 0.0413	0.4188 ± 0.0338	0.6570 ± 0.0203	0.8250 ± 0.0193	0.8611 ± 0.1003	0.6079 ± 0.1000
Reference	0.3892 ± 0.1400	0.3673 ± 0.0512	0.7181 ± 0.0239	0.8728 ± 0.0156	0.7471 ± 0.0159	0.7492 ± 0.1221
Linear	0.3640 ± 0.0270	0.3552 ± 0.0135	0.6958 ± 0.0203	0.9164 ± 0.0148	0.7840 ± 0.0226	0.5905 ± 0.0337
MLP	0.3705 ± 0.0384	0.3229 ± 0.0117	0.7296 ± 0.0083	0.8394 ± 0.0125	0.7247 ± 0.0216	0.5753 ± 0.0498

MLP's Solar improvement is not concentrated in constant histories. This stratification concerns the fixed MLP, not Auto, and does not establish a causal mechanism. Native normalization and nonnegativity-dependent output handling can also interact with a prefix.

C.5 Full Supervision-Budget Results

Nested fitting subsets use the fixed bit-reversal order of the original 256 queries, ensuring that small subsets do not simply take the first few sorted channels. Utility training uses only labels belonging to the selected queries; repeated source–query pairs are not independent target labels. Every candidate source has at least four observed labels at $N = 32$. All budgets additionally use the same 128 selection queries. LoRA retains the rank, learning-rate grid, epoch schedule, and selection checkpoints, but the number of optimizer steps changes with N. This is a supervision-budget comparison rather than a compute-matched comparison.

Across the 48 Auto–LoRA budget comparisons, all four exploratory block-length intervals favor Auto in 11 conditions and favor LoRA in 17; at least one interval includes zero in the remaining conditions. This sensitivity summary is not a multiplicity-corrected win rate. In particular, TimeMoE–Bike-DC changes ranking as the budget grows, so choosing a favorable test budget would change the estimand and overstate deployment performance.

C.6 Same-Dataset and Mixed Archives

Here "same-dataset" means that the archive comes from the target dataset, potentially including other channels; it does not mean only the current target channel or its immediately preceding history. "Cross-dataset" excludes the target dataset but may include another dataset from the same semantic domain. All pools contain 900 distinct candidates. Cross uses 300 windows each from ETTh2, Exchange, and METR-LA. Same uses 900 windows from the first 5% of the target dataset. Mixed uses 450 Same windows and 150 from each external dataset. The same-dataset archive precedes every fitting query's history. All three settings share 128 fitting queries from the subsequent training interval, 128 selection queries, and 1,024 test queries, but utility labels and estimators are rebuilt separately for each pool.

Same-dataset access can help: relative to Cross, Same reduces Auto MAE on Metro-NYC by 8.24% for TimesFM and 3.54% for TimeMoE, while Mixed reduces it by 2.66% and 3.90%. All four

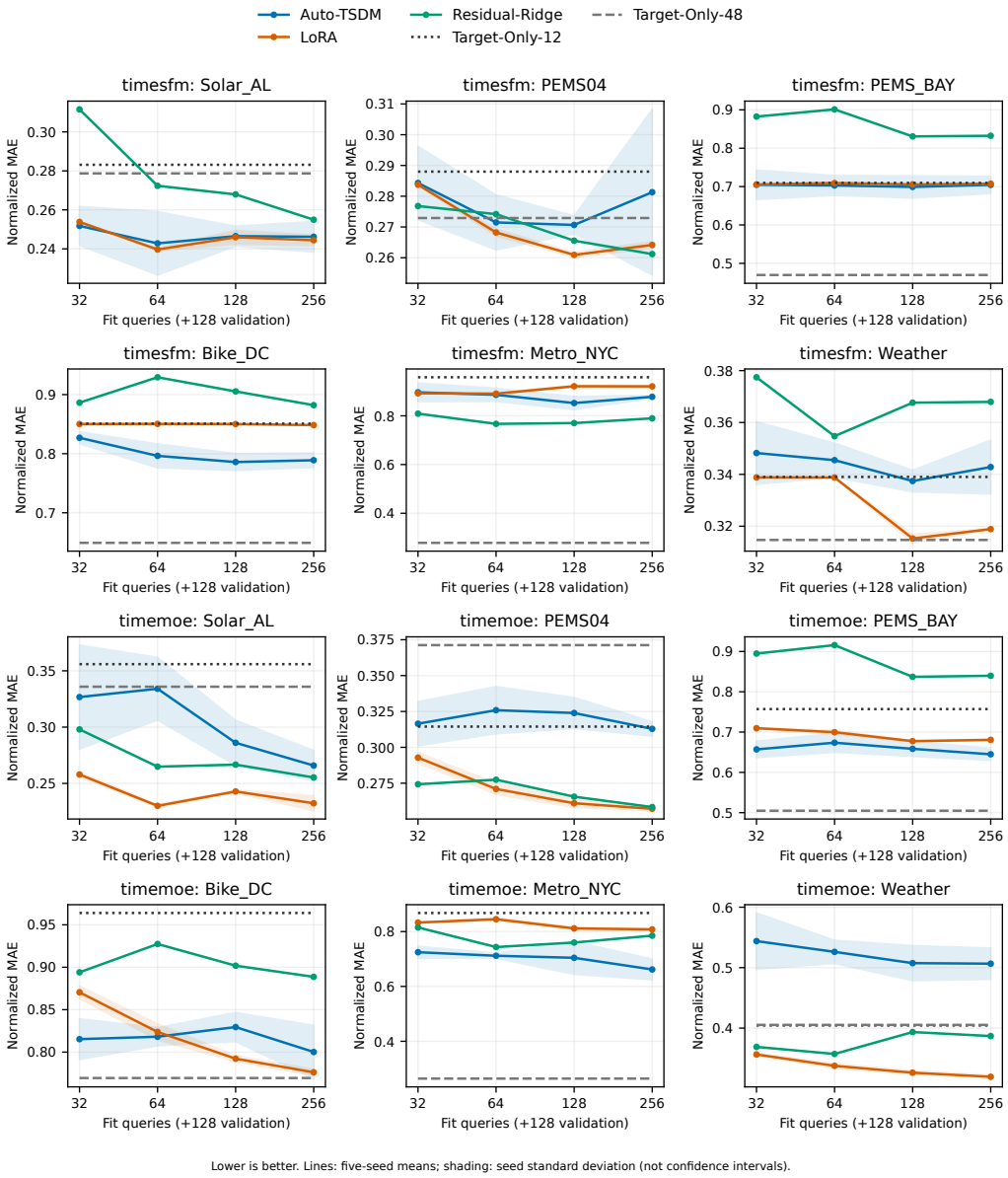


Figure 3: Supervision-budget curves for all six targets and both backbones. Lines show five-seed means and shaded regions show seed standard deviations, not confidence intervals. Every fitting budget additionally uses 128 selection queries.

exploratory block-length intervals support these directions. On TimesFM–Solar, Same and Mixed improve over Cross by 7.57% and 5.62%, respectively. However, TimesFM–PEMS04 deteriorates by 4.69% under Mixed, with all four intervals supporting deterioration. TimeMoE–Weather remains a failure under all three archives: Auto MAE is 0.54489/0.50949/0.55685 for Cross/Same/Mixed, versus 0.43414 for Target-Only-12 and 0.35939 for Residual-Ridge.

The similarity control is selected by validation MAE among raw/feature cosine and Euclidean retrieval and DTW, not all retrieval baselines in the main table. Auto has lower mean MAE than this control in 32/36 archive–target–backbone conditions. These comparisons do not imply that mixed archives improve over same-dataset archives: that claim requires the direct Mixed–Same contrast. Moreover, Mixed replaces half of the same-dataset candidates rather than adding candidates to an intact Same pool. Selection frequencies describe behavior under this fixed-capacity design, not causal

Table 11: Budget study at $N = 64$ fitting queries plus 128 selection queries: normalized MAE, mean $\pm$ sample standard deviation. All budgets share the same 1,024 test queries.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
			TimesFM			
Target-Only-12	0.2831 ± 0.0000	0.2880 ± 0.0000	0.7096 ± 0.0000	0.8509 ± 0.0000	0.9583 ± 0.0000	0.3390 ± 0.0000
Target-Only-48	0.2787 ± 0.0000	0.2729 ± 0.0000	0.4696 ± 0.0000	0.6491 ± 0.0000	0.2786 ± 0.0000	0.3147 ± 0.0000
Auto-TSDM	0.2428 ± 0.0162	0.2715 ± 0.0089	0.7029 ± 0.0259	0.7963 ± 0.0202	0.8862 ± 0.0272	0.3455 ± 0.0066
LoRA	0.2397 ± 0.0006	0.2682 ± 0.0011	0.7093 ± 0.0014	0.8504 ± 0.0002	0.8908 ± 0.0025	0.3388 ± 0.0000
Residual-Ridge	0.2724 ± 0.0000	0.2742 ± 0.0000	0.9011 ± 0.0000	0.9294 ± 0.0000	0.7671 ± 0.0000	0.3547 ± 0.0000
GlobalMean	0.2726 ± 0.0182	0.3041 ± 0.0315	0.7151 ± 0.0464	0.8323 ± 0.0366	0.9675 ± 0.0202	0.3538 ± 0.0142
GP	0.2551 ± 0.0144	0.2714 ± 0.0089	0.7073 ± 0.0296	0.7963 ± 0.0202	0.9304 ± 0.0319	0.3408 ± 0.0086
GR	0.2994 ± 0.0370	0.3279 ± 0.0952	0.7032 ± 0.0300	0.8516 ± 0.0384	0.9397 ± 0.0296	0.3439 ± 0.0050
Reference	0.2480 ± 0.0164	0.3044 ± 0.0128	0.7277 ± 0.0359	0.8697 ± 0.0083	0.8815 ± 0.0245	0.3460 ± 0.0054
Linear	0.2743 ± 0.0111	0.3373 ± 0.0193	0.7266 ± 0.0180	0.9074 ± 0.0173	0.9652 ± 0.0407	0.3657 ± 0.0157
MLP	0.2971 ± 0.0581	0.3237 ± 0.0064	0.7299 ± 0.0225	0.8836 ± 0.0313	0.8746 ± 0.0171	0.3728 ± 0.0206
			TimeMoE			
Target-Only-12	0.3559 ± 0.0000	0.3145 ± 0.0000	0.7571 ± 0.0000	0.9639 ± 0.0000	0.8677 ± 0.0000	0.4044 ± 0.0000
Target-Only-48	0.3359 ± 0.0000	0.3712 ± 0.0000	0.5048 ± 0.0000	0.7695 ± 0.0000	0.2641 ± 0.0000	0.4053 ± 0.0000
Auto-TSDM	0.3341 ± 0.0278	0.3259 ± 0.0164	0.6736 ± 0.0237	0.8182 ± 0.0106	0.7118 ± 0.0095	0.5261 ± 0.0191
LoRA	0.2301 ± 0.0013	0.2710 ± 0.0037	0.6997 ± 0.0044	0.8238 ± 0.0095	0.8450 ± 0.0062	0.3376 ± 0.0039
Residual-Ridge	0.2649 ± 0.0000	0.2774 ± 0.0000	0.9158 ± 0.0000	0.9274 ± 0.0000	0.7436 ± 0.0000	0.3572 ± 0.0000
GlobalMean	0.3470 ± 0.0354	0.3243 ± 0.0145	0.7137 ± 0.0642	0.9260 ± 0.1733	0.8952 ± 0.0395	0.7016 ± 0.0816
GP	0.3693 ± 0.0391	0.3719 ± 0.0471	0.6813 ± 0.0319	0.8184 ± 0.0061	0.7350 ± 0.0186	0.6288 ± 0.0891
GR	0.3721 ± 0.0774	0.3331 ± 0.0226	0.6788 ± 0.0355	0.8109 ± 0.0077	0.7663 ± 0.0455	0.7758 ± 0.0825
Reference	0.3495 ± 0.0244	0.4314 ± 0.0587	0.7301 ± 0.0221	0.8869 ± 0.0331	0.7937 ± 0.0380	0.7173 ± 0.0551
Linear	0.3662 ± 0.0238	0.3457 ± 0.0143	0.6811 ± 0.0054	0.8867 ± 0.0104	0.7754 ± 0.0351	0.5302 ± 0.0132
MLP	0.3157 ± 0.0235	0.3149 ± 0.0095	0.7342 ± 0.0158	0.8309 ± 0.0128	0.7199 ± 0.0191	0.6117 ± 0.0208

Table 12: Budget study at $N = 128$ fitting queries plus 128 selection queries: normalized MAE, mean $\pm$ sample standard deviation. All budgets share the same 1,024 test queries.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
			TimesFM			
Target-Only-12	0.2831 ± 0.0000	0.2880 ± 0.0000	0.7096 ± 0.0000	0.8509 ± 0.0000	0.9583 ± 0.0000	0.3390 ± 0.0000
Target-Only-48	0.2787 ± 0.0000	0.2729 ± 0.0000	0.4696 ± 0.0000	0.6491 ± 0.0000	0.2786 ± 0.0000	0.3147 ± 0.0000
Auto-TSDM	0.2465 ± 0.0051	0.2707 ± 0.0030	0.6991 ± 0.0288	0.7859 ± 0.0142	0.8524 ± 0.0268	0.3374 ± 0.0041
LoRA	0.2460 ± 0.0035	0.2609 ± 0.0008	0.7056 ± 0.0011	0.8502 ± 0.0001	0.9213 ± 0.0053	0.3153 ± 0.0017
Residual-Ridge	0.2679 ± 0.0000	0.2655 ± 0.0000	0.8308 ± 0.0000	0.9054 ± 0.0000	0.7702 ± 0.0000	0.3676 ± 0.0000
GlobalMean	0.2680 ± 0.0235	0.2879 ± 0.0233	0.7126 ± 0.0153	0.8041 ± 0.0274	0.9700 ± 0.0252	0.3696 ± 0.0283
GP	0.2679 ± 0.0195	0.2952 ± 0.0518	0.7082 ± 0.0319	0.7859 ± 0.0142	0.9192 ± 0.0131	0.3682 ± 0.0504
GR	0.2528 ± 0.0106	0.2777 ± 0.0119	0.7172 ± 0.0513	0.8038 ± 0.0236	0.9479 ± 0.0244	0.3641 ± 0.0521
Reference	0.2583 ± 0.0176	0.2972 ± 0.0086	0.7214 ± 0.0160	0.8602 ± 0.0168	0.8641 ± 0.0402	0.3493 ± 0.0052
Linear	0.2674 ± 0.0242	0.3584 ± 0.0163	0.7363 ± 0.0133	0.8894 ± 0.0196	0.9674 ± 0.0439	0.3591 ± 0.0045
MLP	0.3712 ± 0.1777	0.3078 ± 0.0069	0.7210 ± 0.0192	0.8798 ± 0.0386	0.8772 ± 0.0149	0.3765 ± 0.0033
			TimeMoE			
Target-Only-12	0.3559 ± 0.0000	0.3145 ± 0.0000	0.7571 ± 0.0000	0.9639 ± 0.0000	0.8677 ± 0.0000	0.4044 ± 0.0000
Target-Only-48	0.3359 ± 0.0000	0.3712 ± 0.0000	0.5048 ± 0.0000	0.7695 ± 0.0000	0.2641 ± 0.0000	0.4053 ± 0.0000
Auto-TSDM	0.2861 ± 0.0199	0.3239 ± 0.0108	0.6584 ± 0.0186	0.8295 ± 0.0171	0.7043 ± 0.0600	0.5075 ± 0.0289
LoRA	0.2429 ± 0.0012	0.2610 ± 0.0030	0.6775 ± 0.0013	0.7923 ± 0.0034	0.8115 ± 0.0051	0.3262 ± 0.0033
Residual-Ridge	0.2667 ± 0.0000	0.2656 ± 0.0000	0.8371 ± 0.0000	0.9019 ± 0.0000	0.7600 ± 0.0000	0.3934 ± 0.0000
GlobalMean	0.4475 ± 0.0776	0.3431 ± 0.0195	0.6932 ± 0.0265	0.8050 ± 0.0178	0.9093 ± 0.0509	0.6693 ± 0.0783
GP	0.3282 ± 0.0597	0.3723 ± 0.0571	0.6644 ± 0.0171	0.8032 ± 0.0036	0.7418 ± 0.0202	0.7670 ± 0.0959
GR	0.2837 ± 0.0200	0.3656 ± 0.0169	0.6926 ± 0.0737	0.8198 ± 0.0092	0.7166 ± 0.0682	0.6528 ± 0.0924
Reference	0.3579 ± 0.0383	0.4136 ± 0.0258	0.7619 ± 0.0142	0.8741 ± 0.0199	0.7300 ± 0.0275	0.6477 ± 0.0350
Linear	0.3761 ± 0.0221	0.3753 ± 0.0244	0.6774 ± 0.0180	0.8991 ± 0.0258	0.7605 ± 0.0213	0.5075 ± 0.0289
MLP	0.3085 ± 0.0244	0.3239 ± 0.0108	0.7156 ± 0.0069	0.8377 ± 0.0069	0.7370 ± 0.0264	0.6045 ± 0.0574

source contributions. GP/GR may have ties at the top-three boundary; the original stable sorting rule is retained.

C.7 Origin-Phase Robustness

The frozen-model follow-up draws 2,048 queries per target from stride-one candidates covering all 12 origin-modulo-12 phases. Source pools, fitted models, normalization, LoRA weights, and validation choices remain fixed. This is a sample of eligible origins rather than an exhaustive test-

Table 13: Budget study at $N = 256$ fitting queries plus 128 selection queries: normalized MAE, mean $\pm$ sample standard deviation. All budgets share the same 1,024 test queries.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM						
Target-Only-12	0.2831 ± 0.0000	0.2880 ± 0.0000	0.7096 ± 0.0000	0.8509 ± 0.0000	0.9583 ± 0.0000	0.3390 ± 0.0000
Target-Only-48	0.2787 ± 0.0000	0.2729 ± 0.0000	0.4696 ± 0.0000	0.6491 ± 0.0000	0.2786 ± 0.0000	0.3147 ± 0.0000
Auto-TSDM	0.2462 ± 0.0078	0.2813 ± 0.0269	0.7045 ± 0.0222	0.7889 ± 0.0123	0.8781 ± 0.0066	0.3428 ± 0.0104
LoRA	0.2444 ± 0.0027	0.2641 ± 0.0011	0.7076 ± 0.0006	0.8484 ± 0.0002	0.9208 ± 0.0046	0.3189 ± 0.0003
Residual-Ridge	0.2550 ± 0.0000	0.2612 ± 0.0000	0.8324 ± 0.0000	0.8821 ± 0.0000	0.7898 ± 0.0000	0.3679 ± 0.0000
GlobalMean	0.3007 ± 0.0563	0.3266 ± 0.0512	0.7078 ± 0.0354	0.7861 ± 0.0143	0.9529 ± 0.0246	0.4366 ± 0.1185
GP	0.3297 ± 0.1300	0.3272 ± 0.0628	0.6855 ± 0.0133	0.7953 ± 0.0140	0.9208 ± 0.0365	0.3755 ± 0.0622
GR	0.2501 ± 0.0075	0.2733 ± 0.0146	0.6842 ± 0.0153	0.8124 ± 0.0311	0.8880 ± 0.0145	0.3485 ± 0.0105
Reference	0.2664 ± 0.0144	0.3089 ± 0.0111	0.7161 ± 0.0200	0.8524 ± 0.0099	0.8928 ± 0.0182	0.3556 ± 0.0087
Linear	0.2543 ± 0.0162	0.3383 ± 0.0154	0.7289 ± 0.0174	0.8729 ± 0.0209	0.9734 ± 0.0467	0.3523 ± 0.0090
MLP	0.2661 ± 0.0272	0.3140 ± 0.0108	0.7151 ± 0.0217	0.8422 ± 0.0104	0.8846 ± 0.0144	0.3729 ± 0.0074
TimeMoE						
Target-Only-12	0.3559 ± 0.0000	0.3145 ± 0.0000	0.7571 ± 0.0000	0.9639 ± 0.0000	0.8677 ± 0.0000	0.4044 ± 0.0000
Target-Only-48	0.3359 ± 0.0000	0.3712 ± 0.0000	0.5048 ± 0.0000	0.7695 ± 0.0000	0.2641 ± 0.0000	0.4053 ± 0.0000
Auto-TSDM	0.2659 ± 0.0134	0.3129 ± 0.0050	0.6449 ± 0.0152	0.8002 ± 0.0313	0.6619 ± 0.0376	0.5066 ± 0.0259
LoRA	0.2324 ± 0.0066	0.2572 ± 0.0013	0.6807 ± 0.0021	0.7762 ± 0.0035	0.8079 ± 0.0035	0.3195 ± 0.0030
Residual-Ridge	0.2553 ± 0.0000	0.2584 ± 0.0000	0.8397 ± 0.0000	0.8887 ± 0.0000	0.7850 ± 0.0000	0.3869 ± 0.0000
GlobalMean	0.3892 ± 0.0748	0.3775 ± 0.0443	0.6906 ± 0.0457	0.8152 ± 0.0123	0.8283 ± 0.0484	0.7642 ± 0.1933
GP	0.2943 ± 0.0540	0.3932 ± 0.0142	0.6454 ± 0.0151	0.7961 ± 0.0130	0.6851 ± 0.0448	0.6147 ± 0.0981
GR	0.3070 ± 0.0714	0.3382 ± 0.0146	0.6469 ± 0.0152	0.8045 ± 0.0144	0.6839 ± 0.0376	0.6557 ± 0.1084
Reference	0.3780 ± 0.0345	0.4151 ± 0.0149	0.8011 ± 0.0151	0.9100 ± 0.0368	0.6919 ± 0.0174	0.6334 ± 0.0466
Linear	0.3579 ± 0.0337	0.3987 ± 0.0159	0.6865 ± 0.0178	0.8887 ± 0.0241	0.7631 ± 0.0228	0.5196 ± 0.0179
MLP	0.2841 ± 0.0238	0.3149 ± 0.0077	0.7261 ± 0.0096	0.8444 ± 0.0084	0.7113 ± 0.0063	0.5511 ± 0.0100

Table 14: Source availability, cross pool: normalized MAE, mean $\pm$ sample standard deviation. All pools have 900 candidates and share 128 fitting, 128 selection, and 1,024 test queries.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM						
Target-Only-12	0.2920 ± 0.0000	0.2873 ± 0.0000	0.7586 ± 0.0000	0.7444 ± 0.0000	0.9934 ± 0.0000	0.3617 ± 0.0000
Target-Only-48	0.2915 ± 0.0000	0.2702 ± 0.0000	0.5082 ± 0.0000	0.5728 ± 0.0000	0.2975 ± 0.0000	0.3427 ± 0.0000
Random-Prefix	0.3094 ± 0.0023	0.3148 ± 0.0027	0.7828 ± 0.0086	0.7732 ± 0.0079	0.9945 ± 0.0086	0.3859 ± 0.0052
Validation-selected similarity	0.3382 ± 0.0097	0.3023 ± 0.0035	0.8055 ± 0.0054	0.7545 ± 0.0072	1.1576 ± 0.0161	0.4001 ± 0.0054
Residual-Ridge	0.2889 ± 0.0000	0.2757 ± 0.0000	0.7651 ± 0.0000	0.7448 ± 0.0000	0.6924 ± 0.0000	0.3653 ± 0.0000
Auto-TSDM	0.2683 ± 0.0051	0.2636 ± 0.0052	0.7115 ± 0.0083	0.7159 ± 0.0281	0.8631 ± 0.0151	0.3559 ± 0.0088
GlobalMean	0.3284 ± 0.1333	0.2862 ± 0.0216	0.7159 ± 0.0238	0.7246 ± 0.0111	0.9716 ± 0.0171	0.3657 ± 0.0179
GP	0.3000 ± 0.0073	0.2998 ± 0.0615	0.7209 ± 0.0230	0.7022 ± 0.0166	0.9399 ± 0.0267	0.3689 ± 0.0100
GR	0.2673 ± 0.0052	0.2842 ± 0.0313	0.7110 ± 0.0087	0.7870 ± 0.0372	0.9320 ± 0.0327	0.3582 ± 0.0110
Reference	0.2779 ± 0.0056	0.3167 ± 0.0254	0.7178 ± 0.0043	0.7463 ± 0.0082	0.9057 ± 0.0146	0.3903 ± 0.0146
Linear	0.2825 ± 0.0144	0.3350 ± 0.0191	0.7362 ± 0.0134	0.7918 ± 0.0368	0.9787 ± 0.0298	0.3859 ± 0.0037
MLP	0.3210 ± 0.0970	0.3136 ± 0.0086	0.7214 ± 0.0073	0.7429 ± 0.0147	0.8631 ± 0.0151	0.3830 ± 0.0063
TimeMoE						
Target-Only-12	0.3556 ± 0.0000	0.3121 ± 0.0000	0.8144 ± 0.0000	0.8720 ± 0.0000	0.8820 ± 0.0000	0.4341 ± 0.0000
Target-Only-48	0.3409 ± 0.0000	0.3747 ± 0.0000	0.5483 ± 0.0000	0.6573 ± 0.0000	0.2802 ± 0.0000	0.4271 ± 0.0000
Random-Prefix	0.4201 ± 0.0044	0.4457 ± 0.0072	0.7861 ± 0.0040	0.7823 ± 0.0064	0.8329 ± 0.0014	0.7380 ± 0.0096
Validation-selected similarity	0.3707 ± 0.0088	0.3160 ± 0.0030	0.8645 ± 0.0031	0.7824 ± 0.0040	1.0330 ± 0.0162	0.4738 ± 0.0238
Residual-Ridge	0.2767 ± 0.0000	0.2721 ± 0.0000	0.7596 ± 0.0000	0.7470 ± 0.0000	0.6635 ± 0.0000	0.3594 ± 0.0000
Auto-TSDM	0.2843 ± 0.0196	0.3226 ± 0.0106	0.6972 ± 0.0043	0.7222 ± 0.0186	0.6970 ± 0.0160	0.5449 ± 0.0542
GlobalMean	0.5010 ± 0.0572	0.3429 ± 0.0285	0.7444 ± 0.0394	0.7624 ± 0.0479	0.8965 ± 0.0987	0.5780 ± 0.0564
GP	0.3046 ± 0.0475	0.3967 ± 0.0635	0.7004 ± 0.0063	0.7174 ± 0.0199	0.7838 ± 0.0229	0.6689 ± 0.1251
GR	0.3038 ± 0.0370	0.3792 ± 0.0916	0.7063 ± 0.0134	0.7209 ± 0.0092	0.7685 ± 0.0345	0.7757 ± 0.1403
Reference	0.3403 ± 0.0340	0.4390 ± 0.0146	0.7735 ± 0.0266	0.7769 ± 0.0102	0.7200 ± 0.0090	0.6677 ± 0.0454
Linear	0.3328 ± 0.0421	0.4057 ± 0.0291	0.7265 ± 0.0093	0.7715 ± 0.0202	0.7389 ± 0.0136	0.8005 ± 0.0581
MLP	0.2745 ± 0.0104	0.3259 ± 0.0079	0.7219 ± 0.0067	0.7175 ± 0.0072	0.6907 ± 0.0047	0.5506 ± 0.0495

window evaluation. Equal origin phases need not correspond to equal physical daily phases across differently sampled datasets.

Auto has lower mean MAE than Target-Only-12 in 10/12 conditions, Target-Only-48 in 3/12, and LoRA in 5/12. Solar reductions relative to Target-Only-48 remain 8.35% and 20.90% for TimesFM and TimeMoE; their adjusted joint interval lower bounds remain positive at all six block lengths. TimeMoE–Weather deteriorates by 24.24% relative to Target-Only-12. The follow-up therefore preserves the conditional interpretation of the main findings, without establishing new-domain confirmation or correcting the original fitting/selection-origin coverage.

Table 15: Source availability, same pool: normalized MAE, mean ± sample standard deviation. All pools have 900 candidates and share 128 fitting, 128 selection, and 1,024 test queries.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM						
Target-Only-12	0.2920 ± 0.0000	0.2873 ± 0.0000	0.7586 ± 0.0000	0.7444 ± 0.0000	0.9934 ± 0.0000	0.3617 ± 0.0000
Target-Only-48	0.2915 ± 0.0000	0.2702 ± 0.0000	0.5082 ± 0.0000	0.5728 ± 0.0000	0.2975 ± 0.0000	0.3427 ± 0.0000
Random-Prefix	0.3142 ± 0.0106	0.3609 ± 0.0104	0.7240 ± 0.0069	0.7086 ± 0.0034	0.9262 ± 0.0030	0.3940 ± 0.0067
Validation-selected similarity	0.4067 ± 0.0036	0.3081 ± 0.0016	0.8300 ± 0.0050	0.7293 ± 0.0028	1.2587 ± 0.0075	0.3890 ± 0.0079
Residual-Ridge	0.2889 ± 0.0000	0.2757 ± 0.0000	0.7651 ± 0.0000	0.7448 ± 0.0000	0.6924 ± 0.0000	0.3653 ± 0.0000
Auto-TSDM	0.2480 ± 0.0146	0.2656 ± 0.0011	0.7239 ± 0.0158	0.6971 ± 0.0120	0.7919 ± 0.0129	0.3541 ± 0.0075
GlobalMean	0.2953 ± 0.0386	0.2930 ± 0.0208	0.7216 ± 0.0214	0.7116 ± 0.0198	0.9407 ± 0.0500	0.4131 ± 0.1261
GP	0.2856 ± 0.0140	0.3425 ± 0.1672	0.7239 ± 0.0158	0.6925 ± 0.0103	1.0091 ± 0.0119	0.3541 ± 0.0075
GR	0.2604 ± 0.0035	0.2786 ± 0.0245	0.7209 ± 0.0102	0.7797 ± 0.0000	1.0046 ± 0.0189	0.4524 ± 0.0217
Reference	0.2515 ± 0.0165	0.2941 ± 0.0056	0.7169 ± 0.0079	0.7183 ± 0.0159	0.8368 ± 0.0118	0.4132 ± 0.0487
Linear	0.2812 ± 0.0798	0.3024 ± 0.0084	0.7324 ± 0.0076	0.7055 ± 0.0219	0.8576 ± 0.0151	0.4098 ± 0.0288
MLP	0.2618 ± 0.0195	0.3141 ± 0.0064	0.6997 ± 0.0039	0.6982 ± 0.0032	0.7919 ± 0.0129	0.3960 ± 0.0123
TimeMoE						
Target-Only-12	0.3556 ± 0.0000	0.3121 ± 0.0000	0.8144 ± 0.0000	0.8720 ± 0.0000	0.8820 ± 0.0000	0.4341 ± 0.0000
Target-Only-48	0.3409 ± 0.0000	0.3747 ± 0.0000	0.5483 ± 0.0000	0.6573 ± 0.0000	0.2802 ± 0.0000	0.4271 ± 0.0000
Random-Prefix	0.4274 ± 0.0163	0.4717 ± 0.0119	0.7310 ± 0.0044	0.7886 ± 0.0056	0.8135 ± 0.0057	0.7320 ± 0.0156
Validation-selected similarity	0.4193 ± 0.0056	0.3213 ± 0.0012	0.8721 ± 0.0013	0.8374 ± 0.0033	1.1672 ± 0.0094	0.4797 ± 0.0266
Residual-Ridge	0.2767 ± 0.0000	0.2721 ± 0.0000	0.7596 ± 0.0000	0.7470 ± 0.0000	0.6635 ± 0.0000	0.3594 ± 0.0000
Auto-TSDM	0.2721 ± 0.0118	0.3191 ± 0.0075	0.6945 ± 0.0093	0.7155 ± 0.0041	0.6723 ± 0.0099	0.5095 ± 0.0323
GlobalMean	0.3366 ± 0.0579	0.3168 ± 0.0171	0.8066 ± 0.1663	0.7091 ± 0.0044	0.8726 ± 0.0422	0.7260 ± 0.2487
GP	0.2795 ± 0.0029	0.3882 ± 0.0785	0.6945 ± 0.0093	0.7103 ± 0.0000	0.8507 ± 0.0508	0.5556 ± 0.0520
GR	0.2815 ± 0.0028	0.3787 ± 0.0580	0.6975 ± 0.0109	0.7103 ± 0.0000	0.8934 ± 0.0652	0.5918 ± 0.0852
Reference	0.2779 ± 0.0088	0.4321 ± 0.0247	0.7453 ± 0.0276	0.7155 ± 0.0041	0.7087 ± 0.0100	0.7188 ± 0.1205
Linear	0.3052 ± 0.0131	0.3057 ± 0.0032	0.7641 ± 0.0107	0.7012 ± 0.0070	0.7197 ± 0.0075	0.6088 ± 0.1023
MLP	0.2654 ± 0.0104	0.3191 ± 0.0075	0.6996 ± 0.0015	0.6936 ± 0.0042	0.6723 ± 0.0099	0.5651 ± 0.0342

Table 16: Source availability, mixed pool: normalized MAE, mean ± sample standard deviation. All pools have 900 candidates and share 128 fitting, 128 selection, and 1,024 test queries.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM						
Target-Only-12	0.2920 ± 0.0000	0.2873 ± 0.0000	0.7586 ± 0.0000	0.7444 ± 0.0000	0.9934 ± 0.0000	0.3617 ± 0.0000
Target-Only-48	0.2915 ± 0.0000	0.2702 ± 0.0000	0.5082 ± 0.0000	0.5728 ± 0.0000	0.2975 ± 0.0000	0.3427 ± 0.0000
Random-Prefix	0.3157 ± 0.0078	0.3311 ± 0.0093	0.7516 ± 0.0083	0.7361 ± 0.0047	0.9498 ± 0.0100	0.3840 ± 0.0040
Validation-selected similarity	0.3913 ± 0.0069	0.3034 ± 0.0017	0.8286 ± 0.0086	0.7253 ± 0.0033	1.2491 ± 0.0083	0.3890 ± 0.0072
Residual-Ridge	0.2889 ± 0.0000	0.2757 ± 0.0000	0.7651 ± 0.0000	0.7448 ± 0.0000	0.6924 ± 0.0000	0.3653 ± 0.0000
Auto-TSDM	0.2532 ± 0.0081	0.2760 ± 0.0162	0.7140 ± 0.0085	0.7067 ± 0.0213	0.8401 ± 0.0077	0.3507 ± 0.0080
GlobalMean	0.2990 ± 0.0619	0.2680 ± 0.0043	0.7082 ± 0.0192	0.7154 ± 0.0159	0.9672 ± 0.0397	0.3615 ± 0.0099
GP	0.3280 ± 0.0548	0.3657 ± 0.1865	0.7230 ± 0.0105	0.7008 ± 0.0168	0.9572 ± 0.0229	0.3553 ± 0.0085
GR	0.2592 ± 0.0034	0.3567 ± 0.1649	0.7100 ± 0.0061	0.7797 ± 0.0000	0.9522 ± 0.0166	0.4195 ± 0.0667
Reference	0.2724 ± 0.0161	0.2869 ± 0.0115	0.7225 ± 0.0154	0.7284 ± 0.0024	0.8845 ± 0.0191	0.3860 ± 0.0255
Linear	0.2635 ± 0.0232	0.3312 ± 0.0226	0.7324 ± 0.0090	0.7389 ± 0.0266	0.9808 ± 0.0095	0.3908 ± 0.0134
MLP	0.2993 ± 0.0591	0.3088 ± 0.0092	0.7153 ± 0.0039	0.7226 ± 0.0179	0.8401 ± 0.0077	0.3903 ± 0.0105
TimeMoE						
Target-Only-12	0.3556 ± 0.0000	0.3121 ± 0.0000	0.8144 ± 0.0000	0.8720 ± 0.0000	0.8820 ± 0.0000	0.4341 ± 0.0000
Target-Only-48	0.3409 ± 0.0000	0.3747 ± 0.0000	0.5483 ± 0.0000	0.6573 ± 0.0000	0.2802 ± 0.0000	0.4271 ± 0.0000
Random-Prefix	0.4427 ± 0.0134	0.4560 ± 0.0114	0.7602 ± 0.0070	0.7833 ± 0.0072	0.8217 ± 0.0051	0.7325 ± 0.0187
Validation-selected similarity	0.4125 ± 0.0044	0.3211 ± 0.0051	0.8761 ± 0.0081	0.8293 ± 0.0051	1.1555 ± 0.0139	0.4744 ± 0.0312
Residual-Ridge	0.2767 ± 0.0000	0.2721 ± 0.0000	0.7596 ± 0.0000	0.7470 ± 0.0000	0.6635 ± 0.0000	0.3594 ± 0.0000
Auto-TSDM	0.2773 ± 0.0051	0.3125 ± 0.0083	0.6923 ± 0.0044	0.7087 ± 0.0018	0.6698 ± 0.0066	0.5568 ± 0.0155
GlobalMean	0.4181 ± 0.1415	0.3438 ± 0.0293	0.7089 ± 0.0096	0.7340 ± 0.0540	0.8148 ± 0.0863	0.5924 ± 0.1423
GP	0.2801 ± 0.0033	0.3611 ± 0.0420	0.6952 ± 0.0040	0.7103 ± 0.0000	0.8090 ± 0.0219	0.5592 ± 0.0200
GR	0.2793 ± 0.0023	0.3472 ± 0.0303	0.6944 ± 0.0050	0.7103 ± 0.0000	0.7842 ± 0.0069	0.7313 ± 0.1688
Reference	0.2991 ± 0.0412	0.4200 ± 0.0224	0.7558 ± 0.0129	0.7572 ± 0.0073	0.7142 ± 0.0149	0.6839 ± 0.0613
Linear	0.3348 ± 0.0169	0.3979 ± 0.0231	0.7339 ± 0.0114	0.7643 ± 0.0192	0.7424 ± 0.0180	0.6863 ± 0.1337
MLP	0.2704 ± 0.0101	0.3125 ± 0.0083	0.7187 ± 0.0101	0.7132 ± 0.0111	0.6698 ± 0.0066	0.5643 ± 0.0430

C.8 Computation, Memory, and Deployment Volume

Offline measurements rerun seed 41 for all six targets and both backbones. Auto includes one shared utility-labeling pass, GP/GR/Reference statistics, Linear and MLP fitting, and selection among all five estimators. Separate mechanism ablations are not charged to deployment of Auto. LoRA includes the full two-learning-rate search and scheduled validation; Residual-Ridge includes all four regularization coefficients. The reported preparation time starts from standardized inputs. Raw data preparation, source construction, model loading, and warmup are separate stages; adding these stage measurements is not a separately measured cold-start wall-clock experiment.

Table 17: Frozen-model origin-phase follow-up: normalized MAE, mean ± sample standard deviation on 2,048 queries per target. All original estimators and validation choices remain fixed.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
TimesFM						
Target-Only-12	0.3032 ± 0.0000	0.2882 ± 0.0000	0.7424 ± 0.0000	0.7617 ± 0.0000	0.9661 ± 0.0000	0.3099 ± 0.0000
Target-Only-48	0.2902 ± 0.0000	0.2750 ± 0.0000	0.4863 ± 0.0000	0.5773 ± 0.0000	0.2897 ± 0.0000	0.2998 ± 0.0000
Auto-TSDM	0.2659 ± 0.0097	0.2832 ± 0.0269	0.7455 ± 0.0282	0.7156 ± 0.0104	0.8798 ± 0.0077	0.3099 ± 0.0109
LoRA	0.2643 ± 0.0031	0.2636 ± 0.0014	0.7516 ± 0.0012	0.7602 ± 0.0001	0.9002 ± 0.0052	0.2946 ± 0.0004
Residual-Ridge	0.2635 ± 0.0000	0.2622 ± 0.0000	0.8767 ± 0.0000	0.8228 ± 0.0000	0.7960 ± 0.0000	0.3340 ± 0.0000
GP	0.3448 ± 0.1265	0.3303 ± 0.0581	0.7227 ± 0.0243	0.7204 ± 0.0120	0.9176 ± 0.0218	0.3447 ± 0.0657
GR	0.2706 ± 0.0077	0.2752 ± 0.0173	0.7197 ± 0.0239	0.7381 ± 0.0287	0.8905 ± 0.0114	0.3167 ± 0.0134
MLP	0.2775 ± 0.0220	0.3184 ± 0.0047	0.7592 ± 0.0215	0.7718 ± 0.0128	0.8812 ± 0.0104	0.3402 ± 0.0119
Within-source label shuffle	0.3122 ± 0.0174	0.3252 ± 0.0106	0.7978 ± 0.0169	0.8105 ± 0.0170	0.9765 ± 0.0126	0.3445 ± 0.0077
TimeMoE						
Target-Only-12	0.3665 ± 0.0000	0.3159 ± 0.0000	0.8079 ± 0.0000	0.8924 ± 0.0000	0.8691 ± 0.0000	0.3717 ± 0.0000
Target-Only-48	0.3495 ± 0.0000	0.3763 ± 0.0000	0.5411 ± 0.0000	0.6823 ± 0.0000	0.2709 ± 0.0000	0.3706 ± 0.0000
Auto-TSDM	0.2764 ± 0.0118	0.3120 ± 0.0066	0.6733 ± 0.0196	0.7223 ± 0.0303	0.6752 ± 0.0405	0.4618 ± 0.0282
LoRA	0.2482 ± 0.0068	0.2558 ± 0.0013	0.7357 ± 0.0056	0.7121 ± 0.0028	0.8050 ± 0.0023	0.3026 ± 0.0071
Residual-Ridge	0.2626 ± 0.0000	0.2582 ± 0.0000	0.8845 ± 0.0000	0.8304 ± 0.0000	0.7882 ± 0.0000	0.3539 ± 0.0000
GP	0.3073 ± 0.0575	0.3940 ± 0.0128	0.6742 ± 0.0194	0.7205 ± 0.0176	0.6950 ± 0.0467	0.5666 ± 0.0927
GR	0.3171 ± 0.0732	0.3427 ± 0.0154	0.6750 ± 0.0186	0.7325 ± 0.0164	0.6966 ± 0.0374	0.6142 ± 0.1094
MLP	0.2912 ± 0.0188	0.3138 ± 0.0097	0.7692 ± 0.0069	0.7646 ± 0.0067	0.7167 ± 0.0093	0.5071 ± 0.0175
Within-source label shuffle	0.3614 ± 0.0092	0.4455 ± 0.0101	0.7673 ± 0.0172	0.8021 ± 0.0165	0.8564 ± 0.0419	0.6647 ± 0.0267

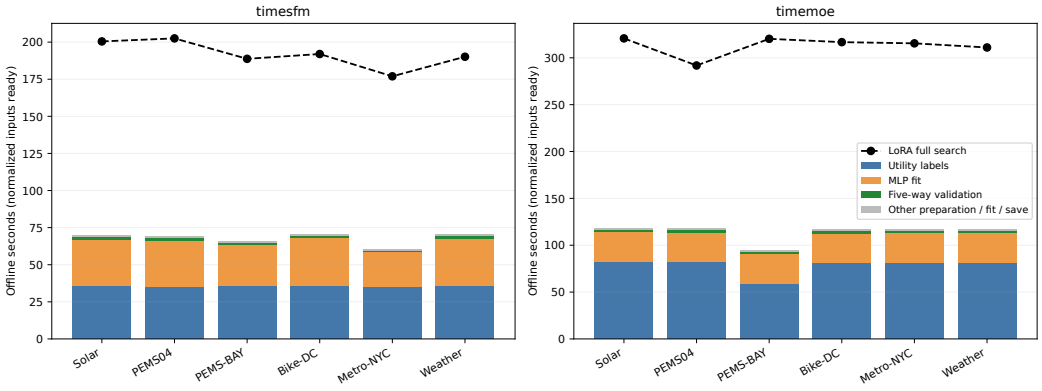


Figure 4: Offline preparation cost decomposition under the fixed implementation and search budgets. Utility labeling and MLP fitting dominate Auto's preparation; a final GP/GR selection does not remove the cost of evaluating the other candidates.

Online measurements keep the model and archive resident in memory and include query processing, retrieval/scoring, and prediction. Auto executes only the estimator selected on validation data for that target–seed configuration; it neither evaluates all candidates nor repeats model selection online. The reported Auto mean therefore aggregates the costs of different selected estimators. Across the 30 configurations per backbone, TimesFM selects GP/GR/Reference/Linear/MLP in 5/16/2/3/4 cases, and TimeMoE in 12/8/0/4/6 cases. When Global is selected, the online execution follows that Global branch without evaluating Reference, Linear, or MLP. This differs from offline preparation, which includes all five candidates. For Auto and LoRA, each target–seed uses the median of three observations before means and medians are computed across the 30 target–seed configurations. Three short observations do not establish a speed significance claim. The batch-128 audit exactly reproduces all 252 original validation prediction files included in the check. Single-query executions exhibit numerical or retrieval differences for some methods and are therefore implementation diagnostics, not a latency result paired with the unchanged batch-128 accuracy.

Median offline PyTorch allocated-memory peaks across targets are 995.6/1,026.0 MiB for Auto and 1,083.3/1,926.7 MiB for LoRA on TimesFM/TimeMoE. Auto labeling uses batch 128 whereas LoRA training uses batch 32, so these are measured configurations rather than an equal-batch architectural memory comparison. They are not total device memory measured by the operating system. The TimeMoE–Solar LoRA timing replay uses a unified expert execution implementation under the

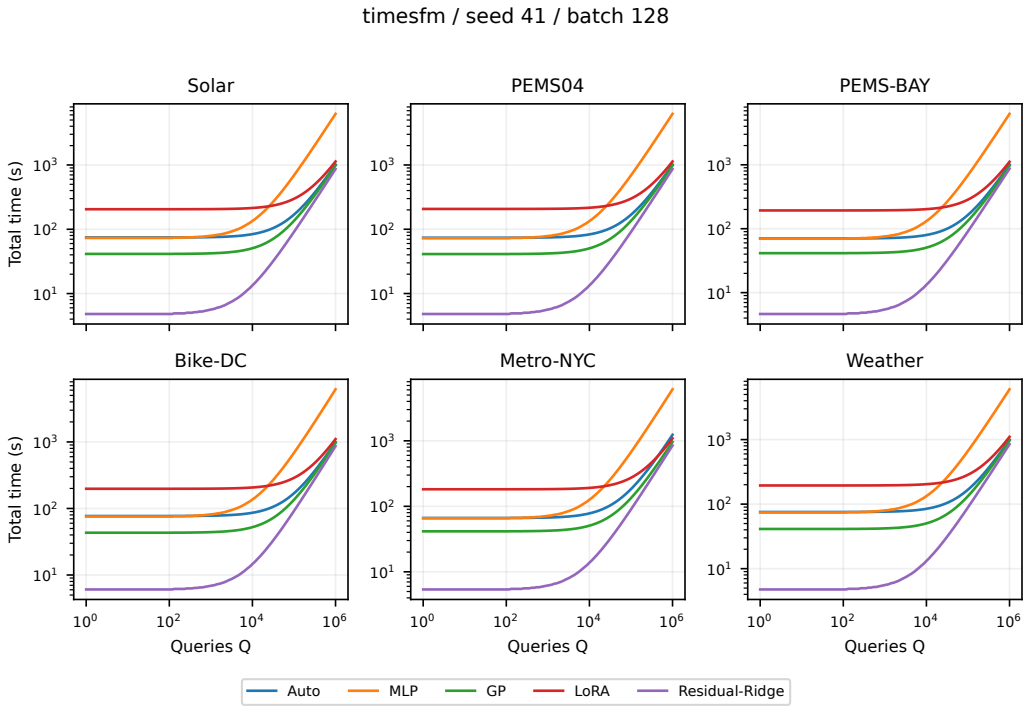


Figure 5: TimesFM preparation plus deployment cost under the measured configuration. Cost intersections do not imply matched forecasting accuracy.

same training schedule; its validation MAE differs from the archived training run by approximately 0.0025. Online timings use the original accuracy weights.

C.9 Single- versus Multi-Dataset Archives under K3

The K3 composition study uses Solar and both backbones with seeds 41–45. The three single-dataset archives each have 900 distinct candidate windows, including the 300 windows used for that dataset in the main three-dataset archive. The additional 600 windows are sampled without replacement using a fixed seed from the remaining eligible training-archive windows. The candidate eligibility rule matches the main study, and all source continuations are withheld. This design holds total candidate count fixed but changes within-dataset coverage as well as archive composition.

Single-dataset estimators are rebuilt using the same single-window utility labels and top-three inference protocol as the main method. All new models and validation choices are locked before evaluating the original 512 test queries. The multi-dataset models and predictions are reused unchanged. The 512-query cohort retains the original stride-12 origin grid; this ablation is not an all-phase evaluation of the newly fitted single-dataset models. Each new label matrix is checked by replaying an original inference batch, and the completed prediction audit recomputes query-level MAE/MSE, checks query alignment and top-three source indices, and verifies the original and new locked artifacts. Statistical comparisons use 5,000 joint seed-and-time bootstrap replicates and six block lengths, reporting exploratory 95% intervals without a global multiplicity guarantee.

The METR-LA-only archive is worse than the three-dataset archive on both backbones across all six exploratory interval settings. The remaining four single-versus-multi comparisons have intervals containing zero at least at one block length. In particular, the lower mean errors of TimesFM–Exchange-only and TimeMoE–ETTh2-only do not establish reliable improvements over the multi-dataset archive. An interval containing zero is not evidence of equivalence. All single-dataset alternatives are reported, rather than selecting a winning archive on the test set.

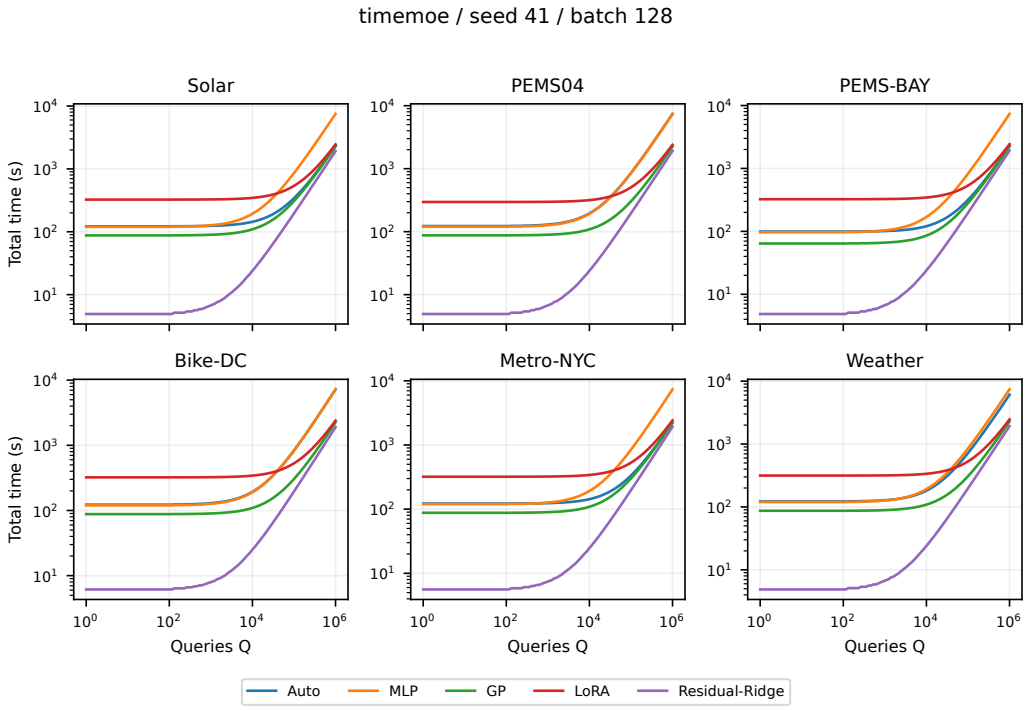


Figure 6: TimeMoE preparation plus deployment cost. For seed-41 PEMS04, Bike-DC, and Weather, estimated intersections with LoRA occur at approximately 33,400, 39,700, and 50,700 queries. Near-zero speed differences are not extrapolated into reliable intersections.

Table 18: Solar source-composition results with TimesFM: normalized MAE, mean ± sample standard deviation over five seeds. All methods share the original 512 test queries. The selected-similarity control uses only the five specified similarity methods.

Method	Three-dataset archive	ETTh2-only	Exchange-only	METR-LA-only
Target-Only-12	0.2780 ± 0.0000	0.2780 ± 0.0000	0.2780 ± 0.0000	0.2780 ± 0.0000
Target-Only-48	0.2795 ± 0.0000	0.2795 ± 0.0000	0.2795 ± 0.0000	0.2795 ± 0.0000
Random-Prefix	0.3005 ± 0.0063	0.3054 ± 0.0033	0.3418 ± 0.0167	0.2929 ± 0.0026
Selected similarity	0.3251 ± 0.0084	0.3660 ± 0.0036	0.3065 ± 0.0043	0.3128 ± 0.0102
Residual-Ridge	0.2540 ± 0.0000	0.2540 ± 0.0000	0.2540 ± 0.0000	0.2540 ± 0.0000
GlobalMean	0.3020 ± 0.0618	0.2928 ± 0.0366	0.2578 ± 0.0160	0.4268 ± 0.2646
GP	0.3314 ± 0.1387	0.2627 ± 0.0389	0.2565 ± 0.0096	0.3377 ± 0.1125
GR	0.2534 ± 0.0096	0.2707 ± 0.0000	0.2513 ± 0.0151	0.2778 ± 0.0126
Reference	0.2627 ± 0.0133	0.2693 ± 0.0139	0.2546 ± 0.0166	0.2940 ± 0.0205
Linear	0.2621 ± 0.0226	0.3119 ± 0.0483	0.2369 ± 0.0059	0.3000 ± 0.0256
MLP	0.2538 ± 0.0272	0.2642 ± 0.0184	0.2480 ± 0.0068	0.2634 ± 0.0109
Auto-TSDM	0.2498 ± 0.0092	0.2539 ± 0.0233	0.2416 ± 0.0080	0.2694 ± 0.0113

The audit covers 40 archive–backbone–seed configurations, 720 prediction files, and 368,640 query–method error records; these records are not independent target observations. Each backbone is evaluated on the same 512 target queries. The 30 newly generated label matrices reproduce their checked original inference batches exactly. Per-query errors are regenerated from saved predictions and serialized at full precision to avoid short-decimal float32 CSV round-trip discrepancies; original CSV exports are retained. The models, predictions, validation choices, and query cohorts are unchanged by this numerical postprocessing.

Auto selections also change with archive composition: on TimeMoE, ETTh2-only selects MLP in all five seeds, whereas the three-dataset archive selects GP/GR in four of five. These comparisons

Table 19: Solar source-composition results with TimeMoE: normalized MAE, mean ± sample standard deviation over five seeds. All methods share the original 512 test queries. The selected-similarity control uses only the five specified similarity methods.

Method	Three-dataset archive	ETTh2-only	Exchange-only	METR-LA-only
Target-Only-12	0.3381 ± 0.0000	0.3381 ± 0.0000	0.3381 ± 0.0000	0.3381 ± 0.0000
Target-Only-48	0.3284 ± 0.0000	0.3284 ± 0.0000	0.3284 ± 0.0000	0.3284 ± 0.0000
Random-Prefix	0.4124 ± 0.0141	0.4340 ± 0.0158	0.4521 ± 0.0203	0.3895 ± 0.0030
Selected similarity	0.3615 ± 0.0094	0.3659 ± 0.0088	0.3705 ± 0.0029	0.3553 ± 0.0076
Residual-Ridge	0.2548 ± 0.0000	0.2548 ± 0.0000	0.2548 ± 0.0000	0.2548 ± 0.0000
GlobalMean	0.3938 ± 0.0754	0.4333 ± 0.0920	0.3014 ± 0.0290	0.3869 ± 0.0667
GP	0.2953 ± 0.0544	0.3444 ± 0.0647	0.2790 ± 0.0161	0.3878 ± 0.0383
GR	0.3078 ± 0.0725	0.3036 ± 0.0242	0.2712 ± 0.0152	0.3641 ± 0.0237
Reference	0.3795 ± 0.0267	0.3839 ± 0.0454	0.3588 ± 0.0214	0.4038 ± 0.0194
Linear	0.3698 ± 0.0398	0.3180 ± 0.0122	0.3101 ± 0.0455	0.3872 ± 0.0550
MLP	0.2825 ± 0.0252	0.2602 ± 0.0088	0.2687 ± 0.0137	0.3263 ± 0.0201
Auto-TSDM	0.2655 ± 0.0135	0.2602 ± 0.0088	0.2670 ± 0.0136	0.3286 ± 0.0230

Table 20: Solar source-composition results with TimesFM: normalized MSE, mean ± sample standard deviation over five seeds. All methods share the original 512 test queries. The selected-similarity control uses only the five specified similarity methods.

Method	Three-dataset archive	ETTh2-only	Exchange-only	METR-LA-only
Target-Only-12	0.2843 ± 0.0000	0.2843 ± 0.0000	0.2843 ± 0.0000	0.2843 ± 0.0000
Target-Only-48	0.2907 ± 0.0000	0.2907 ± 0.0000	0.2907 ± 0.0000	0.2907 ± 0.0000
Random-Prefix	0.2802 ± 0.0088	0.2978 ± 0.0055	0.3485 ± 0.0378	0.2559 ± 0.0036
Selected similarity	0.3364 ± 0.0130	0.4025 ± 0.0060	0.2978 ± 0.0077	0.3085 ± 0.0243
Residual-Ridge	0.1720 ± 0.0000	0.1720 ± 0.0000	0.1720 ± 0.0000	0.1720 ± 0.0000
GlobalMean	0.3111 ± 0.0721	0.3153 ± 0.0815	0.2483 ± 0.0367	0.5317 ± 0.4940
GP	0.3321 ± 0.1372	0.2608 ± 0.0665	0.2523 ± 0.0236	0.3530 ± 0.1906
GR	0.2547 ± 0.0164	0.2877 ± 0.0000	0.2519 ± 0.0297	0.2536 ± 0.0148
Reference	0.2374 ± 0.0244	0.2389 ± 0.0099	0.2227 ± 0.0129	0.2563 ± 0.0301
Linear	0.2157 ± 0.0318	0.2565 ± 0.0206	0.2048 ± 0.0246	0.3110 ± 0.0771
MLP	0.2118 ± 0.0218	0.2152 ± 0.0118	0.2225 ± 0.0254	0.2382 ± 0.0226
Auto-TSDM	0.2385 ± 0.0276	0.2536 ± 0.0469	0.2297 ± 0.0235	0.2475 ± 0.0173

therefore evaluate the full retraining-and-selection procedure for each archive, not the same fixed selector applied to different candidates. Source ordering and the original stable tie-breaking rule are retained; complete selection records and boundary tie counts accompany the results.

C.10 Current-Protocol Direct Forecasting Comparison

All direct predictors use the original 256 fitting queries, 128 validation queries, and 512 test queries per target, with 12-point inputs and a 12-point horizon. These are the current main-study results, not the earlier K1 protocol. The already locked fits and saved predictions are reused and independently checked for query alignment and finite outputs; every reported MAE/MSE is recomputed from those predictions. Direct models are shared across backbone comparisons. The seeds do not resample the fitting-query set; Ridge and the configured XGBoost yield identical outputs across seeds, so zero seed standard deviation does not imply zero data uncertainty.

Ridge selects its regularization from $\{0.1, 1, 10, 100\}$ using validation MAE. XGBoost fits a separate regressor for each forecast step with 200 trees and learning rate 0.05, selecting maximum depth from $\{2, 4\}$. The official DLinear implementation uses a moving-average window of 25, Adam with learning rate 10^{-3}, and MSE training loss; validation MAE selects among epochs 25, 50, and 100. These are specified baseline configurations rather than exhaustive hyperparameter searches. Fixed Utility-MLP is included separately from Auto.

Table 21: Solar source-composition results with TimeMoE: normalized MSE, mean ± sample standard deviation over five seeds. All methods share the original 512 test queries. The selected-similarity control uses only the five specified similarity methods.

Method	Three-dataset archive	ETTh2-only	Exchange-only	METR-LA-only
Target-Only-12	0.3849 ± 0.0000	0.3849 ± 0.0000	0.3849 ± 0.0000	0.3849 ± 0.0000
Target-Only-48	0.3459 ± 0.0000	0.3459 ± 0.0000	0.3459 ± 0.0000	0.3459 ± 0.0000
Random-Prefix	0.3869 ± 0.0334	0.4070 ± 0.0287	0.4843 ± 0.0463	0.3381 ± 0.0075
Selected similarity	0.3829 ± 0.0251	0.3866 ± 0.0176	0.4024 ± 0.0230	0.3569 ± 0.0195
Residual-Ridge	0.1671 ± 0.0000	0.1671 ± 0.0000	0.1671 ± 0.0000	0.1671 ± 0.0000
GlobalMean	0.4674 ± 0.1563	0.5711 ± 0.1807	0.3138 ± 0.0393	0.3946 ± 0.1010
GP	0.3232 ± 0.0774	0.3622 ± 0.0537	0.2952 ± 0.0223	0.3456 ± 0.0359
GR	0.3575 ± 0.1592	0.3338 ± 0.0388	0.2867 ± 0.0272	0.4106 ± 0.0457
Reference	0.3612 ± 0.0489	0.3909 ± 0.0976	0.3641 ± 0.0232	0.4105 ± 0.0322
Linear	0.3527 ± 0.0586	0.2828 ± 0.0174	0.2342 ± 0.0235	0.3656 ± 0.0484
MLP	0.2660 ± 0.0278	0.2352 ± 0.0270	0.2532 ± 0.0332	0.3230 ± 0.0255
Auto-TSDM	0.2644 ± 0.0316	0.2352 ± 0.0270	0.2779 ± 0.0186	0.3218 ± 0.0259

Table 22: Exploratory 95% paired joint seed-and-time intervals for absolute Auto MAE reduction relative to the three-dataset archive on Solar. Positive values favor the single-dataset archive. Columns are temporal block lengths; 5,000 bootstrap replicates are used. These intervals are not globally multiplicity-adjusted.

Single archive	24	72	168	512	1024	2048
TimesFM						
ETTh2-only	$[-0.0302, 0.0235]$	$[-0.0294, 0.0230]$	$[-0.0308, 0.0228]$	$[-0.0292, 0.0223]$	$[-0.0303, 0.0221]$	$[-0.0298, 0.0223]$
Exchange-only	$[-0.0010, 0.0193]$	$[-0.0011, 0.0192]$	$[-0.0015, 0.0199]$	$[-0.0012, 0.0200]$	$[-0.0012, 0.0202]$	$[-0.0011, 0.0201]$
METR-LA-only	$[-0.0365, -0.0019]$	$[-0.0368, -0.0014]$	$[-0.0366, -0.0017]$	$[-0.0359, -0.0020]$	$[-0.0370, -0.0018]$	$[-0.0373, -0.0017]$
TimeMoE						
ETTh2-only	$[-0.0193, 0.0311]$	$[-0.0210, 0.0332]$	$[-0.0201, 0.0315]$	$[-0.0196, 0.0320]$	$[-0.0221, 0.0331]$	$[-0.0233, 0.0368]$
Exchange-only	$[-0.0175, 0.0136]$	$[-0.0183, 0.0134]$	$[-0.0184, 0.0139]$	$[-0.0171, 0.0133]$	$[-0.0189, 0.0133]$	$[-0.0204, 0.0139]$
METR-LA-only	$[-0.0836, -0.0421]$	$[-0.0835, -0.0429]$	$[-0.0840, -0.0424]$	$[-0.0837, -0.0413]$	$[-0.0830, -0.0414]$	$[-0.0838, -0.0413]$

The complete MAE comparison appears in Table 6. Paired comparisons use 5,000 joint seed-and-time bootstrap replicates at block lengths 24, 72, 168, 512, 1024, and 2048. Auto has positive 95% lower bounds at all six block lengths against Ridge in 6/12 conditions, XGBoost in 6/12, and DLinear in 5/12. Conversely, all six interval upper bounds are negative in 2/12, 2/12, and 4/12 conditions, respectively. These are exploratory contrasts without a study-wide multiplicity guarantee; repeated direct-model predictions across backbones are not independent replications.

C.11 Retrieval-Count Sensitivity with Expanded Context

We evaluate $K \in \{1, 3, 5, 7\}$ using a uniform 96-point maximum context and a batch size of 128. Each source window and target history has length 12; the resulting inputs contain 24, 48, 72, and 96 observations. No source prefix is truncated. All predictions are newly computed under the expanded configuration. The original data, query cohorts, candidate archives, and fitted utility estimators are retained. Those estimators were trained using single-window utility labels under the original 64-point configuration. Keeping their parameters fixed isolates the behavior of the existing retrieval models as their inference-time retrieval count changes; it does not assess retraining a joint-K utility model.

For each K, Auto is selected among GP, GR, Reference, Linear, and MLP using the same 128 validation queries. Selected similarity is chosen among raw cosine, raw Euclidean, feature cosine, feature Euclidean, and DTW. All 240 target–backbone–seed–K choices are locked before the new test evaluation. The fixed-K3-Auto control retains the estimator identity selected in the original main study, and Utility-MLP retains the original trained MLP. Rankings use the original stable tie-breaking and concatenation order. Random retrieval samples K distinct candidates using the original per-query seed rule. Source-independent target-history controls share identical newly computed predictions across seeds, rather than using original 64-point outputs.

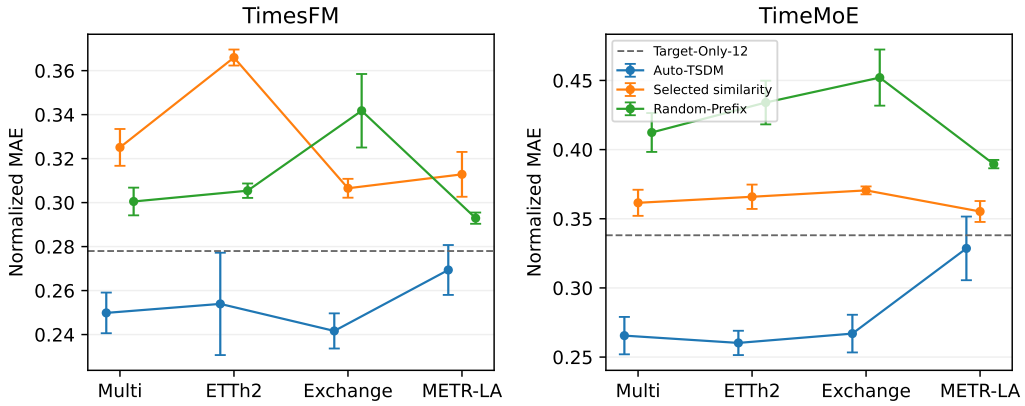


Figure 7: Solar source-archive composition under K3. Error bars are sample standard deviations across five seeds, not confidence intervals. The selected-similarity control is chosen on validation data among the five specified methods. The horizontal line is Target-Only-12.

Table 23: Current-protocol direct-prediction comparison: normalized MSE, five-seed mean and sample standard deviation. Fixed Utility-MLP is reported separately from validation-selected Auto.

Method	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
Direct Ridge	0.1850 ± 0.0000	0.1250 ± 0.0000	3.0271 ± 0.0000	2.5941 ± 0.0000	1.6840 ± 0.0000	0.6080 ± 0.0000
XGBoost	0.2027 ± 0.0000	0.1540 ± 0.0000	2.6445 ± 0.0000	2.7785 ± 0.0000	1.4434 ± 0.0000	1.3028 ± 0.0000
DLinear	0.1946 ± 0.0007	0.1257 ± 0.0002	2.2586 ± 0.0443	2.5365 ± 0.0136	1.6151 ± 0.0612	0.6123 ± 0.0032
timesfm: auto-tsdm	0.2385 ± 0.0276	0.1477 ± 0.0277	2.3409 ± 0.1203	3.0647 ± 0.0933	1.7776 ± 0.1380	0.5108 ± 0.0180
timesfm: utility-mlp	0.2118 ± 0.0218	0.2054 ± 0.0290	2.3878 ± 0.1138	3.0567 ± 0.2310	1.8537 ± 0.1223	0.5538 ± 0.0282
timemoe: auto-tsdm	0.2644 ± 0.0316	0.1820 ± 0.0119	2.1646 ± 0.0510	2.5905 ± 0.1442	0.9883 ± 0.0696	0.7784 ± 0.0872
timemoe: utility-mlp	0.2660 ± 0.0278	0.1885 ± 0.0123	2.3099 ± 0.0451	2.8060 ± 0.0770	1.1434 ± 0.0702	0.9385 ± 0.0848

Comparison with retrieval controls. Figures 8 and 9 compare Auto, validation-selected similarity, Random-Prefix, and Fixed-K3-Auto separately for each target using absolute normalized MAE. Auto and similarity are reselected on validation data at each K; the fixed control changes only the retrieval count while retaining the original $K = 3$ estimator and fitted parameters. This separates within-estimator sensitivity from changes induced by validation-based estimator selection. Each curve shows five-seed means; shading is sample standard deviation, not a confidence interval. Independent vertical scales preserve within-target detail; absolute curve heights should be read from the axes.

Across the 48 target–backbone–K conditions, Auto has lower mean MAE than selected similarity in 42 and Random-Prefix in all 48. The paired exploratory 95% intervals support improvement at all six block lengths in 34 and 43 conditions, respectively. Similarity has lower mean MAE on TimeMoE–PEMS04 at $K = 5, 7$ and TimeMoE–Weather at all four K values; in three of these six conditions, all block lengths support deterioration of Auto relative to similarity. These counts are correlated comparisons over the same queries and do not constitute independent replications or a study-wide significance guarantee.

Against Fixed-K3-Auto, Auto has lower mean MAE in 29 conditions, higher MAE in four, and the same mean within 10^{-8} in 15, including all 12 conditions at $K = 3$. These are descriptive comparisons; the existing paired interval audit does not include this contrast. Auto changes estimator identity in 38/60, 0/60, 28/60, and 29/60 configurations at $K = 1, 3, 5, 7$, respectively. Thus, its curve reflects both retrieval-count changes and validation-based reselection, and should not be interpreted as the response of a single fixed estimator.

Figure 10 additionally includes equal-length continuous target history; the tables also report fixed MLP and Target-Only-12. The latter changes target-history availability: Auto has lower mean MAE in only 12/48 conditions against this reference, so gains over similarity or random external retrieval do not establish superiority when longer target history is available. Complete MAE/MSE and sample standard deviations follow in the tables.

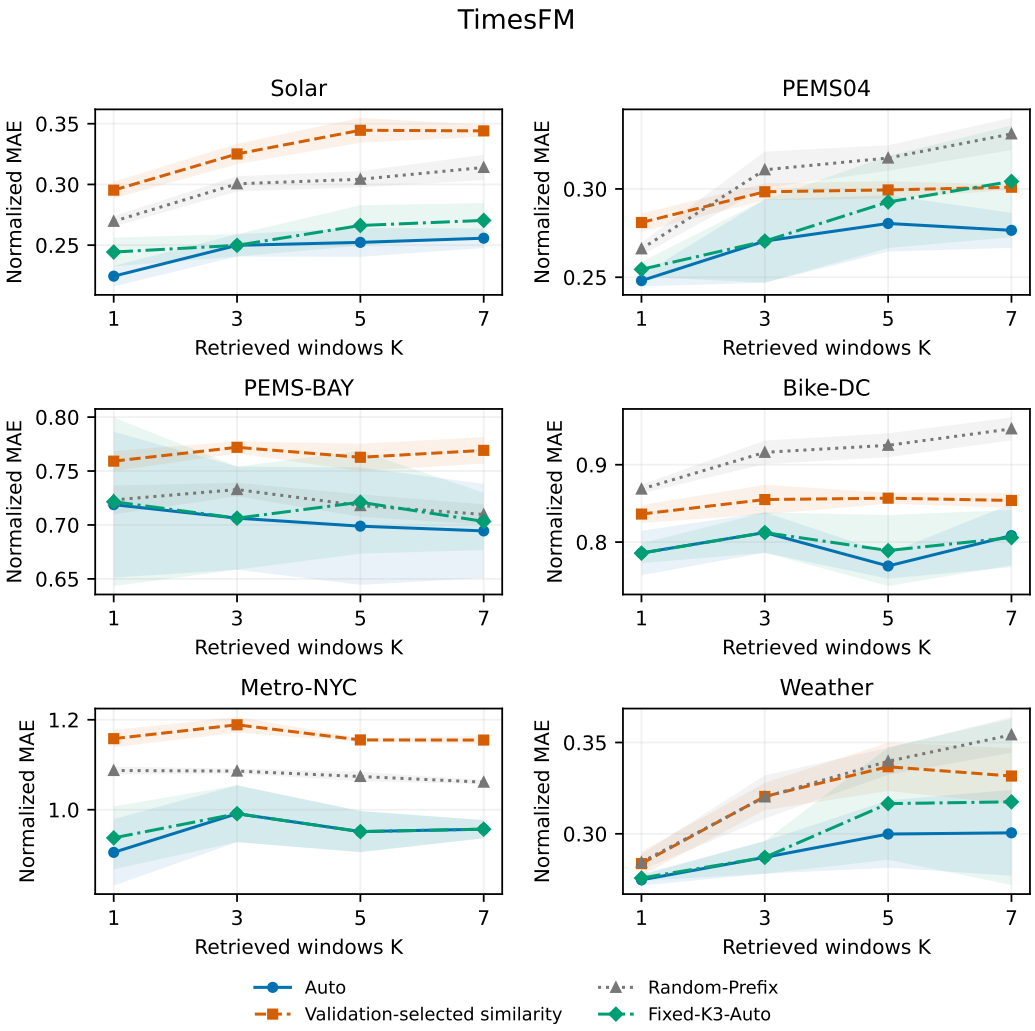


Figure 8: TimesFM: absolute normalized MAE versus retrieved-window count, with Auto, validation-selected similarity, random retrieval, and Fixed-K3-Auto. Each panel is one target; all methods share its test queries and a uniform 96-point maximum context. Shading denotes five-seed sample standard deviation, not confidence intervals. Auto and similarity are selected on validation data separately at each K; Fixed-K3-Auto retains the original $K = 3$ estimator.

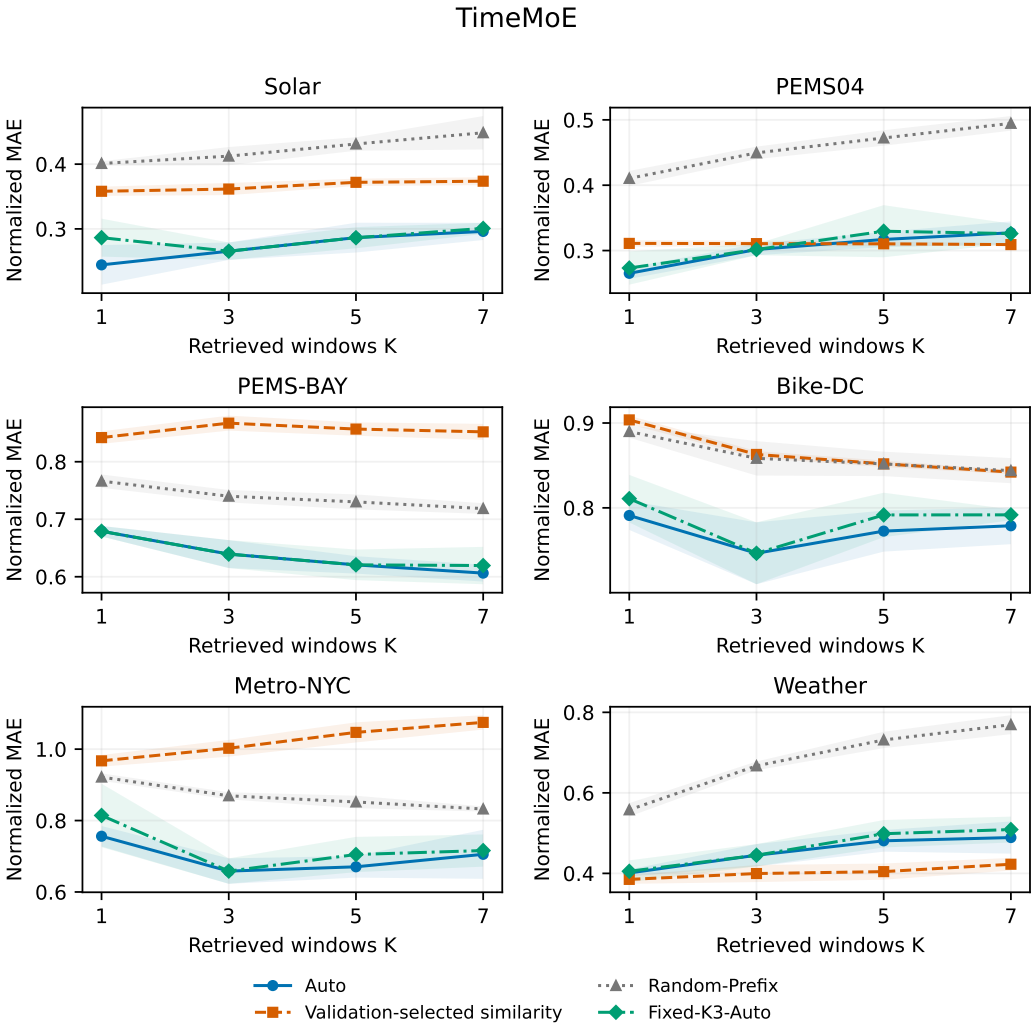


Figure 9: TimeMoE: the same retrieval controls and protocol as Figure 8. Absolute normalized MAE exposes differences in predictive accuracy that normalization to each method's own $K = 3$ value would conceal. Shading denotes five-seed sample standard deviation.

Table 24: Current-protocol K sensitivity for timesfm: normalized MAE, five-seed mean and sample standard deviation.

Method, K	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
auto-tsdm, K=1	0.2245 ± 0.0085	0.2481 ± 0.0033	0.7188 ± 0.0677	0.7859 ± 0.0288	0.9051 ± 0.0739	0.2748 ± 0.0032
auto-tsdm, K=3	0.2499 ± 0.0092	0.2705 ± 0.0236	0.7063 ± 0.0478	0.8123 ± 0.0261	0.9912 ± 0.0632	0.2871 ± 0.0090
auto-tsdm, K=5	0.2523 ± 0.0119	0.2804 ± 0.0159	0.6989 ± 0.0545	0.7691 ± 0.0164	0.9514 ± 0.0457	0.2999 ± 0.0184
auto-tsdm, K=7	0.2557 ± 0.0080	0.2766 ± 0.0100	0.6945 ± 0.0436	0.8084 ± 0.0408	0.9568 ± 0.0203	0.3005 ± 0.0235
fixed-k3-auto, K=1	0.2443 ± 0.0119	0.2545 ± 0.0042	0.7215 ± 0.0782	0.7857 ± 0.0128	0.9374 ± 0.0700	0.2758 ± 0.0015
fixed-k3-auto, K=3	0.2499 ± 0.0092	0.2705 ± 0.0236	0.7063 ± 0.0478	0.8123 ± 0.0261	0.9912 ± 0.0632	0.2871 ± 0.0090
fixed-k3-auto, K=5	0.2662 ± 0.0167	0.2926 ± 0.0261	0.7211 ± 0.0477	0.7890 ± 0.0459	0.9514 ± 0.0457	0.3165 ± 0.0307
fixed-k3-auto, K=7	0.2704 ± 0.0142	0.3043 ± 0.0316	0.7033 ± 0.0265	0.8058 ± 0.0359	0.9568 ± 0.0203	0.3175 ± 0.0455
utility-mlp, K=1	0.2186 ± 0.0090	0.2481 ± 0.0033	0.7198 ± 0.0170	0.8350 ± 0.0096	0.9384 ± 0.0801	0.2782 ± 0.0038
utility-mlp, K=3	0.2538 ± 0.0272	0.3040 ± 0.0121	0.7469 ± 0.0380	0.8443 ± 0.0274	1.0261 ± 0.0403	0.3155 ± 0.0117
utility-mlp, K=5	0.2844 ± 0.0449	0.3206 ± 0.0135	0.7341 ± 0.0372	0.8472 ± 0.0289	0.9870 ± 0.0261	0.3481 ± 0.0181
utility-mlp, K=7	0.2820 ± 0.0395	0.3337 ± 0.0211	0.7426 ± 0.0516	0.8470 ± 0.0143	0.9686 ± 0.0138	0.3621 ± 0.0200
selected-similarity, K=1	0.2953 ± 0.0069	0.2810 ± 0.0052	0.7592 ± 0.0093	0.8362 ± 0.0116	1.1580 ± 0.0193	0.2837 ± 0.0066
selected-similarity, K=3	0.3251 ± 0.0084	0.2984 ± 0.0047	0.7720 ± 0.0056	0.8550 ± 0.0191	1.1888 ± 0.0174	0.3205 ± 0.0075
selected-similarity, K=5	0.3446 ± 0.0103	0.2994 ± 0.0049	0.7627 ± 0.0126	0.8568 ± 0.0072	1.1552 ± 0.0043	0.3367 ± 0.0132
selected-similarity, K=7	0.3441 ± 0.0049	0.3010 ± 0.0015	0.7692 ± 0.0123	0.8538 ± 0.0095	1.1549 ± 0.0090	0.3316 ± 0.0153
random-prefix, K=1	0.2696 ± 0.0043	0.2661 ± 0.0035	0.7232 ± 0.0134	0.8685 ± 0.0056	1.0872 ± 0.0086	0.2845 ± 0.0044
random-prefix, K=3	0.3005 ± 0.0063	0.3108 ± 0.0103	0.7328 ± 0.0063	0.9160 ± 0.0152	1.0859 ± 0.0055	0.3202 ± 0.0118
random-prefix, K=5	0.3043 ± 0.0066	0.3175 ± 0.0072	0.7179 ± 0.0104	0.9251 ± 0.0154	1.0739 ± 0.0104	0.3397 ± 0.0074
random-prefix, K=7	0.3140 ± 0.0104	0.3312 ± 0.0091	0.7098 ± 0.0095	0.9464 ± 0.0148	1.0613 ± 0.0046	0.3542 ± 0.0098
target-equal, K=1	0.2992 ± 0.0000	0.2711 ± 0.0000	0.5826 ± 0.0000	0.7320 ± 0.0000	0.4621 ± 0.0000	0.2850 ± 0.0000
target-equal, K=3	0.2795 ± 0.0000	0.2586 ± 0.0000	0.4498 ± 0.0000	0.5814 ± 0.0000	0.3111 ± 0.0000	0.2821 ± 0.0000
target-equal, K=5	0.2585 ± 0.0000	0.2472 ± 0.0000	0.4405 ± 0.0000	0.5520 ± 0.0000	0.2772 ± 0.0000	0.2750 ± 0.0000
target-equal, K=7	0.2395 ± 0.0000	0.2385 ± 0.0000	0.4095 ± 0.0000	0.5468 ± 0.0000	0.2538 ± 0.0000	0.2677 ± 0.0000
target-12, K=1	0.2780 ± 0.0000	0.2778 ± 0.0000	0.7153 ± 0.0000	0.8991 ± 0.0000	1.1261 ± 0.0000	0.3006 ± 0.0000
target-12, K=3	0.2780 ± 0.0000	0.2778 ± 0.0000	0.7153 ± 0.0000	0.8991 ± 0.0000	1.1261 ± 0.0000	0.3006 ± 0.0000
target-12, K=5	0.2780 ± 0.0000	0.2778 ± 0.0000	0.7153 ± 0.0000	0.8991 ± 0.0000	1.1261 ± 0.0000	0.3006 ± 0.0000
target-12, K=7	0.2780 ± 0.0000	0.2778 ± 0.0000	0.7153 ± 0.0000	0.8991 ± 0.0000	1.1261 ± 0.0000	0.3006 ± 0.0000

Effect of expanding the context configuration. A validation-only check on seed 41 covers all six targets, 128 queries per target, and Target-Only-12 plus MLP retrieval at $K = 1$ and 3. TimeMoE reproduces all checked original predictions exactly. TimesFM changes slightly: the largest absolute forecast difference is 0.00742042, while the largest absolute change in validation MAE is 4.20×10^{-5}. Consequently, the new $K = 3$ point is not assumed identical to the main table. We independently compare all saved $K = 1$ and 3 test predictions with the preceding 64-point study, recording both prediction differences and whether the selected source indices are identical. Auto or similarity estimator changes are distinguished from configuration effects at fixed source inputs. At $K = 3$, Auto selects identical source indices under both configurations in all 60 cases. TimeMoE test predictions are identical; the largest absolute change in TimesFM five-seed mean MAE across the six targets is 2.43×10^{-5}. The main results and direct-prediction comparison retain their original configuration.

Continuous target-history controls use the $12(K + 1)$ observations immediately preceding each origin, with original training-split normalization and no future observations. The control can access history from the preceding chronological partition when needed; the test cohort remains fixed. These are history-available references, not methods under an external-only archive constraint. Boundary counts are recorded per target and K, and the same queries are shared across backbones. Five seeds vary source sampling and estimator initialization, not the target-query sets.

Paired comparisons share query IDs and seed IDs across K, using 5,000 joint seed-and-time bootstrap replicates at block lengths 24, 72, 168, 512, 1024, and 2048. We report exploratory 95% intervals for Auto versus the four controls at each K and for $K \in \{1, 5, 7\}$ versus $K = 3$, without a study-wide multiplicity guarantee. This is an inference-time sensitivity analysis with fixed utility estimates.

Relative to the original main-study estimator identity, Auto changes its choice in 38/60 cases at $K = 1$, 0/60 cases at $K = 3$, 28/60 cases at $K = 5$, 29/60 cases at $K = 7$. The reselected Auto curve therefore includes validation-based estimator changes; fixed-K3-Auto and fixed-MLP controls hold estimator identity and parameters constant.

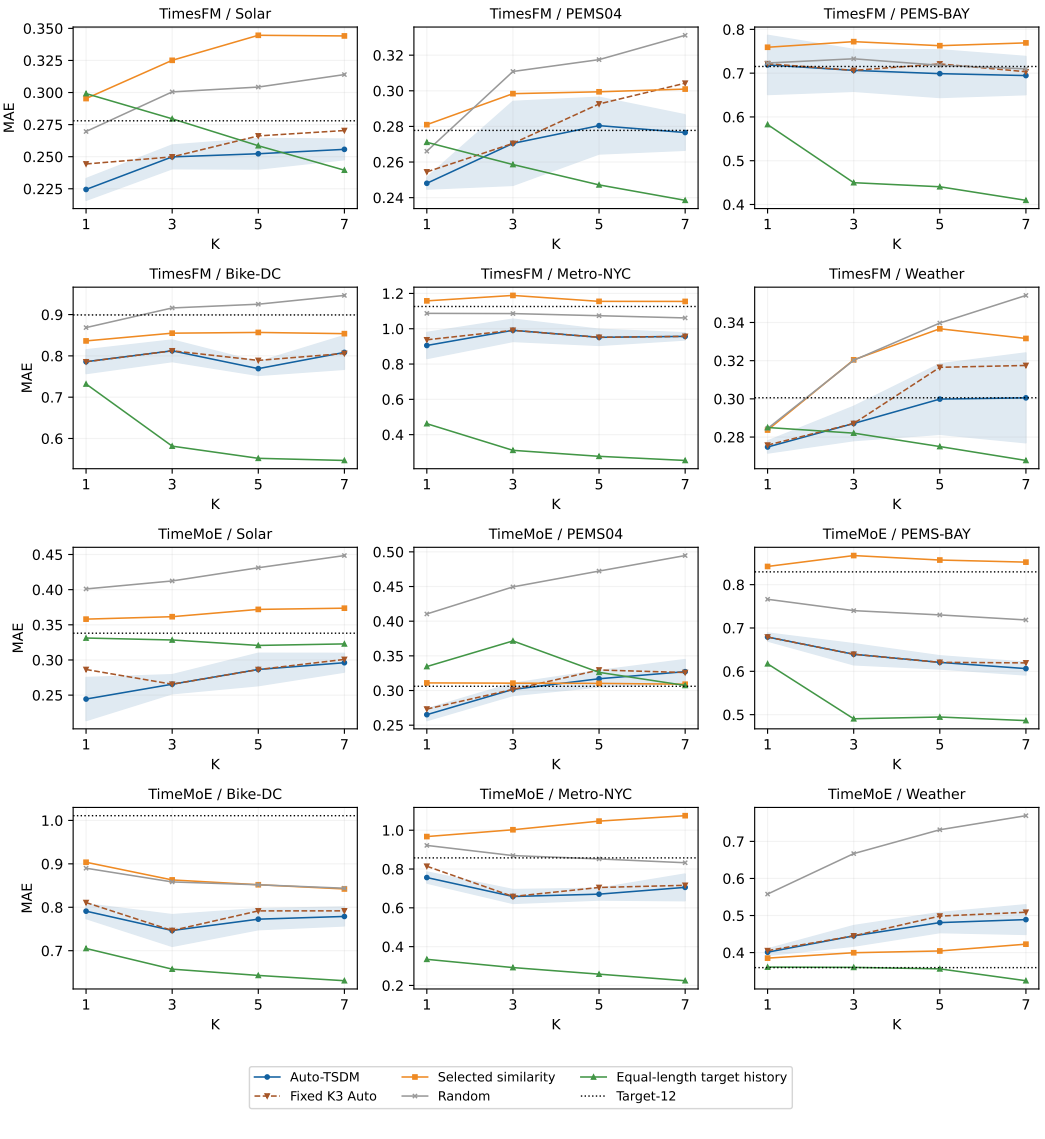


Figure 10: Sensitivity to $K \in \{1, 3, 5, 7\}$ under a uniform 96-point maximum context. Auto is reselected on validation data at each K; Fixed K3 Auto retains the original main-study estimator choice. Shading shows Auto's five-seed sample standard deviation, not a confidence interval. Equal-length target history uses $12(K + 1)$ consecutive target observations.

Table 25: Current-protocol K sensitivity for timemoe: normalized MAE, five-seed mean and sample standard deviation.

Method, K	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
auto-tsdm, K=1	0.2444 ± 0.0306	0.2652 ± 0.0085	0.6791 ± 0.0092	0.7910 ± 0.0171	0.7561 ± 0.0287	0.4011 ± 0.0088
auto-tsdm, K=3	0.2655 ± 0.0135	0.3016 ± 0.0088	0.6394 ± 0.0243	0.7465 ± 0.0365	0.6584 ± 0.0357	0.4451 ± 0.0277
auto-tsdm, K=5	0.2865 ± 0.0230	0.3170 ± 0.0118	0.6206 ± 0.0152	0.7726 ± 0.0242	0.6705 ± 0.0303	0.4809 ± 0.0270
auto-tsdm, K=7	0.2961 ± 0.0133	0.3272 ± 0.0174	0.6063 ± 0.0147	0.7789 ± 0.0217	0.7055 ± 0.0687	0.4890 ± 0.0400
fixed-k3-auto, K=1	0.2865 ± 0.0294	0.2733 ± 0.0255	0.6791 ± 0.0092	0.8108 ± 0.0280	0.8143 ± 0.0889	0.4052 ± 0.0272
fixed-k3-auto, K=3	0.2655 ± 0.0135	0.3016 ± 0.0088	0.6394 ± 0.0243	0.7465 ± 0.0365	0.6584 ± 0.0357	0.4451 ± 0.0277
fixed-k3-auto, K=5	0.2864 ± 0.0164	0.3296 ± 0.0399	0.6207 ± 0.0266	0.7917 ± 0.0262	0.7049 ± 0.0494	0.4985 ± 0.0341
fixed-k3-auto, K=7	0.3008 ± 0.0086	0.3260 ± 0.0153	0.6195 ± 0.0326	0.7918 ± 0.0065	0.7160 ± 0.0452	0.5089 ± 0.0322
utility-mlp, K=1	0.2318 ± 0.0033	0.2652 ± 0.0085	0.7008 ± 0.0254	0.7910 ± 0.0171	0.7831 ± 0.0283	0.4054 ± 0.0153
utility-mlp, K=3	0.2825 ± 0.0252	0.3024 ± 0.0099	0.7169 ± 0.0208	0.7969 ± 0.0128	0.7390 ± 0.0236	0.4928 ± 0.0264
utility-mlp, K=5	0.3086 ± 0.0523	0.3170 ± 0.0118	0.7303 ± 0.0188	0.7994 ± 0.0101	0.7201 ± 0.0204	0.5214 ± 0.0208
utility-mlp, K=7	0.3187 ± 0.0134	0.3272 ± 0.0174	0.7238 ± 0.0150	0.7949 ± 0.0115	0.7092 ± 0.0268	0.5574 ± 0.0212
selected-similarity, K=1	0.3581 ± 0.0071	0.3110 ± 0.0049	0.8420 ± 0.0109	0.9037 ± 0.0050	0.9672 ± 0.0163	0.3851 ± 0.0111
selected-similarity, K=3	0.3615 ± 0.0094	0.3106 ± 0.0047	0.8671 ± 0.0131	0.8630 ± 0.0040	1.0024 ± 0.0234	0.3996 ± 0.0213
selected-similarity, K=5	0.3719 ± 0.0057	0.3103 ± 0.0037	0.8568 ± 0.0116	0.8519 ± 0.0038	1.0469 ± 0.0282	0.4044 ± 0.0201
selected-similarity, K=7	0.3736 ± 0.0061	0.3091 ± 0.0018	0.8520 ± 0.0139	0.8421 ± 0.0024	1.0749 ± 0.0201	0.4227 ± 0.0158
random-prefix, K=1	0.4010 ± 0.0035	0.4104 ± 0.0115	0.7662 ± 0.0116	0.8899 ± 0.0073	0.9217 ± 0.0105	0.5578 ± 0.0166
random-prefix, K=3	0.4124 ± 0.0141	0.4495 ± 0.0096	0.7400 ± 0.0108	0.8583 ± 0.0203	0.8693 ± 0.0099	0.6670 ± 0.0108
random-prefix, K=5	0.4312 ± 0.0104	0.4723 ± 0.0118	0.7302 ± 0.0130	0.8518 ± 0.0144	0.8521 ± 0.0174	0.7314 ± 0.0203
random-prefix, K=7	0.4485 ± 0.0258	0.4947 ± 0.0112	0.7186 ± 0.0093	0.8438 ± 0.0148	0.8324 ± 0.0072	0.7693 ± 0.0235
target-equal, K=1	0.3312 ± 0.0000	0.3346 ± 0.0000	0.6173 ± 0.0000	0.7051 ± 0.0000	0.3346 ± 0.0000	0.3609 ± 0.0000
target-equal, K=3	0.3284 ± 0.0000	0.3715 ± 0.0000	0.4905 ± 0.0000	0.6575 ± 0.0000	0.2921 ± 0.0000	0.3601 ± 0.0000
target-equal, K=5	0.3207 ± 0.0000	0.3262 ± 0.0000	0.4945 ± 0.0000	0.6430 ± 0.0000	0.2582 ± 0.0000	0.3561 ± 0.0000
target-equal, K=7	0.3228 ± 0.0000	0.3073 ± 0.0000	0.4863 ± 0.0000	0.6309 ± 0.0000	0.2243 ± 0.0000	0.3241 ± 0.0000
target-12, K=1	0.3381 ± 0.0000	0.3061 ± 0.0000	0.8295 ± 0.0000	1.0109 ± 0.0000	0.8574 ± 0.0000	0.3595 ± 0.0000
target-12, K=3	0.3381 ± 0.0000	0.3061 ± 0.0000	0.8295 ± 0.0000	1.0109 ± 0.0000	0.8574 ± 0.0000	0.3595 ± 0.0000
target-12, K=5	0.3381 ± 0.0000	0.3061 ± 0.0000	0.8295 ± 0.0000	1.0109 ± 0.0000	0.8574 ± 0.0000	0.3595 ± 0.0000
target-12, K=7	0.3381 ± 0.0000	0.3061 ± 0.0000	0.8295 ± 0.0000	1.0109 ± 0.0000	0.8574 ± 0.0000	0.3595 ± 0.0000

Table 26: Current-protocol K sensitivity for timesfm: normalized MSE, five-seed mean and sample standard deviation.

Method, K	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
auto-tsdm, K=1	0.2010 ± 0.0213	0.1285 ± 0.0064	2.4143 ± 0.1737	2.9769 ± 0.1472	1.5526 ± 0.1795	0.4904 ± 0.0118
auto-tsdm, K=3	0.2385 ± 0.0276	0.1477 ± 0.0277	2.3409 ± 0.1203	3.0650 ± 0.0934	1.7776 ± 0.1380	0.5108 ± 0.0180
auto-tsdm, K=5	0.2380 ± 0.0309	0.1508 ± 0.0101	2.2955 ± 0.0847	2.7667 ± 0.0602	1.6684 ± 0.0992	0.5170 ± 0.0363
auto-tsdm, K=7	0.2314 ± 0.0207	0.1453 ± 0.0095	2.2785 ± 0.0610	2.9429 ± 0.1722	1.6650 ± 0.0561	0.5217 ± 0.0306
fixed-k3-auto, K=1	0.2477 ± 0.0359	0.1327 ± 0.0027	2.4242 ± 0.2067	2.8951 ± 0.0264	1.6209 ± 0.1814	0.4930 ± 0.0060
fixed-k3-auto, K=3	0.2385 ± 0.0276	0.1477 ± 0.0277	2.3409 ± 0.1203	3.0650 ± 0.0934	1.7776 ± 0.1380	0.5108 ± 0.0180
fixed-k3-auto, K=5	0.2679 ± 0.0344	0.1645 ± 0.0340	2.3556 ± 0.1626	2.8609 ± 0.1876	1.6684 ± 0.0992	0.5746 ± 0.0506
fixed-k3-auto, K=7	0.2759 ± 0.0302	0.1745 ± 0.0368	2.2717 ± 0.0672	2.8306 ± 0.1305	1.6650 ± 0.0561	0.5510 ± 0.0421
utility-mlp, K=1	0.1808 ± 0.0047	0.1285 ± 0.0064	2.3509 ± 0.0476	3.1183 ± 0.0901	1.6331 ± 0.2010	0.5067 ± 0.0187
utility-mlp, K=3	0.2118 ± 0.0218	0.2054 ± 0.0290	2.3878 ± 0.1138	3.0567 ± 0.2309	1.8538 ± 0.1223	0.5538 ± 0.0282
utility-mlp, K=5	0.2414 ± 0.0237	0.2210 ± 0.0334	2.3418 ± 0.1333	3.0284 ± 0.1900	1.7463 ± 0.0823	0.5991 ± 0.0308
utility-mlp, K=7	0.2307 ± 0.0240	0.2313 ± 0.0372	2.3695 ± 0.1460	3.0021 ± 0.1115	1.6948 ± 0.0429	0.6015 ± 0.0443
selected-similarity, K=1	0.2974 ± 0.0128	0.1684 ± 0.0068	2.4680 ± 0.0265	2.9215 ± 0.0176	2.1783 ± 0.0473	0.4981 ± 0.0176
selected-similarity, K=3	0.3364 ± 0.0130	0.1860 ± 0.0068	2.4891 ± 0.0302	3.1184 ± 0.2419	2.2242 ± 0.0468	0.5456 ± 0.0109
selected-similarity, K=5	0.3635 ± 0.0166	0.1863 ± 0.0059	2.4487 ± 0.0408	2.9744 ± 0.1522	2.1017 ± 0.0286	0.5733 ± 0.0335
selected-similarity, K=7	0.3533 ± 0.0107	0.1879 ± 0.0036	2.4606 ± 0.0332	2.9568 ± 0.0802	2.0889 ± 0.0363	0.5610 ± 0.0176
random-prefix, K=1	0.2597 ± 0.0068	0.1455 ± 0.0047	2.3698 ± 0.0300	3.0733 ± 0.0324	1.9871 ± 0.0270	0.5123 ± 0.0106
random-prefix, K=3	0.2802 ± 0.0088	0.2017 ± 0.0143	2.3957 ± 0.0193	3.2831 ± 0.1000	2.0022 ± 0.0162	0.5609 ± 0.0337
random-prefix, K=5	0.2694 ± 0.0149	0.2094 ± 0.0124	2.3575 ± 0.0385	3.2460 ± 0.1820	1.9595 ± 0.0309	0.5919 ± 0.0185
random-prefix, K=7	0.2758 ± 0.0181	0.2215 ± 0.0149	2.3460 ± 0.0889	3.3202 ± 0.1241	1.9100 ± 0.0196	0.6068 ± 0.0357
target-equal, K=1	0.3343 ± 0.0000	0.1564 ± 0.0000	2.1108 ± 0.0000	2.6247 ± 0.0000	0.6239 ± 0.0000	0.5322 ± 0.0000
target-equal, K=3	0.2907 ± 0.0000	0.1401 ± 0.0000	1.7092 ± 0.0000	1.9944 ± 0.0000	0.3359 ± 0.0000	0.5626 ± 0.0000
target-equal, K=5	0.2709 ± 0.0000	0.1323 ± 0.0000	1.7015 ± 0.0000	1.8404 ± 0.0000	0.2275 ± 0.0000	0.5527 ± 0.0000
target-equal, K=7	0.2476 ± 0.0000	0.1250 ± 0.0000	1.2325 ± 0.0000	1.8170 ± 0.0000	0.1855 ± 0.0000	0.5479 ± 0.0000
target-12, K=1	0.2843 ± 0.0000	0.1571 ± 0.0000	2.3755 ± 0.0000	3.2056 ± 0.0000	2.1705 ± 0.0000	0.5413 ± 0.0000
target-12, K=3	0.2843 ± 0.0000	0.1571 ± 0.0000	2.3755 ± 0.0000	3.2056 ± 0.0000	2.1705 ± 0.0000	0.5413 ± 0.0000
target-12, K=5	0.2843 ± 0.0000	0.1571 ± 0.0000	2.3755 ± 0.0000	3.2056 ± 0.0000	2.1705 ± 0.0000	0.5413 ± 0.0000
target-12, K=7	0.2843 ± 0.0000	0.1571 ± 0.0000	2.3755 ± 0.0000	3.2056 ± 0.0000	2.1705 ± 0.0000	0.5413 ± 0.0000

C.12 Detailed Utility-Family Comparison

We follow the three estimator categories used in the main text: Global, Reference, and Parametric. Global utility is still utility-based retrieval: GP and GR aggregate observed source utilities into query-independent scores $\hat{u}(s)$. Reference is query-conditioned but nonparametric; it weights fitting-query utilities according to similarity to the current query. Linear and MLP learn parametric source–query scores $\hat{u}_\theta(s, q)$. Auto is a validation-based selection procedure over these estimators,

Table 27: Current-protocol K sensitivity for timemoe: normalized MSE, five-seed mean and sample standard deviation.

Method, K	Solar	PEMS04	PEMS-BAY	Bike-DC	Metro-NYC	Weather
auto-tsdm, K=1	0.2182 ± 0.0619	0.1452 ± 0.0065	2.1842 ± 0.0274	2.7400 ± 0.0431	1.1800 ± 0.0515	0.6762 ± 0.0282
auto-tsdm, K=3	0.2644 ± 0.0316	0.1820 ± 0.0119	2.1646 ± 0.0510	2.5905 ± 0.1442	0.9883 ± 0.0696	0.7784 ± 0.0872
auto-tsdm, K=5	0.2944 ± 0.0401	0.2084 ± 0.0195	2.1174 ± 0.0514	2.6858 ± 0.1344	1.0141 ± 0.0732	0.8734 ± 0.0529
auto-tsdm, K=7	0.3193 ± 0.0104	0.2197 ± 0.0202	2.1220 ± 0.0087	2.7566 ± 0.1580	1.0882 ± 0.1140	0.8954 ± 0.1276
fixed-k3-auto, K=1	0.3043 ± 0.0624	0.1506 ± 0.0180	2.1842 ± 0.0274	2.5813 ± 0.1595	1.2935 ± 0.1561	0.6905 ± 0.0556
fixed-k3-auto, K=3	0.2644 ± 0.0316	0.1820 ± 0.0119	2.1646 ± 0.0510	2.5905 ± 0.1442	0.9883 ± 0.0696	0.7784 ± 0.0872
fixed-k3-auto, K=5	0.2861 ± 0.0282	0.2122 ± 0.0269	2.1216 ± 0.0436	2.7274 ± 0.1099	1.0627 ± 0.0864	0.8975 ± 0.0484
fixed-k3-auto, K=7	0.3072 ± 0.0238	0.2118 ± 0.0112	2.1195 ± 0.0077	2.7572 ± 0.1400	1.0976 ± 0.0637	0.9300 ± 0.0883
utility-mlp, K=1	0.1922 ± 0.0042	0.1452 ± 0.0065	2.2992 ± 0.0775	2.7400 ± 0.0431	1.2298 ± 0.0671	0.6907 ± 0.0458
utility-mlp, K=3	0.2660 ± 0.0278	0.1885 ± 0.0123	2.3099 ± 0.0451	2.8060 ± 0.0770	1.1434 ± 0.0702	0.9385 ± 0.0848
utility-mlp, K=5	0.2932 ± 0.0559	0.2084 ± 0.0195	2.3449 ± 0.0711	2.8288 ± 0.0509	1.0991 ± 0.0686	0.9722 ± 0.0638
utility-mlp, K=7	0.3134 ± 0.0206	0.2197 ± 0.0202	2.3131 ± 0.0720	2.8176 ± 0.0492	1.0762 ± 0.0718	1.0910 ± 0.0912
selected-similarity, K=1	0.4037 ± 0.0188	0.2005 ± 0.0069	2.5049 ± 0.0323	2.7331 ± 0.0350	1.6245 ± 0.0427	0.6194 ± 0.0317
selected-similarity, K=3	0.3829 ± 0.0251	0.1982 ± 0.0061	2.5546 ± 0.0393	2.7620 ± 0.0417	1.6759 ± 0.0519	0.6737 ± 0.0489
selected-similarity, K=5	0.3906 ± 0.0055	0.1974 ± 0.0053	2.5664 ± 0.0579	2.7434 ± 0.0283	1.7754 ± 0.0727	0.6925 ± 0.0456
selected-similarity, K=7	0.3896 ± 0.0098	0.1945 ± 0.0020	2.5582 ± 0.0485	2.7331 ± 0.0450	1.8476 ± 0.0493	0.7292 ± 0.0426
random-prefix, K=1	0.3873 ± 0.0043	0.3306 ± 0.0212	2.2904 ± 0.0300	2.6143 ± 0.0614	1.5432 ± 0.0253	0.9660 ± 0.0823
random-prefix, K=3	0.3869 ± 0.0334	0.3669 ± 0.0158	2.2334 ± 0.0452	2.7344 ± 0.1394	1.4064 ± 0.0241	1.3296 ± 0.0526
random-prefix, K=5	0.3904 ± 0.0170	0.3978 ± 0.0173	2.2037 ± 0.0123	2.7000 ± 0.0794	1.3591 ± 0.0438	1.4750 ± 0.0734
random-prefix, K=7	0.4116 ± 0.0474	0.4253 ± 0.0160	2.1700 ± 0.0218	2.6891 ± 0.0648	1.3063 ± 0.0113	1.5876 ± 0.0849
target-equal, K=1	0.3620 ± 0.0000	0.2263 ± 0.0000	2.0200 ± 0.0000	1.7257 ± 0.0000	0.3587 ± 0.0000	0.6507 ± 0.0000
target-equal, K=3	0.3459 ± 0.0000	0.2721 ± 0.0000	1.7124 ± 0.0000	1.5812 ± 0.0000	0.3132 ± 0.0000	0.6450 ± 0.0000
target-equal, K=5	0.3354 ± 0.0000	0.2206 ± 0.0000	1.6635 ± 0.0000	1.5814 ± 0.0000	0.2325 ± 0.0000	0.6417 ± 0.0000
target-equal, K=7	0.3364 ± 0.0000	0.2050 ± 0.0000	1.6497 ± 0.0000	1.5866 ± 0.0000	0.1643 ± 0.0000	0.5562 ± 0.0000
target-12, K=1	0.3849 ± 0.0000	0.1996 ± 0.0000	2.5995 ± 0.0000	3.0262 ± 0.0000	1.3882 ± 0.0000	0.6404 ± 0.0000
target-12, K=3	0.3849 ± 0.0000	0.1996 ± 0.0000	2.5995 ± 0.0000	3.0262 ± 0.0000	1.3882 ± 0.0000	0.6404 ± 0.0000
target-12, K=5	0.3849 ± 0.0000	0.1996 ± 0.0000	2.5995 ± 0.0000	3.0262 ± 0.0000	1.3882 ± 0.0000	0.6404 ± 0.0000
target-12, K=7	0.3849 ± 0.0000	0.1996 ± 0.0000	2.5995 ± 0.0000	3.0262 ± 0.0000	1.3882 ± 0.0000	0.6404 ± 0.0000

Table 28: Validation-selected utility families: five-seed mean normalized MAE. G: GP/GR; Q: Reference/Linear/MLP; P: Linear/MLP. These are within-family validation selections, not test-set oracle minima.

Backbone	Target	G	Q	P	MLP	Auto
TimeMoE	Bike-DC	0.7430	0.7969	0.7969	0.7969	0.7465
TimeMoE	Metro-NYC	0.6584	0.7291	0.7390	0.7390	0.6584
TimeMoE	PEMS04	0.3320	0.3024	0.3024	0.3024	0.3016
TimeMoE	PEMS-BAY	0.6394	0.6765	0.6765	0.7169	0.6394
TimeMoE	Solar-AL	0.2704	0.2825	0.2825	0.2825	0.2655
TimeMoE	Weather	0.5090	0.4663	0.4663	0.4928	0.4451
TimesFM	Bike-DC	0.8123	0.8452	0.8559	0.8443	0.8123
TimesFM	Metro-NYC	1.0471	0.9912	1.0261	1.0261	0.9912
TimesFM	PEMS04	0.2642	0.2945	0.3032	0.3040	0.2705
TimesFM	PEMS-BAY	0.6680	0.7393	0.7343	0.7469	0.7063
TimesFM	Solar-AL	0.2534	0.2435	0.2435	0.2538	0.2498
TimesFM	Weather	0.2871	0.2991	0.2941	0.3155	0.2871

not a sixth utility model or per-query model gate. GlobalMean remains an internal control, not an Auto candidate.

To avoid comparing a rich model only against a weak global variant, we form TSDM-G by choosing GP or GR on the original 128 validation queries, TSDM-Q by choosing Reference, Linear, or MLP, and TSDM-P by choosing Linear or MLP. This reanalysis uses existing locked fits and predictions on the original 512 test queries; family membership is an exploratory reporting choice, and no test error selects a family member. All models use the original 256 fitting queries, 900-source archives, and five seeds. Fixed MLP and Auto are reported separately. The main table reports Global (G), fixed Reference, and Parametric (P), with Auto as a cross-category selection procedure. Q combines Reference and Parametric only as an auxiliary query-conditioned control; it is not a fourth estimator category. Fixed Reference MAE and uncertainty are included in Table 8.

Q and P each improve over G in mean MAE in 4/12 conditions; fixed MLP does so in 3/12. For each of Q, P, and MLP, only TimeMoE–PEMS04 has positive exploratory paired 95% interval lower bounds at every block length (24, 72, 168, 512, 1024, 2048; 5,000 joint seed/time replicates). There,

Table 29: Standalone estimator online cost (ms/query): six-target mean for seed 41, batch 128, RTX 3090; each target uses the median of three repeats. Retrieval/scoring and native forecasting are included. This table is not the five-seed Auto cost audit.

Estimator	TimesFM	TimeMoE
Global-Positive	0.927	2.173
Global-Risk	0.930	2.176
Reference	1.181	2.474
Linear	4.649	6.009
MLP	6.106	7.271

G has MAE 0.33198 versus 0.30238 for MLP, and MLP is selected in all five within-family Q/P choices. Conversely, Q, P, and MLP have negative upper bounds at every block length in 6, 5, and 7 conditions, respectively. There is no study-wide multiplicity guarantee. The uncertainty in the remaining conditions is not evidence of equivalence.

This comparison distinguishes several claims. Beating shuffled utility labels supports a contribution from observed source–query pairing, but does not imply outperforming a strong global selector. Q improving on G would support query conditioning for that setting, while P or MLP improving on G is needed to support the parametric alternative. Neither implies that greater nonlinearity is universally beneficial. The current results do not establish a rule based on semantic domain, query heterogeneity, or a minimum label count. Auto itself minimizes validation MAE without a latency penalty; choosing a simpler global model when its validation performance is adequate is a deployment consideration, not an evaluated new selection algorithm.

Global source scores can be reused across queries, although forecasting still incurs a cost. Table 29 reports lower measured online costs for GP/GR than for MLP under the same seed-41 protocol. These are batch-amortized times including retrieval and backbone forecasting, not isolated scorer times or single-query response latency. They should not be interpreted as a paired speedup against the five-seed Auto average. A deployment restricted to Global must be evaluated using its own forecasts; Auto accuracy cannot be paired with a standalone Global timing. Parametric scorers evaluate source–query features for the archive and are slower in this implementation. Global estimators still require utility-label construction; they are not training-free. Auto's offline preparation includes labeling, fitting all five candidates, validation selection, and selector persistence even when the chosen estimator is global. It would be incorrect to report a standalone GP/GR cost as the total cost of Auto, or to assign Auto's mixed-estimator mean latency to every variant. Full preparation and LoRA search accounting remain in Appendix C.8.

Within-family offline accounting. The dagger-marked offline entries in Table 2 are sums of measured seed-41 stages, averaged across six targets, rather than separately timed standalone end-to-end family runs. Let B be selector-feature construction, utility-label generation, and the jointly measured GP/GR/Reference statistics stage. Global is charged B plus both GP and GR validation passes; Reference is charged B plus its validation pass; Parametric is charged B, both Linear and MLP fits, and both validation passes. Thus, the Global and Parametric costs include evaluating every candidate in their respective validation-selected families, not merely the winning candidate. The shared statistics stage is conservatively included in all three families; labels are charged once per family, not once per candidate. The existing validation-pass timers cover prediction; the final scalar MAE reduction and argmin are outside those timers. We do not claim a separately measured full wall-clock total.

The resulting Global/Reference/Parametric accounted times are 36.68/36.58/67.24 s for TimesFM and 79.43/79.16/112.39 s for TimeMoE. Raw data preparation, archive construction, model loading and warmup are excluded. Optional standalone-estimator serialization is also excluded; Auto's reported 67.75/113.47 s includes the measured full-selector save, and LoRA's 191.80/312.66 s includes checkpoint operations inside its full two-learning-rate search. Direct predictors include their measured fitting/selection routines. These persistence boundaries differ and are disclosed rather than equating the entries to identical end-to-end deployment pipelines. The main evidence is the lower accounted preparation cost of Global under this implementation, not a universal speedup, a zero-cost global estimator, or a matched-accuracy claim. Online Global remains slightly slower than LoRA, and its mean MAE is lower in only 5/12 conditions.

Timing in the supervision-matched table. Table 2 uses seed 41 for every online-time entry to compare standalone estimators and Auto under the same seed coverage. Each entry averages six target-specific medians of three repeats at batch 128. MAE entries retain the original five-seed means; timing does not have five-seed coverage for all standalone variants. Global and Parametric times match each target's seed-41 validation-selected GP/GR or Linear/MLP identity, respectively, to its measured branch cost. Reference uses its fixed estimator timing. We do not average candidate times indiscriminately or substitute MLP timing for Parametric. Global and Parametric here are TSDM-G and TSDM-P, respectively, as defined in Appendix C.12. Direct Ridge, XGBoost, and DLinear timings use the measured CPU implementation; retrieval with pretrained backbones and LoRA use RTX 3090. Thus, these are implementation costs, not hardware-matched algorithm comparisons. The five-seed Auto/LoRA audit remains separate: their mean online costs are 1.97/3.74 and 0.91/2.12 ms/query, respectively, for TimesFM/TimeMoE. These differ from the seed-41 table because they average all five seeds and their selected branches. Global's lower mean MAE than LoRA in 5/12 conditions is descriptive, not a paired significance claim; its measured online cost is slightly higher than LoRA on both backbones. No Auto accuracy is paired with a Global cost.