DUAL-CONTEXT ANALOG RETRIEVAL FOR TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most long-term time-series forecasting models map the look-back window directly to the full horizon in a single pass. While efficient, this design does not explicitly determine which historical state is most relevant to each future segment or consider what followed that state. Analog forecasting can provide such evidence by retrieving a past state similar to the present and examining its observed continuation. However, relying on a single nearest match can be unreliable, and overlapping patches may yield highly similar or near-duplicate candidates. To address these limitations, we propose DuoTS, a Dual-Context Time Series forecasting model, which leverages retrieved evidence without depending on it exclusively. DuoTS first generates a base forecast using a fully parallel patch encoder and a linear prediction head, and then progressively refines this forecast one future patch at a time. Each refinement considers two complementary views. The first view, the current context, attends to recent tokens and captures the latest dynamics of the series. The second view, the detail context, provides distinct retrieved analogs together with the trajectories that followed them. Since the refinement is performed separately for each future patch, the model weighs the two views differently across the forecast horizon and associates each segment of the prediction with the evidence most appropriate to its temporal distance from the present. Extensive experiments on multiple real-world time-series datasets show that DuoTS achieves state-of-the-art performance, while component-wise ablation studies verify the contribution of each context. Moreover, the refinement mechanism is independent of the remaining DuoTS architecture. It requires only the encoded look-back window and the position of the future patch, allowing it to be applied as a model-agnostic module to improve existing forecasting models.

1 INTRODUCTION

Forecasting models predict the future from the past, and much of what they exploit is recurring structure: trends that persist and cycles that repeat. Models differ in how they use this structure. One class learns it implicitly. PatchTST, iTransformer, and TSMixer map the look-back window directly to the horizon and achieve strong accuracy without an explicit notion of trend or cycle (Nie et al., 2023; Liu et al., 2024b; Chen et al., 2023). A second class separates recurring structure explicitly. Decomposition models such as Autoformer, DLinear, and TimeMixer split or mix trend and seasonal components (Wu et al., 2021; Zeng et al., 2023; Wang et al., 2024), while CycleNet learns a recurrent cycle for each channel and adds it back to the prediction (Lin et al., 2024).

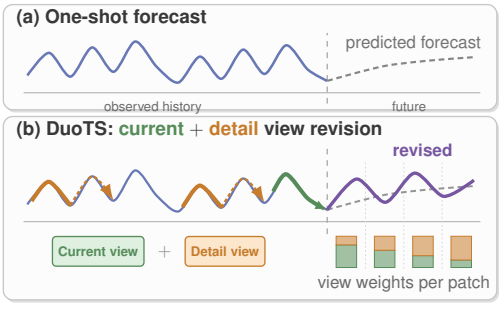


Figure 1: **Forecast, retrieve, revise.** (a) One-shot forecasting predicts the full horizon directly. (b) DuoTS revises each future patch using current and retrieved-detail evidence with learned view weights.

Both classes rely on structure learned across the training set. An implicit model must discover for itself which past states matter for a given window, whereas an explicit cycle describes a typical period and stays fixed after training. Neither is designed to use what the current window itself

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.

reveals. Consider electricity demand during a heat wave. A learned daily cycle describes a typical day, but a heat-wave day is not typical. The most relevant evidence is the earlier days of the same heat wave and how demand evolved after them, and the window already holds it. This is the principle of analog forecasting: find a past state resembling the present and use what followed it (Lorenz, 1969; van den Dool, 1994).

We propose **DuoTS**, which consists of a base forecaster and a corrector. The base forecaster is a fully patch-based model that mixes information sparsely across channels and across time through local dynamics. The corrector makes window-specific evidence explicit and uses it to refine each future patch of the base prediction. It looks at the past through two views: the *current context*, which attends to recent tokens and captures what the series has just been doing, and the *detail context*, which retrieves earlier states resembling the present together with what followed them. Retrieved analogs can nevertheless be misleading: the nearest match is not always the most predictive, overlapping patches yield near-duplicates, and sometimes no good analog exists. The corrector therefore refines the base prediction rather than overriding it, adjusting it only as far as the retrieved evidence warrants. Figure 1 contrasts this forecast–retrieve–revise process with one-shot forecasting.

The corrector reads only patch embeddings, so it can be attached to other patch-based forecasters without changing its design. Attached to SRSNet, TimeKAN, and PatchTST, it improves 15 of 24 tested settings under joint training. The full DuoTS model achieves the lowest MSE in 22 and the lowest MAE in 23 of 32 dataset–horizon settings across eight benchmarks.

Our contributions are as follows.

- We propose DuoTS, which pairs a base forecaster with a corrector that makes window-specific evidence explicit through two views of the past: recent dynamics and retrieved analogs.
- DuoTS achieves state-of-the-art results on eight long-term forecasting benchmarks, obtaining the lowest MSE in 22 and the lowest MAE in 23 of 32 dataset–horizon settings.
- The corrector is model-agnostic: it requires only patch embeddings and transfers across backbones. Attached to SRSNet, TimeKAN, and PatchTST, it improves accuracy in 15 of 24 tested settings. Code is available at `https://anonymous.4open.science/r/DuoTS-244D`.

2 Related Work

Time-series forecasting. Modern forecasting models use several architectural families. DLinear maps decomposed signal components through linear layers (Zeng et al., 2023). TSMixer and TimeMixer use MLP mixing over temporal features or temporal scales (Chen et al., 2023; Wang et al., 2024). PatchTST represents temporal segments as tokens while iTransformer instead represents variables as tokens (Nie et al., 2023; Liu et al., 2024b). Recent models further explore wavelet features in SimpleTM (Chen et al., 2025) and frequency decomposition in TimeKAN (Huang et al., 2025). SRSNet learns which patch representations to retain (Wu et al., 2025). The lightweight TSMixer of Ekambaram et al. (2023) is the closest backbone-level precedent to DuoTS: it applies MLP mixing to patched multivariate series and explicitly models channel interactions with gated components. DuoTS instead combines a direct patch path with a structured path containing sparse cross-channel mixing and explicit first- and second-order patch dynamics, then adds a separate evidence query and dual-context correction for each future patch.

Retrieval-augmented forecasting. Classical analog forecasting uses the continuation of a similar past state (Lorenz, 1969; van den Dool, 1994). Context Parroting demonstrates that related copying strategies remain competitive for scientific time series (Zhang & Gilpin, 2026). Recent methods retrieve evidence from training data or a global memory. RATD guides a diffusion model with retrieved references (Liu et al., 2024a). RAFT supplies matched historical futures (Han et al., 2025). TS-RAG augments a pretrained foundation model with retrieved patterns (Ning et al., 2025). Other methods use a cached nearest-neighbor store (Zhang et al., 2025) or retrieve aligned global segments (Cao et al., 2026). PFRP combines global retrieval with a local model (Du et al., 2026). SARAF promotes both relevance and diversity under nonstationarity (Zhou et al., 2026). DuoTS instead retrieves from the current look-back window for every future patch. Each retrieved analog is paired

with its successors. These successors are the trajectories observed immediately after the matched historical patch. The model places this evidence beside the current context and uses it to refine a direct forecast.

Multi-resolution history modeling. TimeMixer represents a series at several sampling scales (Wang et al., 2024). Pathformer learns pathways across multiple patch sizes (Chen et al., 2024). HDMixer captures dependencies within patches as well as dependencies across patches and variables (Huang et al., 2024). GRACE groups a long user history into journeys and reads several scopes of that history (Ma et al., 2025). DuoTS borrows only the idea of grouping a long sequence into ordered blocks. Those blocks serve as a coarse search index rather than an additional context branch. The search first selects promising blocks and then retrieves individual patches inside them. This design reduces the search space while preserving detailed analog continuations for the final refinement.

3 METHOD

For one sample, let $\mathbf{X} \in \mathbb{R}^{C \times L}$ be a look-back window of L observations from each of C channels, and let $\mathbf{Y} \in \mathbb{R}^{C \times H}$ contain the next H target values. We omit the batch dimension throughout. Reversible instance normalization (RevIN; Kim et al., 2022) normalizes each input channel, and RevIN^{-1} maps a normalized forecast back to its original scale. DuoTS produces a normalized base forecast $\widehat{\mathbf{Y}}^{\mathrm{base}} \in \mathbb{R}^{C \times H}$ and refines it using the dual-context correction $\widehat{\mathbf{E}} \in \mathbb{R}^{C \times H}$. Let P be the patch length and $F = \lceil H/P \rceil$ the number of future patches. Each future patch f has a learned gate $\zeta_f = \tanh(\theta_f)$, repeated over its P output positions to form $\overline{\boldsymbol{\zeta}} \in \mathbb{R}^H$. The final forecast is

$$\widehat{\mathbf{Y}} = \mathrm{RevIN}^{-1}\left(\widehat{\mathbf{Y}}^{\mathrm{base}} + \overline{\boldsymbol{\zeta}} \odot \widehat{\mathbf{E}}\right), \tag{1}$$

where the gate vector is broadcast across channels. The gates share a small initialization but are learned independently; setting them to zero recovers the base forecast.

Figure 2 summarizes the complete forecast–retrieve–revise pipeline.

The normalized window is divided into N overlapping patches of length P and stride $S < P$, where $N = \lceil \max(0, L - P)/S \rceil + 1$. We write $\mathbf{p}_{c,n} \in \mathbb{R}^P$ for patch n of channel c, ordered from oldest to newest. Left padding, used only when needed, ensures that the final patch ends at the forecast boundary. Residual branch weights are initialized at or near zero; λ denotes a tanh-transformed learned scalar, and distinct MLP subscripts denote separate parameters unless stated otherwise.

3.1 TWO-VIEW CORRECTION OF THE ENCODED HISTORY

For each future patch, DuoTS contrasts two views of the encoded history. The current view summarizes the tokens nearest the forecast boundary; the detail view retrieves earlier analogs and the trajectories observed after them. Let $\mathbf{U}_{c,1:N} \in \mathbb{R}^{N \times d}$ denote the patch embeddings supplied by the backbone described in Section 3.2.

Horizon queries and journey blocks. For $f = 1, \ldots, F$, future patch f issues a query anchored at the newest token and offset by a learned horizon embedding $\mathbf{e}^{\mathrm{hor}}_f \in \mathbb{R}^d$:

$$\boldsymbol{\xi}_{c,f} = \mathbf{U}_{c,N} + \mathbf{e}^{\mathrm{hor}}_f \in \mathbb{R}^d. \tag{2}$$

For readability, the equations below suppress c; all operations are applied separately to every channel with shared parameters. Let B_{blk} be the journey-block width. The token sequence is zero-padded only on the left to a multiple of B_{blk} and divided into $M = \lceil N/B_{\mathrm{blk}} \rceil$ ordered blocks $\mathbf{G}_m \in \mathbb{R}^{B_{\mathrm{blk}} \times d}$, $m = 1, \ldots, M$. The compressor and block-axis mixer produce

$$\begin{aligned} \mathbf{b}^0_m &= \mathrm{MLP}_{\mathrm{blk}}(\mathrm{vec}(\mathbf{G}_m)) + \mathbf{e}^{\mathrm{blk}}_m \in \mathbb{R}^d, \\ \mathbf{b}_{1:M} &= \mathbf{b}^0_{1:M} + \lambda_{\mathrm{blk}}\, \mathrm{MLP}_{\mathrm{baxis}}(\mathbf{b}^0_{1:M}) \in \mathbb{R}^{M \times d}, \end{aligned} \tag{3}$$

where $\mathbf{e}^{\mathrm{blk}}_m \in \mathbb{R}^d$ is a learned block-position embedding. The block-axis MLP mixes the M block positions independently for each feature coordinate. The final tokens $\mathbf{b}_m$ form a coarse index for detail retrieval; they are not a third correction view.

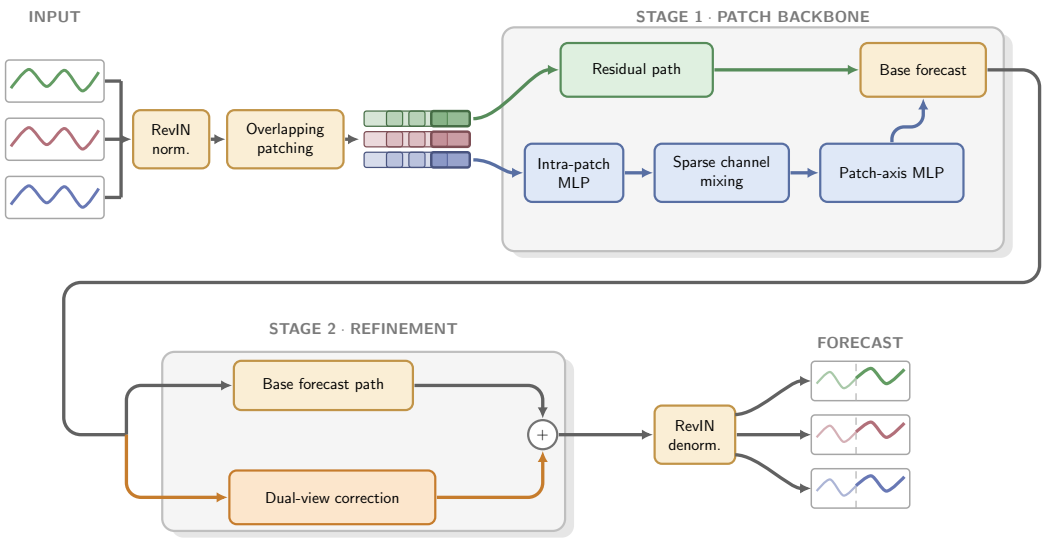


Figure 2: Overview of the DuoTS architecture.

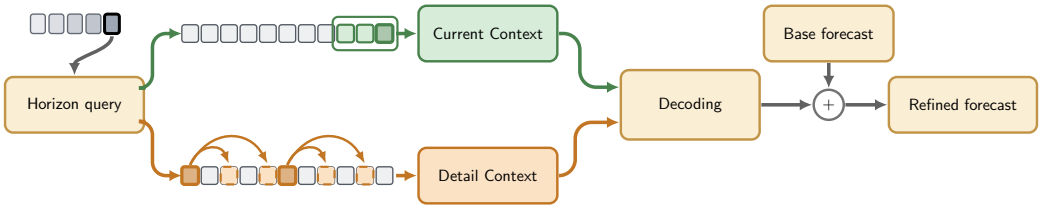


Figure 3: Overview of the dual-context correction. Each future-patch query combines recent tokens with retrieved analogs and their observed successors to produce an additive correction.

Cosine attention. We write $\mathrm{CosSim}(\mathbf{x}, \mathbf{y})$ for the cosine similarity between $\mathbf{x}$ and $\mathbf{y}$. DuoTS uses A-head cosine cross-attention, with head width $d_a = d/A$. For a query $\mathbf{x} \in \mathbb{R}^d$ and memory $\mathbf{Z}_{1:K} = (\mathbf{z}_1, \ldots, \mathbf{z}_K)$, head $a = 1, \ldots, A$ has learned projections $\mathbf{W}_Q^{(a)}, \mathbf{W}_K^{(a)}, \mathbf{W}_V^{(a)} \in \mathbb{R}^{d_a \times d}$ and positive logit scale τ_a^{att}, obtained by exponentiating a learned log-scale and clamping the result above at 100. Its attention weights and the complete operator are defined as follows. Let $\alpha_j^{(a)}$ denote the softmax-normalized weight that head a assigns to memory token j, computed over all K memory tokens from the temperature-scaled cosine similarities $\tau_a^{\mathrm{att}} \mathrm{CosSim}(\mathbf{W}_Q^{(a)}\mathbf{x}, \mathbf{W}_K^{(a)}\mathbf{z}_j)$. Then

$$\mathrm{CosineAttention}(\mathbf{x}, \mathbf{Z}_{1:K}) = \mathbf{W}_O \left[\sum_{j=1}^{K} \alpha_j^{(1)} \mathbf{W}_V^{(1)} \mathbf{z}_j \;\middle\|\; \cdots \;\middle\|\; \sum_{j=1}^{K} \alpha_j^{(A)} \mathbf{W}_V^{(A)} \mathbf{z}_j \right] \in \mathbb{R}^d, \quad (4)$$

where $\mathbf{W}_O \in \mathbb{R}^{d \times d}$ is learned. Attention and output dropout used during training are omitted from equation 4 for readability.

View 1: current context. The state at the forecast boundary is read by multi-head cosine attention from the query to the K_{cur} newest tokens, where K_{cur} is the current-context window size:

$$\mathbf{h}_f^{\mathrm{cur}} = \mathrm{CosineAttention}(\boldsymbol{\xi}_f, \mathbf{U}_{N-K_{\mathrm{cur}}+1:N}) \in \mathbb{R}^d. \quad (5)$$

This context directly summarizes the dynamics closest to the forecast origin.

View 2: detail context. The detail view retrieves past states that resemble $\boldsymbol{\xi}_f$ and reads what followed them. Separate learned query and key projections score journey blocks and eligible historical patches. With $\mathbf{W}_q^x, \mathbf{W}_k^x \in \mathbb{R}^{d\times d}$ and positive logit scales τ_x for $x \in \{\mathrm{blk}, \mathrm{det}\}$, the scores are

$$\begin{aligned} s_{f,m}^{\mathrm{blk}} &= \tau_{\mathrm{blk}} \,\mathrm{CosSim}(\mathbf{W}_q^{\mathrm{blk}}\boldsymbol{\xi}_f, \mathbf{W}_k^{\mathrm{blk}}\mathbf{b}_m), \\ s_{f,n}^{\mathrm{det}} &= \tau_{\mathrm{det}} \,\mathrm{CosSim}(\mathbf{W}_q^{\mathrm{det}}\boldsymbol{\xi}_f, \mathbf{W}_k^{\mathrm{det}}\mathbf{U}_n). \end{aligned} \tag{6}$$

Each τ_x is obtained by exponentiating a learned log-scale and clamping the result above at 100. The block retrieval budget K_{blk} specifies how many highest-scoring valid blocks are retained, and patches are ranked only inside those blocks. A patch is eligible only if all K_{suc} requested successor patches, spaced by $\nu = \lceil P/S \rceil$ patch positions, are observed.

Within the retained blocks, the model greedily selects K_{det} high-scoring patches, where K_{det} is the patch retrieval budget. The exclusion radius ρ is measured in patch positions: the selector prefers a new index more than ρ positions from those already selected. If that preference would leave too few candidates, it falls back to the next highest-scoring valid patch. This avoids filling the evidence set with near-duplicates created by overlapping patches while keeping the requested number of matches whenever possible.

Let $n_f^{(k)}$ be the k-th selected historical index, $k = 1, \ldots, K_{\mathrm{det}}$. The retrieved value combines that state, its ordered successor changes, and the future-patch query:

$$\begin{aligned} \mathbf{v}_f^{(k)} = \mathrm{MLP}_{\mathrm{trans}}\Big[\mathbf{U}_{n_f^{(k)}} \,\|\, (\mathbf{U}_{n_f^{(k)}+\nu} - \mathbf{U}_{n_f^{(k)}}) \,\|\, \cdots \\ \|\, (\mathbf{U}_{n_f^{(k)}+K_{\mathrm{suc}}\nu} - \mathbf{U}_{n_f^{(k)}}) \,\|\, \boldsymbol{\xi}_f\Big] \in \mathbb{R}^d. \end{aligned} \tag{7}$$

Successor changes transfer what happened after the match rather than the absolute level at which the successor ended. Let $\mathcal{V}_f$ contain the valid selected slots, and let $\pi_f^{(k)}$ be the softmax of their retrieval scores after invalid slots are masked. The detail context is

$$\mathbf{h}_f^{\mathrm{det}} = \chi_f^{\mathrm{blk}}\chi_f^{\mathrm{det}} \sum_{k\in\mathcal{V}_f} \pi_f^{(k)}\mathbf{v}_f^{(k)} \in \mathbb{R}^d. \tag{8}$$

The factors χ_f^{blk} and χ_f^{det} are learned sigmoid calibrations of the gap between the best and second-best block and patch scores, respectively. Detail confidence is used only when a single patch is retrieved; otherwise $\chi_f^{\mathrm{det}} = 1$. If no runner-up exists, a large positive margin gives confidence close to one.

Two-view fusion and correction. A sample- and horizon-conditioned two-way softmax blends the current and detail contexts, after which an MLP decodes one correction patch:

$$\begin{aligned} \boldsymbol{\omega}_f &= \mathrm{softmax}\big(\mathrm{MLP}_{\mathrm{view}}([\boldsymbol{\xi}_f \| \mathbf{e}_f^{\mathrm{hor}}])\big) \in \mathbb{R}^2, \\ \overline{\mathbf{h}}_f &= \omega_{f,1}\mathbf{h}_f^{\mathrm{cur}} + \omega_{f,2}\mathbf{h}_f^{\mathrm{det}} \in \mathbb{R}^d, \\ \mathbf{r}_f &= \mathrm{MLP}_{\mathrm{dec}}(\overline{\mathbf{h}}_f) \in \mathbb{R}^P. \end{aligned} \tag{9}$$

Because $\boldsymbol{\xi}_f$ depends on the input and both terms depend on f, the view weights can change across samples and forecast locations. Reintroducing the channel index, the normalized correction is the first H components of the concatenated correction patches:

$$\widehat{\mathbf{E}}_{c,:} = \big[\mathbf{r}_{c,1} \| \cdots \| \mathbf{r}_{c,F}\big]_{1:H} \in \mathbb{R}^H. \tag{10}$$

There is no additional projection after concatenation. Equation 1 then applies the learned future-patch gates and inverse RevIN. Algorithm 1 gives the complete correction procedure.

3.2 Base Forecast: Dual-Path Backbone

The backbone provides both the base forecast and the patch embeddings used by the correction module above. Each patch follows a structured path and a direct path. The structured path mixes information within each patch, across related channels, and across historical patch positions; the direct path preserves a short linear route from the raw patch to the forecast head.

Sparse cross-channel mixing. Each patch first passes through a residual intra-patch MLP. Learned channel embeddings then select the κ_{ch} most related channels and use softmax attention to form their neighborhood consensus. DuoTS passes only the difference between this consensus and the current channel through a shared MLP and adds the result residually with a learned coefficient γ_{ch}, producing $\overline{\mathbf{q}}_{c,n}$. The graph is shared across samples and patch positions; setting $\kappa_{\mathrm{ch}} = 0$ bypasses the update. Appendix B gives the complete construction.

Local and full-window patch dynamics. Let d be the shared token width and Λ the number of structured-path layers. After linearly projecting $\overline{\mathbf{q}}_{c,n}$ to width d and adding a learned patch-position embedding, let $\mathbf{V}_{c,n}^{(0)} \in \mathbb{R}^d$ denote the structured token. At each layer $\ell = 1, \ldots, \Lambda$, DuoTS makes first- and second-order changes explicit:

$$\boldsymbol{\Delta}_{c,n}^{1,(\ell)} = \mathbf{V}_{c,n}^{(\ell-1)} - \mathbf{V}_{c,n-1}^{(\ell-1)}, \qquad \boldsymbol{\Delta}_{c,n}^{2,(\ell)} = \mathbf{V}_{c,n}^{(\ell-1)} - 2\mathbf{V}_{c,n-1}^{(\ell-1)} + \mathbf{V}_{c,n-2}^{(\ell-1)}. \tag{11}$$

Unavailable differences at the first or second patch are set to zero. Learned scalars admit the local trend and curvature signals,

$$\widetilde{\mathbf{V}}_{c,n}^{(\ell)} = \mathbf{V}_{c,n}^{(\ell-1)} + \lambda_\ell^1 \boldsymbol{\Delta}_{c,n}^{1,(\ell)} + \lambda_\ell^2 \boldsymbol{\Delta}_{c,n}^{2,(\ell)}, \tag{12}$$

after which a residual MLP mixes all N patch positions:

$$\mathbf{V}^{(\ell)} = \widetilde{\mathbf{V}}^{(\ell)} + \lambda_\ell^{\mathrm{axis}} \, \mathrm{MLP}_{\mathrm{axis},\ell}(\widetilde{\mathbf{V}}^{(\ell)}). \tag{13}$$

Here $\mathbf{V}^{(\ell)}, \widetilde{\mathbf{V}}^{(\ell)} \in \mathbb{R}^{C \times N \times d}$ stack the tokens for one sample, and d_{ff} is the hidden width of the axis mixer. The mixer applies a shared $N \to d_{\mathrm{ff}} \to N$ MLP independently to every channel and feature coordinate. Thus the differences supply immediate-neighbor dynamics while the axis MLP allows any historical patch to influence any other.

Path fusion and base forecast. The direct token is $\mathbf{D}_{c,n} = \mathbf{W}_{\mathrm{dir}} \mathbf{p}_{c,n} + \mathbf{e}_n^{\mathrm{pos}} \in \mathbb{R}^d$, where $\mathbf{W}_{\mathrm{dir}} \in \mathbb{R}^{d \times P}$ and $\mathbf{e}_n^{\mathrm{pos}} \in \mathbb{R}^d$. The same position embedding was already added to the structured token. The effective scalar λ_{fuse} merges the final structured token with the direct route:

$$\begin{aligned} \mathbf{U}_{c,n} &= \mathbf{D}_{c,n} + \lambda_{\mathrm{fuse}} \left(\mathrm{MLP}_{\mathrm{fuse}}([\mathbf{V}_{c,n}^{(\Lambda)} \| \mathbf{D}_{c,n}]) - \mathbf{D}_{c,n} \right) \in \mathbb{R}^d, \\ \widehat{\mathbf{Y}}_{c,:}^{\mathrm{base}} &= \mathbf{W}_{\mathrm{head}} \, \mathrm{vec}(\mathbf{U}_{c,1:N}) + \mathbf{b}_{\mathrm{head}}. \end{aligned} \tag{14}$$

Here $[\cdot \| \cdot]$ denotes feature concatenation, vec flattens the N tokens in chronological order, $\mathbf{W}_{\mathrm{head}} \in \mathbb{R}^{H \times Nd}$, and $\mathbf{b}_{\mathrm{head}} \in \mathbb{R}^H$. The learned fusion weight λ_{fuse} is initialized at zero, so the backbone initially follows the direct linear patch route. The shared flattening head maps the fused tokens to all H forecast steps in normalized coordinates. The same tokens are also read by the correction module above.

4 Experiments

Evaluation setup and protocol. We evaluate DuoTS on eight standard multivariate benchmarks: ETTh1, ETTh2, ETTm1, and ETTm2 (Zhou et al., 2021), Weather (Max Planck Institute for Biogeochemistry, 2024), Electricity (Trindade, 2015), Traffic (California Department of Transportation (Caltrans), 2024), and Solar (National Renewable Energy Laboratory, 2024). Forecast horizons are $H \in \{96, 192, 336, 720\}$. The look-back length for each setting is selected from $\{96, 336, 512\}$ following Wu et al. (2025). Using one window length for all settings would not be neutral because it affects the cross-channel structure available to a forecasting model (Abdelmalak et al., 2026). Training minimizes mean squared error, and we report MSE and MAE on standardized values. Results are reported as the mean and sample standard deviation over three runs. We compare against six published long-term forecasting baselines: SRSNet (Wu et al., 2025), TimeKAN (Huang et al., 2025), Amplifier (Fei et al., 2025), iTransformer (Liu et al., 2024b), TimeMixer (Wang et al., 2024), and PatchTST (Nie et al., 2023). Appendices C and D provide the benchmark statistics, metric definitions, and search space.

Table 1: Long-term forecasting results for DuoTS at the horizons $H \in \{96, 192, 336, 720\}$ with the horizon average of each dataset in the rows marked *Avg*. The best MSE and MAE per row are in **bold** and the second best are underlined.

		DuoTS (Ours)		SRSNet		TimeKAN		Amplifier		iTransformer		TimeMixer		PatchTST	
Dataset	Horizon	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.363** ±0.0008	**0.388** ±0.0003	0.366	0.394	0.370	0.396	0.373	0.399	0.386	0.405	0.372	0.401	0.377	0.397
	192	**0.400** ±0.0050	**0.413** ±0.0033	**0.400**	0.415	0.403	0.417	0.414	0.420	0.430	0.435	0.413	0.430	0.409	0.425
	336	**0.401** ±0.0002	**0.430** ±0.0004	0.424	**0.430**	0.420	0.432	0.442	0.446	0.450	0.452	0.438	0.450	0.431	0.444
	720	0.441 ±0.0023	0.457 ±0.0016	**0.426**	**0.455**	0.442	0.465	0.455	0.467	0.495	0.487	0.483	0.483	0.457	0.477
Avg		**0.401**	**0.422**	0.404	0.423	0.408	0.427	0.421	0.433	0.440	0.444	0.426	0.441	0.418	0.435
ETTh2	96	**0.263** ±0.0006	**0.332** ±0.0011	0.271	0.338	0.280	0.343	0.287	0.349	0.292	0.347	0.270	0.342	0.274	0.337
	192	**0.312** ±0.0001	**0.367** ±0.0002	0.324	0.372	0.329	0.382	0.348	0.393	0.348	0.384	0.349	0.387	0.348	0.384
	336	**0.314** ±0.0018	**0.377** ±0.0012	0.349	0.394	0.370	0.412	0.383	0.423	0.372	0.407	0.367	0.410	0.377	0.416
	720	**0.383** ±0.0027	**0.424** ±0.0023	0.391	0.434	0.420	0.450	0.407	0.444	0.424	0.444	0.401	0.436	0.406	0.441
Avg		**0.318**	**0.375**	0.333	0.384	0.349	0.396	0.356	0.402	0.359	0.395	0.346	0.393	0.351	0.394
ETTm1	96	**0.284** ±0.0019	0.342 ±0.0010	0.288	**0.341**	0.290	0.348	0.292	0.346	0.287	0.342	0.293	0.345	0.289	0.343
	192	0.334 ±0.0005	0.366 ±0.0009	**0.326**	**0.364**	0.332	0.368	0.327	0.365	0.331	0.371	0.335	0.372	0.329	0.368
	336	0.364 ±0.0002	**0.383** ±0.0005	0.365	0.386	**0.354**	0.386	0.365	0.386	0.358	0.384	0.368	0.386	0.362	0.390
	720	0.419 ±0.0020	0.419 ±0.0013	0.426	0.420	**0.401**	0.417	0.427	0.419	0.412	**0.416**	0.426	0.417	0.416	0.423
Avg		0.350	**0.377**	0.351	**0.377**	**0.344**	0.379	0.352	0.379	0.347	0.378	0.355	0.380	0.349	0.381
ETTm2	96	0.166 ±0.0003	0.255 ±0.0004	**0.164**	**0.254**	**0.164**	**0.254**	**0.164**	**0.254**	0.168	0.262	0.165	0.256	0.165	0.255
	192	**0.220** ±0.0011	**0.293** ±0.0008	**0.220**	0.296	0.238	0.300	0.226	0.300	0.224	0.295	0.225	0.298	0.221	**0.293**
	336	0.278 ±0.0034	0.330 ±0.0014	**0.271**	**0.327**	0.278	0.331	0.276	0.331	0.274	0.330	0.277	0.332	0.276	**0.327**
	720	0.364 ±0.0015	0.385 ±0.0009	**0.353**	**0.380**	0.359	0.387	0.358	0.388	0.367	0.385	0.360	0.387	0.362	0.381
Avg		0.257	0.316	**0.252**	**0.314**	0.259	0.318	0.256	0.318	0.258	0.318	0.256	0.318	0.256	**0.314**
Weather	96	**0.142** ±0.0017	**0.193** ±0.0015	0.147	0.198	0.151	0.205	0.147	0.199	0.150	0.207	0.151	0.198	0.150	0.200
	192	**0.188** ±0.0019	**0.238** ±0.0019	0.190	0.242	0.195	0.244	0.194	0.245	0.200	0.248	0.191	0.242	0.191	0.239
	336	**0.237** ±0.0038	**0.276** ±0.0031	0.241	0.282	0.242	0.287	0.243	0.282	0.252	0.287	0.244	0.280	0.242	0.279
	720	0.312 ±0.0022	0.330 ±0.0015	0.325	0.340	0.317	0.340	**0.310**	**0.329**	0.320	0.336	0.316	0.331	0.312	0.330
Avg		**0.219**	**0.259**	0.225	0.265	0.226	0.269	0.223	0.263	0.230	0.269	0.225	0.262	0.223	0.262
Electricity	96	**0.127** ±0.0003	**0.222** ±0.0004	0.131	0.226	0.135	0.231	0.132	0.227	0.134	0.230	0.153	0.256	0.143	0.247
	192	**0.146** ±0.0007	**0.241** ±0.0007	0.147	**0.241**	0.149	0.243	0.149	0.243	0.154	0.250	0.168	0.269	0.158	0.260
	336	**0.162** ±0.0006	**0.258** ±0.0007	0.163	**0.258**	0.165	0.260	0.167	0.261	0.169	0.265	0.189	0.281	0.168	0.267
	720	0.206 ±0.0044	0.296 ±0.0032	0.201	0.291	0.206	0.297	0.203	0.292	**0.194**	**0.288**	0.228	0.320	0.214	0.307
Avg		**0.160**	**0.254**	**0.160**	**0.254**	0.163	0.257	0.162	0.255	0.162	0.258	0.184	0.281	0.170	0.270
Traffic	96	**0.355** ±0.0007	**0.252** ±0.0008	0.361	0.254	0.388	0.269	0.391	0.277	0.363	0.265	0.369	0.256	0.370	0.262
	192	**0.380** ±0.0006	**0.259** ±0.0004	**0.380**	0.263	0.411	0.286	0.405	0.283	0.384	0.273	0.400	0.271	0.386	0.269
	336	**0.386** ±0.0016	**0.262** ±0.0002	0.392	0.270	0.425	0.284	0.416	0.290	0.396	0.277	0.407	0.272	0.396	0.275
	720	**0.424** ±0.0032	**0.284** ±0.0021	0.434	0.293	0.455	0.302	0.436	0.312	0.445	0.308	0.462	0.316	0.435	0.295
Avg		**0.387**	**0.264**	0.391	0.270	0.419	0.285	0.412	0.290	0.397	0.280	0.409	0.278	0.396	0.275
Solar	96	**0.162** ±0.0016	**0.218** ±0.0008	0.167	0.222	0.187	0.255	0.175	0.237	0.174	0.229	0.180	0.233	0.170	0.234
	192	**0.177** ±0.0014	**0.236** ±0.0016	0.182	0.237	0.194	0.265	0.198	0.259	0.205	0.270	0.201	0.259	0.204	0.302
	336	**0.185** ±0.0019	**0.241** ±0.0019	0.188	0.245	0.203	0.264	0.213	0.259	0.216	0.282	0.214	0.272	0.212	0.293
	720	0.196 ±0.0006	**0.250** ±0.0032	**0.195**	0.251	0.209	0.269	0.222	0.369	0.211	0.260	0.218	0.278	0.215	0.307
Avg		**0.180**	**0.236**	0.183	0.238	0.198	0.263	0.202	0.281	0.201	0.260	0.203	0.260	0.200	0.284
# Best/32		**22**	**23**	9	9	3	1	2	2	1	2	0	0	0	2

4.1 Main Results

Table 1 reports the full comparison. DuoTS obtains the lowest MAE in 23 of the 32 dataset–horizon settings and the lowest MSE in 22, compared with 9 and 9 for SRSNet, the strongest baseline. Based on the horizon averages, DuoTS leads both metrics on ETTh1, ETTh2, Weather, Traffic, and Solar. It ties SRSNet on Electricity, achieves the best MAE on ETTm1, and trails on ETTm2.

ETTh2 shows the clearest improvement. DuoTS reduces the horizon-average MSE by 4.5% and the MAE by 2.3% relative to SRSNet and leads at all four horizons. Traffic also provides a large and diverse retrieval pool through its 862 sensors, and DuoTS leads at every horizon on both metrics.

The correction is more challenging on the two quarter-hourly ETT subsets. At a 15-minute sampling interval the look-back window spans only a day or two, and adjacent patches often provide similar candidates. DuoTS still matches the best MAE on ETTm1 but gives up 1.7% MSE to TimeKAN. On ETTm2 it trails SRSNet by 2.0% in horizon-average MSE.

4.2 Ablation Study

Table 2 compares DuoTS with four training-time variants over ETTh2, ETTm2, Weather, and Electricity. The variants remove the second-order patch dynamics, sparse cross-channel mixing, or the intra-patch mixer, or disable the patch-axis MLP, respectively. Each dataset column averages the four forecasting horizons.

Table 2: Component ablation results averaged over the four forecasting horizons for each dataset. Each variant was tuned separately for this ablation study. The best MSE and MAE in each column are in **bold** and the second best are underlined. The DuoTS row exactly reproduces the corresponding *Avg* rows of Table 1.

Case	ETTh2		ETTm2		Weather		Electricity	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
DuoTS	**0.318**	0.375	0.257	0.316	**0.219**	**0.259**	0.160	0.254
w/o second-order dynamics	0.319	**0.374**	0.259	0.318	0.227	0.265	0.160	0.253
w/o channel mixing	0.319	0.375	**0.256**	**0.314**	0.227	0.265	0.166	0.260
w/o intra-patch mixer	0.322	0.376	0.263	0.319	0.224	0.262	0.164	0.258
w/o patch-axis MLP	0.330	0.383	0.263	0.318	0.224	0.263	**0.158**	**0.252**

The full configuration is best on both metrics for Weather and has the lowest MSE on ETTh2. Channel mixing is most useful on Electricity, while removing it gives a small improvement on ETTm2. Second-order dynamics are beneficial in Weather and ETTm2. Removing them improves MAE by 0.001 on ETTh2 and Electricity and ties the full model in Electricity MSE. The intra-patch mixer helps on every reported dataset. Disabling the patch-axis MLP degrades ETTh2, ETTm2, and Weather, but improves Electricity by 0.002 in both MSE and MAE. Thus each component contributes overall, although none yields a uniform gain on every metric and dataset.

Which correction contexts matter? Table 3 removes the indicated context components at inference time. ETTm2, Solar, and Weather show that removing either individual context worsens every horizon. Averaged over the four horizons, removing the current context increases MSE by 0.005–0.014, while removing the retrieved detail context increases it by 0.004–0.012. Removing both contexts also degrades all three datasets, by 0.005–0.019. Thus the two views provide complementary evidence rather than an inactive auxiliary path.

Table 3: Context-component ablation on representative datasets. Entries are ΔMSE ($\times 10^{-3}$), averaged over four horizons. Positive values mean that removing the indicated component hurts.

Component	ETTm2	Solar	Weather
w/o Current context	+5.25	+8.64	+13.82
w/o Detail context	+5.89	+4.20	+11.77
w/o Both	+5.37	+9.46	+18.85

Does the view router depend on forecast position? Figure 4 shows representative full patch-wise router trajectories for Weather and Solar. Across all 32 dataset–horizon cells, the detail-view weight varies by at least 0.02 in 24 cells. Among those 24, 21 assign more detail weight farther into the forecast and three assign less. The trajectories show both smooth horizon-dependent routing and non-monotone, patch-specific routing, supporting an adaptive mixture across the forecast.

4.3 Transferring the Correction to Other Backbones

We add the Dual-Context Correction from DuoTS to SRSNet, TimeKAN, and PatchTST. Across the 24 backbone–dataset–horizon cells in Table 4, the correction lowers MSE in 15. The transfer is not uniform. On ETTm1, adding the correction to the strong SRSNet backbone reduces MSE from 0.426 to 0.423 at $H = 720$. However, only two of the six ETTm1 cells improve, while four become worse. The 15-minute sampling interval may help explain this result. With $L = 512$, adjacent patches densely cover roughly 5.3 days and may yield highly redundant retrieved contexts. This interpretation remains a hypothesis rather than causal evidence. Full DuoTS has the lowest MSE in four of the eight dataset–horizon settings.

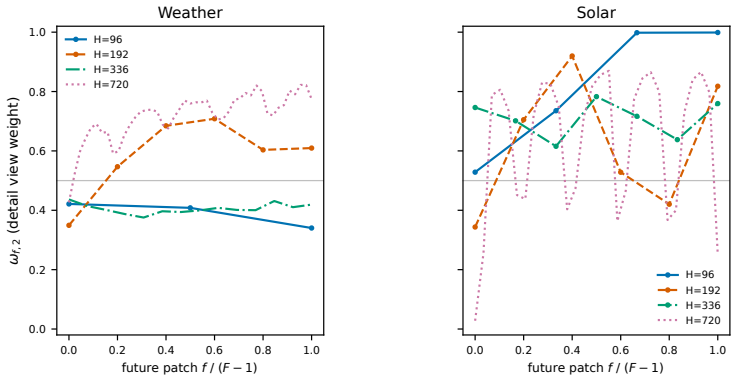


Figure 4: **Full patch-wise view-weight profiles.** Mean weight $\omega_{f,2}$ assigned to the retrieved-detail view at future patch f for Weather and Solar. The current view receives $1-\omega_{f,2}$. Curves correspond to forecast horizons.

Table 4: Correction transfer. Within each arrow, the lower endpoint is bold. The DuoTS column is from Table 1 and is bold when it is lower than all three correction-equipped backbones. The second-lowest value among the three correction-equipped endpoints and Full DuoTS is underlined.

		Original → +Dual-Context Correction from DuoTS			
Dataset	H	SRSNet	TimeKAN	PatchTST	Full DuoTS
ETTh2	96	0.271 → **0.270**	0.280 → **0.267**	**0.274** → 0.275	**0.263**
	720	0.391 → **0.382**	0.420 → **0.409**	0.406 → **0.393**	0.383
Weather	96	0.147 → **0.145**	0.151 → **0.141**	0.150 → **0.143**	0.142
	720	0.325 → **0.314**	0.317 → **0.315**	**0.312** → 0.315	**0.312**
Electricity	96	0.131 → **0.130**	**0.135** → 0.135	0.143 → **0.129**	**0.127**
	720	**0.201** → 0.206	**0.206** → 0.215	0.214 → **0.205**	0.206
ETTm1	96	**0.288** → 0.295	0.290 → **0.287**	**0.289** → 0.296	**0.284**
	720	0.426 → **0.423**	**0.401** → 0.405	**0.416** → 0.425	0.419

5 Conclusion

DuoTS combines a parallel patch backbone with a horizon-specific dual-context correction. Each future patch balances recent information with retrieved analogs and their observed successors to refine the base forecast. Across eight long-term benchmarks, DuoTS obtains the lowest MSE in 22 and the lowest MAE in 23 of 32 dataset–horizon settings. The ablation results show that the second-order dynamics, channel mixing, and intra-patch mixing are generally useful, while context-component ablation confirms that both correction views affect the forecast. Gains across SRSNet, TimeKAN, and PatchTST further support the correction as a model-agnostic module for forecasting architectures.

Limitations & Future Work. DuoTS is limited to analogs within the finite look-back window, and weak or duplicate matches can make the detail evidence uninformative. Its fixed-width projections and patch-wise correction also add overhead on small-channel series, while selected lags represent latent evidence rather than causal local analogs. Future work will add diversity-aware retrieval and abstention, extend memory beyond the input window, and develop lighter corrections with more interpretable local matches and successor tests.

AI USE STATEMENT

Generative AI tools were used to draft portions of the manuscript, revise and polish the writing, and assist with the retrieval and discovery of related work. The authors checked the suggested literature against the original sources and manually reviewed and revised all AI-assisted text. Generative AI tools also supported checks of implementation consistency, experimental setup, and the completeness of the model and accompanying code. The authors made the research decisions, designed the model and experiments, and reviewed and validated the final implementation and results. Generative AI tools did not generate datasets or independently interpret experimental results. The authors take responsibility for the paper's text, claims, citations, code, and artifacts.

REPRODUCIBILITY STATEMENT

The DuoTS architecture and its implementation-level computation are specified in Section 3; Appendix A gives pseudocode for the complete dual-context correction. Section 4 reports the experiments, while Appendices C and D document their setup. The anonymous GitHub repository supplied with the submission contains the code and configurations needed to reproduce the reported results.

REFERENCES

Ibram Abdelmalak, Kiran Madhusudhanan, Jungmin Choi, Christian Klötergens, Vijaya Krishna Yalavarthi, Maximilian Stubbemann, and Lars Schmidt-Thieme. Channel dependence, limited lookback windows, and the simplicity of datasets: How biased is time series forecasting? In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 585–597. Springer, 2026.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019.

California Department of Transportation (Caltrans). Performance measurement system (PeMS). `https://pems.dot.ca.gov/`, 2024. Accessed: 2024-05-20.

Fanpu Cao, Lu Dai, Jindong Han, and Hui Xiong. Enhancing multivariate time series forecasting with global temporal retrieval. In *International Conference on Learning Representations*, 2026.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=oANkBaVci5`.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations*, 2024.

Si-An Chen, Chun-Liang Li, Sercan O Arik, Nathanael Christian Yoder, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023.

Dazhao Du, Tao Han, and Song Guo. Predicting the future by retrieving the past. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20896–20904, 2026. doi: 10.1609/aaai.v40i25.39230.

Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. TSMixer: Lightweight MLP-mixer model for multivariate time series forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, pp. 459–469, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599533. URL `https://doi.org/10.1145/3580305.3599533`.

Jingru Fei, Kun Yi, Wei Fan, Qi Zhang, and Zhendong Niu. Amplifier: Bringing attention to neglected low-energy components in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11645–11653, 2025. doi: 10.1609/aaai.v39i11.33267.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21774–21797, 2025.

Qihe Huang, Lei Shen, Ruixin Zhang, Jiahuan Cheng, Shouhong Ding, Zhengyang Zhou, and Yang Wang. Hdmixer: Hierarchical dependency with extendable patch for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12608–12616, 2024.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. TimeKAN: KAN-based frequency decomposition learning architecture for long-term time series forecasting. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wTLc79YNbh.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. CycleNet: Enhancing time series forecasting through modeling periodic patterns. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106315–106345, 2024.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2766–2786, 2024a. doi: 10.52202/079017-0091.

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*, 2024b. URL https://openreview.net/forum?id=JePfAI8fah.

Edward N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646, 1969.

Luyi Ma, Wanjia Zhang, Kai Zhao, Abhishek Kulkarni, Lalitesh Morishetti, Anjana Ganesh, Ashish Ranjan, Aashika Padmanabhan, Jianpeng Xu, Jason Cho, Praveen Kanumala, Kaushiki Nag, Sumit Dutta, Kamiya Motwani, Malay Patel, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. GRACE: Generative recommendation via journey-aware sparse attention on chain-of-thought tokenization. In *Proceedings of the 19th ACM Conference on Recommender Systems*, pp. 135–144, 2025. doi: 10.1145/3705328.3748056.

Max Planck Institute for Biogeochemistry. Jena climate dataset. https://www.bgc-jena.mpg.de/wetter/, 2024. Accessed: 2024-05-20.

National Renewable Energy Laboratory. Solar power data for integration studies. https://www.nrel.gov/grid/solar-power-data.html, 2024. Accessed: 2024-05-20.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Y. 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. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C58C86.

H. M. van den Dool. Searching for analogues, how long must we wait? *Tellus A: Dynamic Meteorology and Oceanography*, 46(3):314–324, 1994.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Xingjian Wu, Xiangfei Qiu, Hanyin Cheng, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Enhancing time series forecasting through selective representation spaces: A patch perspective. In *Advances in Neural Information Processing Systems*, volume 38, pp. 23328–23354, 2025. doi: 10.52202/085713-0789.

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. doi: 10.1609/aaai.v37i9.26317.

Huiliang Zhang, Ping Nie, Lijun Sun, and Benoit Boulet. Nearest neighbor multivariate time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):12606–12618, 2025. doi: 10.1109/TNNLS.2024.3490603.

Yuanzhao Zhang and William Gilpin. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. In *International Conference on Learning Representations*, 2026.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Shiqiao Zhou, Holger Schöner, Zipeng Wu, Edouard Fouché, IAG Wilson, and Shuo Wang. Stationarity-aware retrieval-augmented time series forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3817813.

A Correction Procedure in Pseudocode

Algorithm 1 writes the implementation as a loop over future patches while remaining vectorized over channels. A block is valid if it contains at least one historical patch whose requested successors are all observed. $\mathrm{TopKValid}(s, K)$ returns the indices of the $\min(K, \text{number of valid blocks})$ highest-scoring valid blocks. Masked blocks are never returned. $\mathrm{Eligible}(\mathcal{J}, K_{\mathrm{suc}}, \nu)$ then returns the patch indices assigned to blocks in $\mathcal{J}$ that satisfy $n + \ell\nu \le N$ for every $\ell = 1, \dots, K_{\mathrm{suc}}$.

$\mathrm{Top2Margin}$ ignores masked scores and returns the best valid score minus the second-best valid score. If only one valid score exists, it uses a large positive sentinel margin, so the sole valid match is not compared with a masked entry. $\mathrm{Conf}(m; \varphi, \psi)$ is its learned sigmoid calibration, $\sigma(\exp(\varphi)m + \psi)$. $\mathrm{DiverseTopK}_\rho$ greedily selects the highest-scoring remaining candidate more than ρ patch positions from every previous selection. If none satisfies that constraint, it selects the highest-scoring remaining valid candidate. Thus it never chooses a masked candidate or a duplicate when enough valid candidates exist. $\mathrm{MaskedSoftmax}$ assigns zero weight to any unavailable selected slot. Finally, $\mathrm{LeftPad}$ and Block create the consecutive width-B_{blk} blocks, $\mathrm{TransitionValue}$ is defined in equation 7, and $\mathrm{Truncate}_H$ keeps the first H entries after the correction patches are concatenated in future order.

Algorithm 1 Dual-context analog correction for one sample

Require: tokens $\mathbf{U}_{c,1:N}$ for $c = 1, \dots, C$, horizon embeddings $\mathbf{e}^{\mathrm{hor}}_{1:F}$
Ensure: normalized correction $\widehat{\mathbf{E}} \in \mathbb{R}^{C \times H}$

$\nu \leftarrow \lceil P/S \rceil$
$\mathbf{G}_{1:M} \leftarrow \mathrm{Block}(\mathrm{LeftPad}(\mathbf{U}_{1:N}), B_{\mathrm{blk}})$
$\mathbf{b}^0_m \leftarrow \mathrm{MLP}_{\mathrm{blk}}(\mathrm{vec}(\mathbf{G}_m)) + \mathbf{e}^{\mathrm{blk}}_m$ for $m = 1, \dots, M$
$\mathbf{b}_{1:M} \leftarrow \mathbf{b}^0_{1:M} + \lambda_{\mathrm{blk}} \mathrm{MLP}_{\mathrm{baxis}}(\mathbf{b}^0_{1:M})$
for $f = 1$ **to** F **do**
  $\boldsymbol{\xi}_f \leftarrow \mathbf{U}_N + \mathbf{e}^{\mathrm{hor}}_f$
  $\mathbf{h}^{\mathrm{cur}}_f \leftarrow \mathrm{CosineAttention}(\boldsymbol{\xi}_f, \mathbf{U}_{N-K_{\mathrm{cur}}+1:N})$
  $s^{\mathrm{blk}}_{f,m} \leftarrow \tau_{\mathrm{blk}} \mathrm{CosSim}(\mathbf{W}^{\mathrm{blk}}_q \boldsymbol{\xi}_f, \mathbf{W}^{\mathrm{blk}}_k \mathbf{b}_m)$ for $m = 1, \dots, M$
  $\mathcal{J}_f \leftarrow \mathrm{TopKValid}(s^{\mathrm{blk}}_{f,1:M}, K_{\mathrm{blk}})$
  $m^{\mathrm{blk}}_f \leftarrow \mathrm{Top2Margin}(s^{\mathrm{blk}}_{f,1:M})$
  $\chi^{\mathrm{blk}}_f \leftarrow \mathrm{Conf}(m^{\mathrm{blk}}_f; \varphi^{\chi}_{\mathrm{blk}}, \psi^{\chi}_{\mathrm{blk}})$
  $\mathcal{C}_f \leftarrow \mathrm{Eligible}(\mathcal{J}_f, K_{\mathrm{suc}}, \nu)$
  $s^{\mathrm{det}}_{f,n} \leftarrow \tau_{\mathrm{det}} \mathrm{CosSim}(\mathbf{W}^{\mathrm{det}}_q \boldsymbol{\xi}_f, \mathbf{W}^{\mathrm{det}}_k \mathbf{U}_n)$ for $n \in \mathcal{C}_f$
  $\mathcal{I}_f \leftarrow \mathrm{DiverseTopK}_\rho(s^{\mathrm{det}}_{f,\mathcal{C}_f}, K_{\mathrm{det}})$
  for $k = 1$ **to** K_{det} **do**
    $n^{(k)}_f \leftarrow \mathcal{I}_f[k]$
    $\mathbf{v}^{(k)}_f \leftarrow \mathrm{TransitionValue}(\mathbf{U}, n^{(k)}_f, \nu, K_{\mathrm{suc}}, \boldsymbol{\xi}_f)$
  end for
  $\boldsymbol{\pi}_f \leftarrow \mathrm{MaskedSoftmax}(s^{\mathrm{det}}_{f,\mathcal{I}_f})$
  $\mathbf{h}^{\mathrm{det}}_f \leftarrow \chi^{\mathrm{blk}}_f \sum_{k=1}^{K_{\mathrm{det}}} \pi^{(k)}_f \mathbf{v}^{(k)}_f$
  if $K_{\mathrm{det}} = 1$ **then**
    $m^{\mathrm{det}}_f \leftarrow \mathrm{Top2Margin}(s^{\mathrm{det}}_{f,\mathcal{C}_f})$
    $\mathbf{h}^{\mathrm{det}}_f \leftarrow \mathrm{Conf}(m^{\mathrm{det}}_f; \varphi^{\chi}_{\mathrm{det}}, \psi^{\chi}_{\mathrm{det}}) \mathbf{h}^{\mathrm{det}}_f$
  end if
  $\boldsymbol{\omega}_f \leftarrow \mathrm{softmax}(\mathrm{MLP}_{\mathrm{view}}([\boldsymbol{\xi}_f \| \mathbf{e}^{\mathrm{hor}}_f]))$
  $\mathbf{r}_f \leftarrow \mathrm{MLP}_{\mathrm{dec}}(\omega_{f,1} \mathbf{h}^{\mathrm{cur}}_f + \omega_{f,2} \mathbf{h}^{\mathrm{det}}_f)$
end for
$\widehat{\mathbf{E}} \leftarrow \mathrm{Truncate}_H(\mathrm{Concat}(\mathbf{r}_1, \dots, \mathbf{r}_F))$
return $\widehat{\mathbf{E}}$

B Sparse Channel Mixer Details

The main text retains only the channel residual update. For each channel c, let $\mathbf{e}_c^q, \mathbf{e}_c^k \in \mathbb{R}^{d_g}$ be learned query and key embeddings of graph width d_g, and let $\beta \in \mathbb{R}$ be the learned self-channel logit. The omitted graph construction is

$$\begin{aligned} \mathbf{q}_{c,n} &= \mathbf{p}_{c,n} + \mathrm{MLP}_{\mathrm{in}}(\mathbf{p}_{c,n}), \\ s_{c,j} &= \frac{(\mathbf{e}_c^q)^\top \mathbf{e}_j^k}{\sqrt{d_g}}, \qquad \mathcal{N}_c = \mathrm{TopK}_{j\neq c}(s_{c,j}, \kappa_{\mathrm{ch}}), \\ (a_{c,c}, (a_{c,j})_{j\in\mathcal{N}_c}) &= \mathrm{softmax}\left(\beta, (s_{c,j})_{j\in\mathcal{N}_c}\right), \quad \mathbf{m}_{c,n} = \sum_{j\in\mathcal{N}_c\cup\{c\}} a_{c,j}\mathbf{q}_{j,n}. \end{aligned} \tag{15}$$

The residual coefficient γ_{ch} is also learned. The graph is shared across samples and patch positions. Section 3.2 describes how its consensus enters the residual channel update.

C Reproducibility Details

C.1 Benchmarks and Their Statistics

Table 5 summarizes the eight public benchmarks. The four ETT datasets contain seven transformer load and oil-temperature signals. ETTh1 and ETTh2 are sampled hourly, while ETTm1 and ETTm2 use 15-minute intervals (Zhou et al., 2021). Weather contains 21 meteorological variables recorded every ten minutes (Max Planck Institute for Biogeochemistry, 2024). Electricity reports hourly demand from 321 clients (Trindade, 2015), Traffic contains hourly occupancy from 862 freeway sensors (California Department of Transportation (Caltrans), 2024), and Solar records ten-minute generation from 137 photovoltaic plants (National Renewable Energy Laboratory, 2024).

C.2 Reported Error Measures

Let i index the n test windows, h the forecast steps, and c the channels. We report

$$\mathrm{MSE} = \frac{1}{nHC}\sum_{i,h,c}\left(\widehat{Y}_{i,h,c} - Y_{i,h,c}\right)^2, \qquad \mathrm{MAE} = \frac{1}{nHC}\sum_{i,h,c}\left|\widehat{Y}_{i,h,c} - Y_{i,h,c}\right|. \tag{16}$$

We report MSE and MAE averaged over all test windows, forecast steps, and channels.

D Evaluation Setup Details

D.1 Procedure and Search Space

We use Optuna (Akiba et al., 2019) with the Tree-structured Parzen Estimator, a multivariate group sampler, 15 startup trials, and a fixed sampler seed. A separate study is run for each benchmark, look-back candidate, and forecast horizon. Each study minimizes the best validation MSE reached during training. Most studies use 50 trials, while a small number use 20 to 30 trials. The configuration with the lowest validation MSE is evaluated on the test split. Table 6 lists the complete search space.

Two constraints keep sampled configurations feasible. A trial is pruned when its requested successor span or current-context window exceeds the available patch grid. The exclusion radius is sampled only when more than one detail patch is retrieved because a single match does not require a diversity constraint.

D.2 Selected Configurations

Tables 7 and 8 report the best configuration for every benchmark and forecast horizon.

Table 5: Statistics of the eight long-term forecasting benchmarks. C is the number of channels and T is the number of observations. Dataset partitions follow chronological order.

Benchmark	Recorded quantity	C	T	Interval	Train / val / test
ETTh1 (Zhou et al., 2021)	Transformer load and oil temperature	7	17,420	1 h	8,640 / 2,880 / 2,880
ETTh2 (Zhou et al., 2021)	Transformer load and oil temperature	7	17,420	1 h	8,640 / 2,880 / 2,880
ETTm1 (Zhou et al., 2021)	Transformer load and oil temperature	7	69,680	15 min	34,560 / 11,520 / 11,520
ETTm2 (Zhou et al., 2021)	Transformer load and oil temperature	7	69,680	15 min	34,560 / 11,520 / 11,520
Weather (Max Planck Institute for Biogeochemistry, 2024)	Meteorological indicators	21	52,696	10 min	36,887 / 5,270 / 10,539
Electricity (Trindade, 2015)	Client-level power demand	321	26,304	1 h	18,412 / 2,632 / 5,260
Traffic (California Department of Transportation (Caltrans), 2024)	Freeway sensor occupancy	862	17,544	1 h	12,280 / 1,756 / 3,508
Solar (National Renewable Energy Laboratory, 2024)	Photovoltaic plant output	137	52,560	10 min	36,792 / 5,256 / 10,512

Table 6: Unified hyperparameter search space for DuoTS. Derived values are computed from the sampled coordinates, and retrieval ranges are capped by the available patch grid.

Coordinate	Symbol	Range, cap, or fixed value
Sampled		
Learning rate	η	log-uniform, $[2 \times 10^{-5}, 3 \times 10^{-3}]$
Path width	d	$\{32, 64, 128, 256, 384, 512\}$
Mixer hidden width	d_{ff}	$\{16, 32, 64, 128, 256, 512, 1024\}$
Structured-path layers	Λ	$\{1, 2, 3\}$
Dropout		uniform, $[0, 0.40]$
Patch length	P	$\{12, 16, 24, 32, 48, 64, 96, 112, 128, 144, 168, 192, 256\}$, capped at L
Stride ratio	r	$\{0.30, 0.50, 0.75\}$
Journey block width	B_{blk}	$\{2, 3, 4, 6, 8\}$, capped at N
Current-context window	K_{cur}	$\{2, 3, 4, 6\}$, capped at N
Successor span	K_{suc}	integer in $[1, 4]$, capped at $\lfloor (N-1)/\nu \rfloor$
Block retrieval budget	K_{blk}	$\{1, 2\}$
Patch retrieval budget	K_{det}	$\{1, 2, 3\}$
Exclusion radius	ρ	$\{1, 2, 3, 4\}$, sampled only when $K_{\text{det}} > 1$
Channel neighbourhood	κ_{ch}	$\{0, 1, 2, 4, 6, 8, 12, 16, 20, 32, 64, 128\}$, capped at $C - 1$
Channel graph width	d_g	$\{4, 8, 16, 32, 64\}$
Initial channel gate		log-uniform, $[10^{-4}, 5 \times 10^{-3}]$
Initial self logit	β	uniform, $[1, 4]$
Initial correction gate		$\{0, 0.01, 0.03, 0.05\}$
Initial patch-axis gate		$\{0.025, 0.05, 0.10, 0.20\}$
Derived or fixed		
Stride	S	$\max(1, \text{round}(rP))$, derived
Patch count	N	$\lceil \max(0, L - P)/S \rceil + 1$, derived
Successor spacing	ν	$\lceil P/S \rceil$, derived
Correction heads		fixed at 4
Fusion-gate initialization	λ_{fuse}	fixed at 0 and learned thereafter
Weight decay		fixed at 0
Normalization		fixed, RevIN without a learned affine map

Table 7: Backbone settings of the reported DuoTS models. The look-back length L is chosen per cell from $\{96, 336, 512\}$, P and S are the patch length and stride, d is the shared path width, d_{ff} the hidden width of the mixer MLPs, and Λ the number of structured-path layers. Dropout is shared by the backbone MLPs, the channel mixer, and the correction attention. η is the Adam learning rate and B the batch size.

Benchmark	H	L	P	S	d	d_{ff}	Λ	Dropout	η	B
ETTh1	96	336	16	8	512	256	2	0.274	2.585×10^{-4}	32
ETTh1	192	512	32	10	384	64	2	0.200	1.601×10^{-4}	32
ETTh1	336	512	112	56	384	256	2	0.050	9.300×10^{-5}	32
ETTh1	720	336	16	8	128	128	2	0.256	2.443×10^{-4}	32
ETTh2	96	512	48	14	256	256	1	0.200	2.324×10^{-3}	256
ETTh2	192	512	24	7	32	32	1	0.300	2.299×10^{-4}	256
ETTh2	336	512	48	24	128	512	2	0.200	1.720×10^{-4}	256
ETTh2	720	336	32	16	32	16	2	0.116	9.332×10^{-4}	32
ETTm1	96	336	12	6	256	256	2	0.252	9.929×10^{-4}	32
ETTm1	192	512	24	18	384	256	1	0.300	5.821×10^{-5}	128
ETTm1	336	512	32	16	256	128	2	0.400	1.397×10^{-4}	256
ETTm1	720	336	16	8	64	1024	2	0.239	1.590×10^{-4}	32
ETTm2	96	336	12	6	256	16	2	0.339	5.270×10^{-5}	32
ETTm2	192	512	12	6	64	128	2	0.400	9.969×10^{-5}	256
ETTm2	336	336	12	9	128	512	3	0.200	4.689×10^{-5}	128
ETTm2	720	336	32	16	256	256	2	0.001	1.154×10^{-4}	32
Weather	96	512	32	16	128	32	2	0.304	6.539×10^{-4}	64
Weather	192	512	32	10	128	32	2	0.300	6.472×10^{-4}	64
Weather	336	512	24	18	256	256	2	0.400	7.906×10^{-4}	64
Weather	720	512	16	8	32	64	2	0.244	3.245×10^{-4}	32
Electricity	96	512	48	24	256	32	2	0.068	3.719×10^{-4}	32
Electricity	192	512	112	84	256	16	1	0.300	1.151×10^{-3}	16
Electricity	336	512	112	34	32	512	2	0.050	8.048×10^{-4}	16
Electricity	720	512	16	8	32	128	2	0.159	8.650×10^{-4}	32
Traffic	96	512	128	38	256	128	2	0.200	1.811×10^{-3}	16
Traffic	192	512	96	72	256	16	2	0.250	2.107×10^{-3}	16
Traffic	336	512	48	24	64	512	2	0.250	2.706×10^{-3}	8
Traffic	720	512	128	96	256	128	2	0.150	2.333×10^{-3}	8
Solar	96	512	24	18	512	16	1	0.050	4.754×10^{-4}	16
Solar	192	512	32	10	256	16	3	0.100	5.044×10^{-4}	16
Solar	336	512	48	14	32	256	1	0.400	6.104×10^{-4}	16
Solar	720	512	24	7	64	32	1	0.000	5.151×10^{-4}	16

Table 8: Correction and channel settings of the reported DuoTS models. B_{blk} is the journey block width, K_{cur} the number of newest tokens read by the current context, K_{suc} the successor span of a retrieved analog, K_{blk} and K_{det} the block and patch retrieval budgets, and ρ the exclusion radius, which is inactive whenever $K_{\mathrm{det}} = 1$. The channel columns give the neighbourhood size κ_{ch}, the graph width d_g, the initial self logit β, and the initial channel gate. The final two columns give the initial correction gate and the initial patch-axis gate.

Benchmark	H	B_{blk}	K_{cur}	K_{suc}	K_{blk}	K_{det}	ρ	κ_{ch}	d_g	β	Gate init	Corr. init	Patch init
ETTh1	96	3	4	4	1	1	0	0	4	1.27	3.03×10^{-3}	0.000	0.100
ETTh1	192	4	4	3	2	1	0	0	8	2.23	8.08×10^{-4}	0.000	0.200
ETTh1	336	4	2	1	1	2	1	1	8	2.64	2.35×10^{-3}	0.000	0.200
ETTh1	720	6	3	3	2	1	0	0	4	3.01	4.90×10^{-3}	0.000	0.100
ETTh2	96	4	6	3	1	1	0	2	4	1.71	8.45×10^{-4}	0.010	0.050
ETTh2	192	4	2	3	2	1	0	4	4	2.29	6.25×10^{-4}	0.000	0.100
ETTh2	336	6	6	3	2	3	4	0	32	1.22	1.01×10^{-4}	0.050	0.100
ETTh2	720	2	3	4	2	2	2	4	4	2.02	3.66×10^{-4}	0.000	0.100
ETTm1	96	6	3	3	1	1	0	0	32	3.88	1.75×10^{-3}	0.000	0.100
ETTm1	192	2	6	2	2	1	0	2	8	2.74	1.67×10^{-4}	0.050	0.050
ETTm1	336	6	2	2	1	2	2	4	4	3.96	1.44×10^{-4}	0.050	0.100
ETTm1	720	3	4	2	2	2	4	0	32	3.87	2.83×10^{-4}	0.000	0.100
ETTm2	96	4	3	3	2	3	1	2	32	3.67	3.17×10^{-4}	0.000	0.100
ETTm2	192	3	2	2	1	3	3	4	4	3.10	5.42×10^{-4}	0.050	0.100
ETTm2	336	8	3	4	1	3	3	4	8	1.51	2.06×10^{-4}	0.050	0.200
ETTm2	720	2	2	1	1	3	3	0	32	1.37	4.49×10^{-4}	0.000	0.100
Weather	96	6	2	3	1	2	3	16	64	2.47	2.79×10^{-3}	0.000	0.100
Weather	192	2	4	4	2	2	4	12	4	2.42	1.99×10^{-4}	0.010	0.025
Weather	336	4	3	2	2	1	0	12	32	2.17	6.73×10^{-4}	0.010	0.025
Weather	720	3	2	3	1	1	0	8	8	1.87	2.05×10^{-3}	0.000	0.100
Electricity	96	3	4	4	1	3	2	128	8	3.64	1.01×10^{-4}	0.000	0.100
Electricity	192	3	3	1	1	1	0	64	32	1.93	6.30×10^{-4}	0.050	0.200
Electricity	336	3	3	2	2	2	3	0	4	1.29	2.19×10^{-4}	0.030	0.200
Electricity	720	6	2	2	2	1	0	8	32	2.96	1.27×10^{-3}	0.000	0.100
Traffic	96	3	3	2	1	2	3	2	32	1.06	1.51×10^{-3}	0.030	0.100
Traffic	192	6	2	1	2	3	4	8	16	1.88	2.22×10^{-4}	0.010	0.025
Traffic	336	2	6	4	2	3	2	4	16	2.72	2.75×10^{-4}	0.010	0.050
Traffic	720	4	4	2	1	1	0	2	16	3.24	3.56×10^{-3}	0.010	0.100
Solar	96	2	3	2	2	1	0	16	16	3.21	1.66×10^{-4}	0.010	0.100
Solar	192	8	3	4	2	1	0	16	64	1.08	2.31×10^{-3}	0.000	0.025
Solar	336	6	3	3	1	3	4	8	16	2.11	4.36×10^{-4}	0.030	0.050
Solar	720	6	2	4	2	2	3	16	32	1.57	1.01×10^{-3}	0.050	0.050

D.3 Meaning of Each Searched Coordinate

The current-context window must fit the patch grid ($K_{\mathrm{cur}} \leq N$), and the successor span is capped at $\lfloor (N-1)/\nu \rfloor$ so retrieved patches have observed continuations. Retrieval budgets are capped by the available blocks and eligible patches; the exclusion radius ρ is sampled only for $K_{\mathrm{det}} > 1$. Setting $\kappa_{\mathrm{ch}} = 0$ disables cross-channel mixing. The initial channel, correction, and patch-axis gates are searched as listed in Table 6; the underlying parameter of λ_{fuse} in equation 14 is fixed at zero initially and then learned.

E Comparison with Retrieval-Augmented Forecasters

Table 9 collects the complete reported results of RAFT and SARAF on the seven datasets shared with DuoTS. The table retains every forecast horizon and the dataset averages. The source papers use a look-back of 720 whereas DuoTS selects its look-back per cell. Every selected DuoTS look-back is shorter at 96, 336, or 512. Despite using less historical context, DuoTS has the lowest reported MSE in 18 of the 28 dataset–horizon rows and the lowest or tied-lowest MAE in 22. These values are still a contextual cross-paper comparison rather than a controlled rerun.

Table 9: Reported full-horizon comparison with retrieval-augmented forecasters (lower is better). DuoTS uses the per-cell look-back selected from $\{96, 336, 512\}$. RAFT and SARAF use $L = 720$. RAFT values are from Tables 16–17 of Han et al. (2025), and SARAF values are from Table 4 of Zhou et al. (2026). These are cross-paper reported values rather than a controlled rerun. Within the three displayed methods, the best value is bold and the second-best is underlined.

		DuoTS		RAFT		SARAF	
Dataset	H	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.363**	**0.388**	0.367	0.397	0.371	0.399
	192	**0.400**	**0.413**	0.411	0.427	0.404	0.420
	336	**0.401**	**0.430**	0.436	0.442	0.435	0.447
	720	**0.441**	**0.457**	0.467	0.478	0.451	0.473
	Avg	**0.401**	**0.422**	0.420	0.436	0.415	0.435
ETTh2	96	**0.263**	**0.332**	0.276	0.344	0.271	0.335
	192	**0.312**	**0.367**	0.347	0.393	0.331	0.378
	336	**0.314**	**0.377**	0.376	0.425	0.359	0.408
	720	**0.383**	**0.424**	0.436	0.473	0.399	0.447
	Avg	**0.318**	**0.375**	0.359	0.409	0.340	0.392
ETTm1	96	**0.284**	**0.342**	0.302	0.349	0.297	0.348
	192	0.334	**0.366**	**0.329**	0.367	0.331	0.370
	336	0.364	**0.383**	**0.355**	**0.383**	0.357	0.386
	720	0.419	0.419	0.406	0.413	**0.401**	**0.410**
	Avg	0.350	**0.377**	0.348	0.378	**0.346**	0.378
ETTm2	96	0.166	0.255	0.164	0.256	**0.160**	**0.252**
	192	0.220	0.293	0.219	0.296	**0.215**	**0.292**
	336	0.278	0.330	0.275	0.336	**0.268**	**0.329**
	720	0.364	0.385	0.359	0.392	**0.349**	**0.383**
	Avg	0.257	0.316	0.254	0.320	**0.248**	**0.314**
Electricity	96	**0.127**	**0.222**	0.133	0.232	0.130	0.228
	192	0.146	**0.241**	0.149	0.247	**0.143**	**0.241**
	336	0.162	**0.258**	0.161	0.259	**0.160**	0.260
	720	0.206	0.296	0.197	0.297	**0.173**	**0.274**
	Avg	0.160	0.254	0.160	0.259	**0.152**	**0.251**
Traffic	96	**0.355**	**0.252**	0.378	0.273	0.371	0.267
	192	**0.380**	**0.259**	0.391	0.277	0.385	0.274
	336	**0.386**	**0.262**	0.402	0.282	0.398	0.282
	720	**0.424**	**0.284**	0.434	0.297	0.425	0.295
	Avg	**0.387**	**0.264**	0.402	0.282	0.395	0.280
Solar	96	**0.162**	**0.218**	0.192	0.251	0.182	0.244
	192	**0.177**	**0.236**	0.247	0.323	0.204	0.254
	336	**0.185**	**0.241**	0.240	0.300	0.216	0.263
	720	**0.196**	**0.250**	0.246	0.311	0.224	0.270
	Avg	**0.180**	**0.236**	0.231	0.296	0.206	0.258

F Retrieval and Successor Evidence

We apply four inference-time perturbations to trained DuoTS checkpoints, modifying only detail retrieval. *Random eligible* replaces the top-scoring selection with candidates drawn uniformly from the same valid search pool. *Bottom-K* selects the lowest-scoring valid candidates. *No diversity* sets the exclusion radius ρ to zero while keeping K_{det} unchanged; it removes the diversity constraint and does not reduce retrieval to a single nearest match. *Successor-index shuffle* keeps the selected candidates but permutes the successor-index tensor across samples in a batch before successor states are gathered, so a retained candidate is paired with successor positions drawn from another sample. These controls target candidate ranking, the diversity constraint, and the alignment between a candidate and its continuation, respectively. These perturbations are diagnostic rather than causal.

Random eligible changes 27 of the 32 settings and increases MSE in 25 of them, with a mean increase of 0.00070. Bottom-K changes 27 settings and increases MSE in 22, with a mean increase of 0.00135. Replacing or inverting the learned ranking therefore degrades the forecast in most settings, indicating that the ranking supplies useful retrieval evidence. Successor-index shuffle changes 28 settings and increases MSE in 19, with a smaller mean increase of 0.00036, which suggests that correctly aligned continuations also contribute, although less uniformly. No diversity changes only five settings (two worse, three better) and shows no consistent effect. Overall, the correction depends mainly on meaningful candidate selection and benefits additionally from the successor evidence attached to each candidate. Changes are counted on unrounded values for all settings except Weather at $H = 336$, for which only four-decimal values are available and all five entries coincide; it is counted as unchanged. All results use a single seed, and the effects are small, so we interpret the aggregate pattern rather than individual differences.

G Additional Mechanism Diagnostics

Effect of the correction on the base forecast. Table 11 compares the direct base prediction with the corrected output of the same trained DuoTS model. Based on the displayed three-decimal errors, the correction reduces MSE in 22 of 32 dataset–horizon settings and MAE in 21 of 32. Both metrics improve at every horizon on ETTm2, Solar, and Weather. Electricity MSE improves at every horizon, while its MAE improves at three and ties at $H = 720$. The largest MSE reduction is 14.970% on Weather at $H = 96$. ETTh2 MSE rises at three horizons and ties at $H = 192$. This paired comparison measures the correction's contribution within the jointly trained model.

Base forecast against the baselines. Table 12 repeats the comparison of Table 1 with the full DuoTS forecast replaced by its base forecast. Without the correction, the base forecast has the lowest MSE in 14 of the 32 dataset–horizon settings and the lowest MAE in 11, compared with 22 and 23 for the full DuoTS forecast. SRSNet obtains 16 and 15 in this comparison. The correction therefore accounts for most of the settings in which DuoTS leads, particularly on Weather and Solar.

Learned correction scale. Across all 430 dataset–horizon–future-patch combinations, the mean absolute correction scale is 0.040. The horizon averages are 0.041 on Solar, 0.056 on Weather, and 0.031 on Traffic. Eight of the 32 cells learn a negative scale at every future patch.

Confidence-map calibration. Of 64 dataset–horizon–level pairs, seven stay exactly at their zero initialization and 37 remain within 0.05 of it. At $H = 336$, the journey-level (s, b) parameters are $(-0.074, -0.629)$ on Solar and $(-0.387, -0.603)$ on Traffic, while Weather learns nonzero maps at both levels.

Where does the correction help? Table 13 compares the corrected and base forecasts on Solar, Traffic, and Weather. Solar and Weather have lower displayed MSE at all four horizons; Traffic improves at three and ties at $H = 192$. Weather improves at every forecast step. Solar's gain is distributed across nearly all channels, and Traffic also improves across most steps and channels.

Table 10: Detail-retrieval perturbations on all 32 dataset–horizon settings (test MSE). Bold marks the lowest value in each row at the displayed precision; ties are all bold.

Dataset	H	Default	Random eligible	Bottom-K	No diversity	Successor-index shuffle	Dataset	H	Default	Random eligible	Bottom-K	No diversity	Successor-index shuffle
ETTh1	96	0.3634	0.3666	0.3690	0.3634	**0.3632**	Weather	96	**0.1428**	0.1432	0.1444	**0.1428**	0.1429
	192	0.4009	0.4027	0.4080	0.4009	**0.3989**		192	0.1889	0.1905	0.1926	0.1889	**0.1888**
	336	0.4017	0.4017	**0.4015**	0.4017	**0.4015**		336	**0.2372**	**0.2372**	**0.2372**	**0.2372**	**0.2372**
	720	**0.4412**	0.4416	0.4416	**0.4412**	0.4413		720	**0.3126**	**0.3126**	**0.3126**	**0.3126**	**0.3126**
ETTh2	96	**0.2632**	**0.2632**	**0.2632**	**0.2632**	0.2634	Electricity	96	**0.1279**	**0.1279**	**0.1279**	**0.1279**	0.1281
	192	**0.3125**	**0.3125**	**0.3125**	**0.3125**	**0.3125**		192	**0.1465**	**0.1465**	**0.1465**	**0.1465**	**0.1465**
	336	0.3144	0.3152	**0.3087**	0.3144	0.3130		336	**0.1629**	0.1634	0.1639	**0.1629**	**0.1629**
	720	0.3835	0.3829	**0.3823**	0.3834	0.3829		720	0.2070	0.2070	**0.2069**	0.2070	0.2070
ETTm1	96	**0.2849**	0.2851	0.2851	**0.2849**	0.2850	Traffic	96	**0.3559**	0.3562	0.3564	**0.3559**	0.3560
	192	**0.3340**	0.3347	0.3346	**0.3340**	0.3350		192	**0.3808**	**0.3808**	**0.3808**	**0.3808**	**0.3808**
	336	**0.3643**	0.3654	0.3659	**0.3643**	0.3648		336	**0.3868**	**0.3868**	0.3869	**0.3868**	**0.3868**
	720	**0.4193**	0.4196	0.4202	0.4194	0.4213		720	**0.4249**	**0.4249**	**0.4249**	**0.4249**	**0.4249**
ETTm2	96	**0.1665**	0.1669	0.1677	**0.1665**	0.1675	Solar	96	**0.1625**	0.1704	0.1807	**0.1625**	0.1635
	192	**0.2203**	0.2206	0.2206	**0.2203**	0.2214		192	**0.1778**	0.1790	0.1800	**0.1778**	0.1781
	336	**0.2789**	0.2795	0.2798	**0.2789**	0.2797		336	**0.1852**	0.1853	0.1854	**0.1852**	**0.1852**
	720	**0.3644**	**0.3644**	**0.3644**	**0.3644**	0.3724		720	**0.1968**	0.1980	0.2010	**0.1968**	0.1969

Table 11: Matched comparison of the direct base forecast and the corrected DuoTS forecast at all 32 dataset–horizon settings. The two outputs come from the same trained model; the correction is not retrained separately. For each metric, $\Delta_{\%} = 100(\mathrm{DuoTS} - \mathrm{Base})/\mathrm{Base}$. Errors are truncated to three decimal places. Percentage changes are computed from the displayed three-decimal errors and then truncated to three decimal places. The final row counts strict improvements using these same displayed errors; equal values are ties. Negative percentage changes (bold) indicate an improvement.

Dataset	H	Base forecast		DuoTS (Full)		$\Delta_{\%}$		Dataset	H	Base forecast		DuoTS (Full)		$\Delta_{\%}$	
		MSE	MAE	MSE	MAE	MSE	MAE			MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.363	0.389	0.363	0.388	0.000	**-0.257**	Electricity	96	0.130	0.228	0.127	0.222	**-2.307**	**-2.631**
	192	0.406	0.417	0.400	0.413	**-1.477**	**-0.959**		192	0.150	0.247	0.146	0.241	**-2.666**	**-2.429**
	336	0.400	0.429	0.401	0.430	+0.250	+0.233		336	0.163	0.261	0.162	0.258	**-0.613**	**-1.149**
	720	0.441	0.457	0.441	0.457	0.000	0.000		720	0.207	0.296	0.206	0.296	**-0.483**	0.000
ETTh2	96	0.262	0.332	0.263	0.332	+0.381	0.000	Solar	96	0.176	0.231	0.162	0.218	**-7.954**	**-5.627**
	192	0.312	0.367	0.312	0.367	0.000	0.000		192	0.191	0.248	0.177	0.236	**-7.329**	**-4.838**
	336	0.307	0.377	0.314	0.377	+2.280	0.000		336	0.187	0.244	0.185	0.241	**-1.069**	**-1.229**
	720	0.380	0.422	0.383	0.424	+0.789	+0.473		720	0.205	0.260	0.196	0.250	**-4.390**	**-3.846**
ETTm1	96	0.283	0.341	0.284	0.342	+0.353	+0.293	Traffic	96	0.359	0.255	0.355	0.252	**-1.114**	**-1.176**
	192	0.332	0.365	0.334	0.366	+0.602	+0.273		192	0.380	0.259	0.380	0.259	0.000	0.000
	336	0.371	0.389	0.364	0.383	**-1.886**	**-1.542**		336	0.387	0.263	0.386	0.262	**-0.258**	**-0.380**
	720	0.424	0.421	0.419	0.419	**-1.179**	**-0.475**		720	0.425	0.284	0.424	0.284	**-0.235**	0.000
ETTm2	96	0.168	0.257	0.166	0.255	**-1.190**	**-0.778**	Weather	96	0.167	0.224	0.142	0.193	**-14.970**	**-13.839**
	192	0.222	0.296	0.220	0.293	**-0.900**	**-1.013**		192	0.212	0.262	0.188	0.238	**-11.320**	**-9.160**
	336	0.280	0.332	0.278	0.330	**-0.714**	**-0.602**		336	0.262	0.298	0.237	0.276	**-9.541**	**-7.382**
	720	0.380	0.394	0.364	0.385	**-4.210**	**-2.284**		720	0.320	0.338	0.312	0.330	**-2.500**	**-2.366**
Improved settings (out of 32): MSE **22**, MAE **21**															

Table 12: Long-term forecasting results of the DuoTS base forecast, i.e., the direct prediction before the correction, against the baselines of Table 1. Base-forecast values are taken from Table 11; baseline values are identical to Table 1. Averages are truncated to three decimal places. The best MSE and MAE per row are in **bold** and the second best are underlined.

		DuoTS (Base)		SRSNet		TimeKAN		Amplifier		iTransformer		TimeMixer		PatchTST	
Dataset	Horizon	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.363**	**0.389**	0.366	0.394	0.370	0.396	0.373	0.399	0.386	0.405	0.372	0.401	0.377	0.397
	192	0.406	0.417	**0.400**	**0.415**	0.403	0.417	0.414	0.420	0.430	0.435	0.413	0.430	0.409	0.425
	336	**0.400**	**0.429**	0.424	0.430	0.420	0.432	0.442	0.446	0.450	0.452	0.438	0.450	0.431	0.444
	720	0.441	0.457	**0.426**	**0.455**	0.442	0.465	0.455	0.467	0.495	0.487	0.483	0.483	0.457	0.477
Avg		**0.402**	**0.423**	0.404	**0.423**	0.408	0.427	0.421	0.433	0.440	0.444	0.426	0.441	0.418	0.435
ETTh2	96	**0.262**	**0.332**	0.271	0.338	0.280	0.343	0.287	0.349	0.292	0.347	0.270	0.342	0.274	0.337
	192	**0.312**	**0.367**	0.324	0.372	0.329	0.382	0.348	0.393	0.348	0.384	0.349	0.387	0.348	0.384
	336	**0.307**	**0.377**	0.349	0.394	0.370	0.412	0.383	0.423	0.372	0.407	0.367	0.410	0.377	0.416
	720	**0.380**	**0.422**	0.391	0.434	0.420	0.450	0.407	0.444	0.424	0.444	0.401	0.436	0.406	0.441
Avg		**0.315**	**0.374**	0.333	0.384	0.349	0.396	0.356	0.402	0.359	0.395	0.346	0.393	0.351	0.394
ETTm1	96	**0.283**	**0.341**	0.288	**0.341**	0.290	0.348	0.292	0.346	0.287	0.342	0.293	0.345	0.289	0.343
	192	0.332	0.365	**0.326**	**0.364**	0.332	0.368	0.327	0.365	0.331	0.371	0.335	0.372	0.329	0.368
	336	0.371	0.389	0.365	0.386	**0.354**	0.386	0.365	0.386	0.358	**0.384**	0.368	0.386	0.362	0.390
	720	0.424	0.421	0.426	0.420	**0.401**	0.417	0.427	0.419	0.412	**0.416**	0.426	0.417	0.416	0.423
Avg		0.352	0.379	0.351	**0.377**	**0.344**	0.379	0.352	0.379	0.347	0.378	0.355	0.380	0.349	0.381
ETTm2	96	0.168	0.257	**0.164**	**0.254**	**0.164**	**0.254**	**0.164**	**0.254**	0.168	0.262	0.165	0.256	0.165	0.255
	192	0.222	0.296	**0.220**	0.296	0.238	0.300	0.226	0.300	0.224	0.295	0.225	0.298	0.221	**0.293**
	336	0.280	0.332	**0.271**	**0.327**	0.278	0.331	0.276	0.331	0.274	0.330	0.277	0.332	0.276	**0.327**
	720	0.380	0.394	**0.353**	**0.380**	0.359	0.387	0.358	0.388	0.367	0.385	0.360	0.387	0.362	0.381
Avg		0.262	0.319	**0.252**	**0.314**	0.259	0.318	0.256	0.318	0.258	0.318	0.256	0.318	0.256	**0.314**
Weather	96	0.167	0.224	**0.147**	**0.198**	0.151	0.205	**0.147**	0.199	0.150	0.207	0.151	**0.198**	0.150	0.200
	192	0.212	0.262	**0.190**	0.242	0.195	0.244	0.194	0.245	0.200	0.248	0.191	0.242	0.191	**0.239**
	336	0.262	0.298	**0.241**	0.282	0.242	0.287	0.243	0.282	0.252	0.287	0.244	0.280	0.242	**0.279**
	720	0.320	0.338	0.325	0.340	0.317	0.340	**0.310**	**0.329**	0.320	0.336	0.316	0.331	0.312	0.330
Avg		0.240	0.280	0.225	0.265	0.226	0.269	**0.223**	0.263	0.230	0.269	0.225	**0.262**	**0.223**	**0.262**
Electricity	96	**0.130**	0.228	0.131	**0.226**	0.135	0.231	0.132	0.227	0.134	0.230	0.153	0.256	0.143	0.247
	192	0.150	0.247	**0.147**	**0.241**	0.149	0.243	0.149	0.243	0.154	0.250	0.168	0.269	0.158	0.260
	336	**0.163**	0.261	**0.163**	**0.258**	0.165	0.260	0.167	0.261	0.169	0.265	0.189	0.281	0.168	0.267
	720	0.207	0.296	0.201	0.291	0.206	0.297	0.203	0.292	**0.194**	**0.288**	0.228	0.320	0.214	0.307
Avg		0.162	0.258	**0.160**	**0.254**	0.163	0.257	0.162	0.255	0.162	0.258	0.184	0.281	0.170	0.270
Traffic	96	**0.359**	0.255	0.361	**0.254**	0.388	0.269	0.391	0.277	0.363	0.265	0.369	0.256	0.370	0.262
	192	**0.380**	**0.259**	**0.380**	0.263	0.411	0.286	0.405	0.283	0.384	0.273	0.400	0.271	0.386	0.269
	336	**0.387**	**0.263**	0.392	0.270	0.425	0.284	0.416	0.290	0.396	0.277	0.407	0.272	0.396	0.275
	720	**0.425**	**0.284**	0.434	0.293	0.455	0.302	0.436	0.312	0.445	0.308	0.462	0.316	0.435	0.295
Avg		**0.387**	**0.265**	0.391	0.270	0.419	0.285	0.412	0.290	0.397	0.280	0.409	0.278	0.396	0.275
Solar	96	0.176	0.231	**0.167**	**0.222**	0.187	0.255	0.175	0.237	0.174	0.229	0.180	0.233	0.170	0.234
	192	0.191	0.248	**0.182**	**0.237**	0.194	0.265	0.198	0.259	0.205	0.270	0.201	0.259	0.204	0.302
	336	**0.187**	**0.244**	0.188	0.245	0.203	0.264	0.213	0.259	0.216	0.282	0.214	0.272	0.212	0.293
	720	0.205	0.260	**0.195**	**0.251**	0.209	0.269	0.222	0.369	0.211	0.260	0.218	0.278	0.215	0.307
Avg		0.189	0.245	**0.183**	**0.238**	0.198	0.263	0.202	0.281	0.201	0.260	0.203	0.260	0.200	0.284
# Best/32		14	11	**16**	**15**	3	1	3	2	1	3	0	1	0	4

Table 13: Correction decomposition averaged over four horizons. ΔMSE is corrected minus base ($\times 10^{-3}$). The last columns report where the correction has larger squared error.

Dataset	ΔMSE	Steps hurt (%)	Channels hurt (%)
Solar	-9.45	24.5	0.2
Traffic	-1.25	12.9	12.4
Weather	-17.50	0.0	16.7

H COMPUTATIONAL COST

Table 14: Computational cost with $L = 512$ and $H = 96$ on one RTX 4060 Laptop GPU. ETT uses batch size 32, Electricity uses batch size 4, and Weather uses batch size 16. Smaller values are better.

Dataset	Model	Params (M)	FLOPs (G)	FLOPs/C (G/ch.)	Epoch (s)	Latency (ms/sample)	Memory (MB)
ETTh2 ($C = 7$, batch size 32)							
	DLinear	0.10	0.001	0.0001	0.59	0.003	21
	TimeKAN	0.05	0.035	0.0050	8.54	0.410	554
	iTransformer	0.95	0.021	0.0030	1.53	0.021	47
	PatchTST	7.40	1.971	0.2816	10.96	0.348	573
	DuoTS (Ours)	12.97	2.648	0.3783	34.95	0.942	2057
Electricity ($C = 321$, batch size 4)							
	DLinear	0.10	0.063	0.0002	12.10	0.044	37
	TimeKAN	0.06	2.170	0.0068	1395.76	32.348	4180
	iTransformer	5.05	3.918	0.0122	81.12	1.368	274
	PatchTST	7.40	90.365	0.2815	1323.39	24.970	2729
	DuoTS (Ours)	3.04	13.229	0.0412	1151.09	10.051	2783
Weather ($C = 21$, batch size 16)							
	DLinear	0.10	0.004	0.0002	5.38	0.007	23
	TimeKAN	0.06	0.173	0.0082	200.28	2.348	1408
	iTransformer	1.21	0.061	0.0029	19.04	0.044	55
	PatchTST	9.46	17.418	0.8294	486.50	3.438	3600
	DuoTS (Ours)	1.09	0.268	0.0127	51.94	0.188	304

Table 14 uses a common look-back window of 512, horizon 96, and a single RTX 4060 Laptop GPU. On Electricity, DuoTS has 3.04M parameters and uses 13.229G FLOPs, compared with PatchTST's 7.40M parameters and 90.365G FLOPs. On Weather, DuoTS lies between iTransformer and PatchTST in FLOPs, epoch time, and latency. On ETTh2, DuoTS is less efficient than all listed baselines. Its seven channels do not amortize the fixed wide-representation and correction costs.