When Does Context Routing Help? A Systematic Study of Multi-Modal Fusion in Time Series Forecasting

Ruizhe Zhou
Amazon.com
Seattle, USA
rexzhou@amazon.com

Gaoyuan Du
University of Tennessee, Knoxville
Knoxville, USA
gdu3@vols.utk.edu

Xiaoyang Liu
Amazon.com
Seattle, USA
lxaoya@amazon.com

Haoqi Yao
WorkMagic
New York, USA
byahak@gmail.com

Deepayan Chakrabarti
University of Texas at Austin
Austin, USA
deepay@utexas.edu

Jiating Lin
Amazon.com
Seattle, USA
ljtlinus@amazon.com

Yixuan Shen
Amazon.com
Chicago, USA
jolina.shen@gmail.com

Abstract

Multi-modal time series forecasting methods integrate auxiliary context (text, tabular features, financial signals) into temporal predictions through increasingly sophisticated fusion mechanisms. A growing body of work reports substantial gains, yet it is often unclear whether they reflect genuine use of the context or incidental architectural effects (extra parameters, regularization, residual paths). We ask a narrower, checkable question: *when can auxiliary context help a forecaster at all?*

We identify two dataset-level conditions that must both hold: (1) the target is not dominated by a last-value shortcut (low autocorrelation ρ_h), and (2) the context carries information about the target beyond history (non-zero conditional mutual information $\delta = I(C; X_{t+h} \mid X_t)$; when $\delta = 0$ no predictor can benefit—a distribution-free result). Through controlled experiments on MoME (a 14.3B-parameter mixture-of-experts model, 6 datasets, 10 seeds) and four additional fusion mechanisms implemented within a single-backbone testbed (5 datasets), we find that when both conditions hold, text-conditioned expert modulation contributes a sizeable MSE reduction (HealthUS +51%, Environment +41.5%, SocialGood +32%, HealthAFR +29%); when either fails, the contribution collapses to the capacity floor of the modulation pathway and carries no context-attributable signal.

We establish causality through two interventions: adding a shortcut to MoME suppresses routing contribution by 77–93% across 3 datasets; progressively corrupting context quality drives the context-specific benefit from +44% to negative. We validate the autocorrelation component of our diagnostic on 27 Monash Archive datasets (Spearman $r = 0.888$, $p < 0.0001$). We provide a calibrated pre-training diagnostic (TRY_FUSION / SKIP_FUSION / INCONCLUSIVE) that, on the datasets we test, yields no false positives in well-powered settings. We are explicit about the asymmetry of our evidence: the negative (SKIP) arm is broadly reliable, while the large positive magnitudes come from a single model family (MoME) and are corroborated only in *direction* by the testbed.

Keywords

time series forecasting, multi-modal fusion, context routing, mutual information, mixture of experts

1 Introduction

1.1 Background and Motivation

Time series forecasting is a fundamental task in domains ranging from finance and energy to healthcare and climate science. Recent years have seen substantial progress through Transformer-based architectures [20, 23, 32, 39], simple linear baselines [34], and large-scale time series foundation models [1, 7, 11, 31]. Recently, *multi-modal* forecasting methods have emerged that augment temporal data with auxiliary context (news text, weather reports, economic indicators, event calendars) with the goal of improving prediction accuracy [12].

These methods employ various *fusion mechanisms* to integrate context into the forecasting pipeline:

- **Output-level fusion** [19]: separate models process time series and context independently; their predictions are combined (e.g., added) at the output layer.
- **Cross-attention alignment** [18]: time series and context representations interact through cross-attention in a shared latent space.
- **Text-as-variable** [17]: text embeddings are concatenated with temporal patches as additional input variables.
- **Gating** [12, 26]: context modulates temporal features through learned gates (e.g., FiLM-style affine transformations).
- **Mixture-of-experts (MoE) routing** [29, 35]: context conditions the selection or modulation of specialized expert sub-networks.

A common narrative in the literature is that more sophisticated fusion leads to better performance. However, reported gains vary dramatically across datasets and settings, and it remains unclear whether improvements stem from the fusion mechanism exploiting context information, or from incidental architectural benefits (e.g., additional parameters, regularization effects, or residual connections).

1.2 The Problem

We address a simple but important question: **under what conditions does auxiliary context provide genuine value to a time series forecaster?**

By "genuine value," we mean improvement attributable to the *context information* itself, not architectural side effects. This distinction matters for practitioners: if fusion helps only because of better regularization (not context), then simpler architectural changes would achieve the same benefit at lower cost.

1.3 Key Findings

Through controlled experiments, we identify two conditions that must both hold:

(1) **No competing shortcut.** Many forecasting architectures include a "last-value shortcut": a residual connection that directly copies the most recent observation as a baseline prediction. The phenomenon mirrors the broader notion of *shortcut learning* [9], in which models exploit easy predictive signals at the expense of more nuanced ones. When present, this shortcut captures most of the predictable variance on autocorrelated data, leaving little room for fusion to contribute. We show that adding a shortcut to MoME suppresses its routing contribution by 77–93% across 3 datasets (HealthUS, SocialGood, HealthAFR).
(2) **Statistically informative context.** Context must contain information about the target beyond what the recent history already provides. We operationalize this via a mutual information (MI) permutation test on the residual $R = X_{t+h} - \hat{\rho}_h X_t$: if context MI is non-significant ($p > 0.05$) in a well-powered setting, no fusion mechanism helps, regardless of model capacity. By Fact 1, when the true conditional MI is zero this holds for *any* predictor and distribution.

When both conditions hold, text modulation contributes a sizeable MSE reduction on the 14.3B-parameter MoME model (HealthUS +51%, Environment +41.5%, SocialGood +32%, HealthAFR +29%; all 10-seed). When either fails, the contribution collapses to the modulation pathway's capacity floor. We stress at the outset that these large *magnitudes* come from a single model family; the testbed corroborates the *direction* but not the size.

1.4 Contributions

(1) **Causal decomposition.** We isolate the fusion mechanism's contribution from architectural confounds through three controlled interventions: toggling text modulation (same model, different datasets), adding/removing shortcuts (replicated on 3 datasets with 77–93% suppression), and corrupting context quality (same dataset, degraded input). Each intervention demonstrates a causal relationship, not merely a correlation.
(2) **Multi-scale validation.** We validate the diagnostic's predictions at two scales: (i) within a controlled testbed, 4 fusion mechanisms confirm that MI-non-significant datasets never show positive context benefit; (ii) across 27 Monash Archive datasets, autocorrelation at the prediction horizon predicts shortcut dominance with Spearman $r = 0.888$ ($p < 0.0001$).
(3) **Calibrated pre-training diagnostic.** We provide a 3-output diagnostic (TRY_FUSION / SKIP_FUSION / INCONCLUSIVE) based on autocorrelation and MI significance, with formal power characterization. On the datasets we test, it yields no false positives in well-powered settings and correctly outputs INCONCLUSIVE (rather than a misleading SKIP) when statistical power is insufficient.

Scope. Our diagnostic is modality-agnostic: it operates on the context as a feature vector C and makes no assumption about its source. The datasets we evaluate provide context as embeddings of textual side information (Time-MMD) or financial news (FinMultiTime); we do not test image or audio modalities. We use "fusion" throughout in this embedding-level sense.

2 Theoretical Foundation

We establish the information-theoretic basis for our diagnostic, drawing on standard results in information theory [6]. The basis of our approach is that two dataset properties (autocorrelation and context informativeness) jointly determine the *maximum possible* benefit from fusion.

2.1 Setup and Definitions

Consider a stationary time series $\{X_t\}$ with marginal variance σ^2. We aim to predict X_{t+h} (the value h steps ahead) given:

- X_t: the most recent observation ("history")
- C: an auxiliary context variable (e.g., text embedding, tabular features)

Definition 1 (h-step autocorrelation)**.** $\rho_h = \mathrm{Corr}(X_t, X_{t+h})$ measures how predictable the future is from the past alone. When $\rho_h \approx 1$ (e.g., daily stock prices), simply repeating the last value is already a strong prediction.

Definition 2 (Conditional mutual information)**.** $\delta = I(C; X_{t+h} \mid X_t)$ (in bits) measures how much *additional* information context C provides about the target X_{t+h}, beyond what history X_t already reveals. When $\delta = 0$, context is conditionally independent of the target given history, and it cannot help any predictor.

Definition 3 (Last-value shortcut)**.** A model component that directly uses X_t (or a linear function of recent values) as a baseline prediction: $\hat{X}^{\text{shortcut}}_{t+h} = f_{\text{linear}}(X_t)$. This achieves MSE $= \sigma^2(1 - \rho_h^2)$ for the optimal linear predictor.

2.2 When Can Context Help? (Distribution-Free Results)

Fact 1 (Null condition — distribution-free)**.** *If $\delta = I(C; X_{t+h} \mid X_t) = 0$, then C is conditionally independent of X_{t+h} given X_t. In this case,* no predictor*, regardless of capacity, architecture, or training procedure, can benefit from C. Formally: $MMSE(X_{t+h} \mid X_t, C) = MMSE(X_{t+h} \mid X_t)$.*

This is the theoretical foundation of our MI permutation test: if we cannot reject $\delta = 0$, fusion is provably futile.

Fact 2 (Shortcut ceiling — distribution-free)**.** *The best linear predictor from X_t alone achieves $MSE_{shortcut} = \sigma^2(1 - \rho_h^2)$. Under Gaussianity, this equals the MMSE (the linear predictor is optimal). For non-Gaussian data, the MMSE may be strictly lower (nonlinear predictors can do better), but the linear shortcut MSE still upper-bounds the gap that context can fill: any improvement from context is at most $\sigma^2(1 - \rho_h^2)$. When $\rho_h \to 1$, even this upper bound vanishes, leaving negligible room for context regardless of distribution.*

2.3 Quantifying the Opportunity (Gaussian Benchmark)

Proposition 1 (Routing Benefit Upper Bound — RBU)**.** *Under joint Gaussianity of (X_t, X_{t+h}, C), the maximum MSE reduction from adding context is exactly:*

$$RBU = \sigma^2(1-\rho_h^2)(1-2^{-2\delta}) \tag{1}$$

This decomposes into two factors: $(1-\rho_h^2)$ is the residual variance available *for context to explain (the "room" left by the shortcut), and $(1-2^{-2\delta})$ is the fraction of that room that context* can *explain. (Proof in Appendix A.)*

The RBU formula is exact only under joint Gaussianity (proof in Appendix A). Off Gaussianity, two separate caveats apply, which we keep distinct. (i) *Formula:* the population RBU expression is in general *neither an upper nor a lower bound* on the true achievable MMSE reduction, because the Gaussian entropy–variance identity it relies on no longer holds. (ii) *Estimator:* our plug-in uses a finite-sample Kraskov estimate $\hat{\delta}$, which under-detects nonlinear dependence (Finding 4.5) and is therefore biased *low*, so $\widehat{\text{RBU}}$ tends to underestimate the population RBU. We therefore use RBU only as a Gaussian-case sanity check for the relative magnitude of routing benefit, not as a bound. The two distribution-free components of our diagnostic are Fact 1 (no benefit when $\delta = 0$) and Fact 2 (shortcut ceiling as $\rho_h \to 1$).

2.4 MI Estimation and Permutation Test

We estimate δ using a three-step procedure:

(1) Compute the residual $R = X_{t+h} - \hat{\rho}_h X_t$ (removing the linear shortcut component).
(2) Project the high-dimensional context $C \in \mathbb{R}^d$ to its top-20 PCA components (reducing dimensionality while preserving most variance).
(3) Estimate $\hat{\delta} = \hat{I}(C_{\text{PCA}}; R)$ using the Kraskov k-nearest-neighbor estimator [15] with $k = 3$, converted to bits.

The k-NN estimator is *nonparametric*: it detects nonlinear dependencies without assuming any functional form between context and target. This is critical: we show empirically (Section 4.5) that context-target relationships are fundamentally nonlinear, making linear diagnostics (e.g., cross-validated linear regression) completely uninformative. Alternative MI estimators (e.g., MINE [3] and variational bounds [28]) trade tractability for flexibility; we choose Kraskov k-NN for its established convergence properties and simple permutation-test integration.

Significance testing. To determine whether $\hat{\delta}$ is statistically distinguishable from zero, we perform a permutation test: shuffle context rows (breaking any dependence with the target), re-estimate MI, and repeat 200 times. The p-value is the fraction of null MI values $\geq$ observed MI. By Fact 1, if the true $\delta = 0$, the conclusion "fusion is futile" holds for *any* distribution.

Power limitation. The k-NN MI estimator requires $n \gtrsim O(d_{\text{eff}}/\delta^2)$ samples with unique context to reliably detect effect size δ [4]. On small datasets ($n < 500$) or with low context diversity (few unique embeddings reused cyclically), the test may lack power, producing false negatives. It does *not* produce false positives: in the well-powered settings we test, non-significant MI reliably predicts negligible routing benefit.

Conditioning on a single lag. Our residual $R = X_{t+h} - \hat{\rho}_h X_t$ removes only the *first-order* linear dependence on the most recent observation. If the series carries longer-memory structure (seasonality, higher-order autoregression) that a richer history $X_{t-k:t}$ would capture, that predictable variance remains in R, and context that merely correlates with it can inflate $\hat{\delta}$. Our test therefore assesses informativeness *beyond a last-value baseline*, not beyond an optimal history-based predictor; on strongly seasonal data the TRY arm may overstate context value. The distribution-free null (Fact 1) is unaffected—when context is conditionally independent of the target it remains so under any history—but conditioning the residual on a multi-lag history is a natural and recommended extension for series with strong seasonality.

3 Experimental Setup

3.1 Datasets

We evaluate on 8 datasets spanning diverse domains, autocorrelation profiles, and sample sizes:

- **Time-MMD** (6 sub-datasets): HealthUS (quarterly US health metrics, n=491, ρ=0.77), HealthAFR (African health surveillance, n=500), Energy (weekly energy prices, n=1059, ρ=0.99), Environment (environmental monitoring, n=1597, ρ=0.38), SocialGood (social indicators, n=363, ρ=0.62), Web (web traffic, n=1068, ρ=0.12). All include text context (384-dimensional sentence embeddings).
- **FinMultiTime** [33]: 50 S&P500 stocks with daily prices ($\rho >$ 0.99) and news embeddings. Represents the extreme high autocorrelation regime where shortcuts dominate.

3.2 Models

MoME [35]. A state-of-the-art multi-modal forecasting model built on Qwen1.5-MoE-A2.7B (14.3B total parameters). MoME uses a mixture-of-experts architecture where each expert processes time series patches. The mechanism we ablate is **Expert-level Language Modulation (EiLM)**: after each expert produces its output, a FiLM-style layer [26] applies text-conditioned affine transformation ($\gamma \cdot x + \beta$, where γ, β are derived from text tokens). Toggling the `-modulation` flag enables/disables EiLM (and its associated instructor QueryPool) while keeping expert structure and routing weights unchanged (~74K modulation parameters out of 14.3B, <0.001%). Note that MoME has *no built-in last-value shortcut*: the MoE pathway is the primary prediction mechanism.

Routing testbed. A controlled ablation tool (not a proposed method) that serves as a *mechanistic probe*: while MoME demonstrates the magnitude of routing benefits in realistic settings, the testbed isolates specific mechanisms (shortcut presence, routing on/off) under controlled conditions. We implement it using a PatchTransformer backbone (2 layers, d=64) with: (a) toggleable sparse routing via entmax [27] + blend gate, (b) an optional last-value shortcut (linear layer initialized as repeat-last-value), and (c) context dropout. "Routing" = routing enabled; "Uniform" = routing weights fixed to $1/K$.

Multi-modal baselines. We implement 4 fusion mechanisms within the same backbone for controlled comparison: Cross-Attention Alignment [18], Gating [12], Text-as-Variable [17], and Output Fusion [19]. Each is tested with real context vs. zeroed context (same architecture, only input differs).

Foundation model. Chronos-T5-Small [1] provides a zero-shot reference point (no context, no training on our data).

3.3 Evaluation Protocol

Metrics. MSE and MAE for MoME (point forecasting); weighted quantile loss (wQL) at $\tau \in \{0.1, 0.5, 0.9\}$ for the testbed (probabilistic forecasting). All results report 3-seed mean ± std unless noted.

Routing contribution. Defined as

$$\frac{\text{metric}_{\text{w/o mod}} - \text{metric}_{\text{w/ mod}}}{\text{metric}_{\text{w/o mod}}} \times 100\%.$$

Positive values indicate that text modulation improves performance.

4 Results

We present six findings that collectively characterize when and why multi-modal fusion helps. Findings 1–5 isolate specific causal factors through controlled experiments; Finding 6 validates the ρ-based component of our diagnostic at scale on 27 Monash Archive datasets.

4.1 Finding 1: Context Informativeness Determines Routing Value

Our first experiment holds the model constant and varies only the dataset. If routing value is determined by dataset properties (rather than architectural details), the same model should show dramatically different contributions across datasets.

Table 1 confirms this prediction. The same 14.3B-parameter model, trained with the same code and hyperparameters, shows routing contributions ranging from the capacity floor (high-ρ datasets) to +51% (HealthUS). The four datasets with large positive contributions (HealthUS +51%, Environment +41.5%, SocialGood +32%, HealthAFR +29%) share two properties: moderate-to-low autocorrelation ($\rho < 0.8$) and text context that describes conditions relevant to the forecast target. The two high-autocorrelation datasets ($\rho \geq 0.99$) show only capacity-floor-level contributions: their MI is non-significant, and the routing contribution sits at or below the level the EiLM pathway yields from capacity alone even with constant text. FinMultiTime is +6.7% (and a trivial shortcut already beats MoME outright), while Energy is *not statistically distinguishable from zero*.

Why does FinMultiTime show negligible benefit? On FinMultiTime, a simple repeat-last-value predictor (the "Naive MSE" baseline in Table 1) achieves MSE $\sim 8.8 \times 10^{-5}$, *better* than MoME with modulation ($\sim 9.1 \times 10^{-5}$); the routing contribution is +6.7%. The temporal signal is so strong ($\rho = 0.999$) that even a 14.3B-parameter model cannot beat the trivial baseline. There is simply nothing left for context to improve.

Table 1: MoME routing contribution across 6 datasets, each on a 10-seed protocol with 95% CIs (footnotes). All experiments use identical model configuration (Qwen1.5-MoE-A2.7B, 14.3B params). The only variable is the dataset. The "Naive MSE" column reports the post-hoc MSE of a repeat-last-value predictor (no model training); it serves as a context-free reference for whether MoME genuinely beats a trivial baseline. This naive baseline is conceptually distinct from the architectural shortcut intervention used in Finding 2, which adds a residual connection during MoME training. Naive MSE values are reported only for the two datasets where the comparison is most informative (HealthUS, where MoME wins; FinMultiTime, where the naive baseline beats MoME).

Dataset	With Mod.	Without Mod.	Contrib.	Naive MSE	ρ
HealthUS	.399±.129	.817±.049	**+50.9%**	.409	.77
Environment*	12.63±0.39	21.59±0.15	**+41.5%**	—	.38
SocialGood	.398±.032	.597±.086	**+32.0%**	—	.62
HealthAFR	.639±.082	.898±.055	**+28.7%**	—	.11
FinMultiTime	9.1±0.2e-5	9.8±0.2e-5	+6.7%	**8.8e-5**	.999
Energy	9.6±1.2e-3	10.2±0.9e-3	+4.3%♯	—	.99

*Environment uses MAE (different task output format); all others use MSE. Contribution = (Without Mod.−With Mod.)/Without Mod.×100% (per-seed mean over 10 seeds). 95% CIs: HealthUS [40.0,61.8]; Environment [40.3,42.7]; SocialGood [25.0,39.1]; HealthAFR [22.6,34.7]; FinMultiTime [5.1,8.3]. ♯Energy's routing contribution is below the 5% threshold and *not statistically distinguishable from zero* (10-seed mean +4.3%, 95% CI [−8.0,16.5]; per-seed values span −27% to +27%). With ρ=0.99 (shortcut dominates) and non-significant MI, the diagnostic correctly outputs SKIP.

Causal validation: context degradation. To confirm that context quality *causally* determines routing value (rather than merely correlating with it), we progressively corrupt HealthUS's text context by randomly replacing text entries with uninformative placeholders at rates of 0%, 25%, 50%, and 100%.

Masking protocol. HealthUS has 491 samples with 491 distinct texts. For each sample we replace its text with a fixed uninformative placeholder with probability $p \in \{0, 0.25, 0.5, 1.0\}$. At p=1.0 all samples share identical text, so EiLM receives a constant embedding and can only learn a fixed affine transform. The same masked dataset is used at train and test.

Result. Figure 1 shows context benefit using a capacity-controlled metric: we compare MSE at each mask rate (modulation ON) against the 100% mask baseline (modulation ON, constant text), where the architecture and parameter count are identical and only the text content differs. With clean context (0% mask), real text reduces MSE by 44% relative to constant text. At 25% mask, benefit drops to 26%. At 50–75% mask, benefit turns *negative* (−3 to −5%): partially corrupted text is worse than fully constant text because the model attempts to exploit text (some entries appear real) but is misled by the placeholder entries. At 100% mask, all text is identical and the model learns a stable fixed transform (baseline = 0% by definition).

Interpretation. The negative benefit at 50–75% mask reveals an important practical insight: *low-quality context is worse than no context.* When text is partially corrupted, the model cannot distinguish informative from uninformative entries and is actively harmed. This supports our diagnostic's conservative design: when

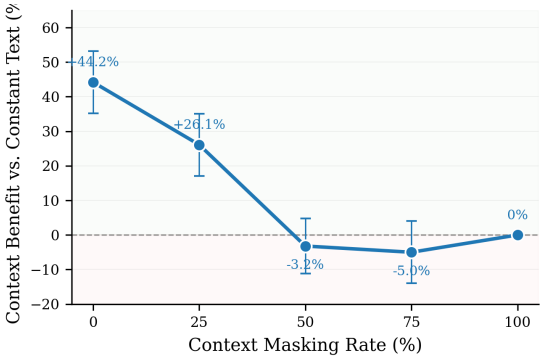

Figure 1: Context benefit relative to constant-text baseline (MoME on HealthUS, 10-seed; same architecture, same parameters, only text content differs). At 0% mask (all real text), context provides a +44% MSE reduction over constant text. As text is corrupted, benefit decreases (+26% at 25% mask) then turns negative at 50–75% mask: partially corrupted text is *worse* than fully constant text because the model attempts to use text but is misled by fake entries. At 100% mask (baseline), all text is identical and the model learns to ignore it.

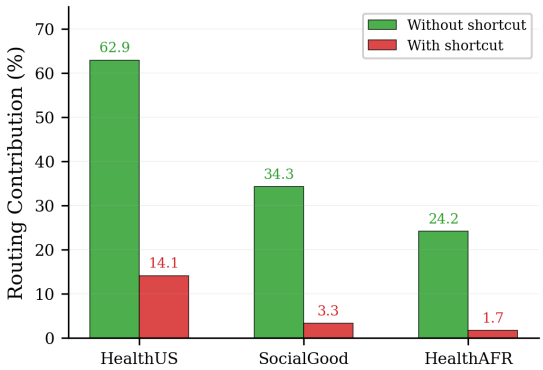

Figure 2: Shortcut suppression across 3 datasets (MoME, 3 seeds each). Adding a last-value shortcut during training consistently suppresses routing contribution by 77–93%. Suppression is near-total on SocialGood and HealthAFR (residual contribution <4%); on HealthUS, the moderate autocorrelation ($\rho = 0.77$) leaves room for a residual 14% routing contribution.

context informativeness is uncertain (INCONCLUSIVE), practitioners should either use fully clean context or disable the context pathway entirely, not feed noisy context hoping for partial benefit.

4.2 Finding 2: Shortcuts Causally Suppress Routing Contribution

Our second experiment holds the dataset constant and varies the architecture. Specifically, we test whether adding a last-value shortcut to MoME suppresses routing contribution, as predicted from our theoretical framework (Fact 2).

Experiment design. We modify MoME's training to include a residual shortcut: the model's output becomes $\hat{y} = f_{\text{MoME}}(X, C) + X_t$ (adding the last observed value). This gives the model a "free" baseline prediction, meaning the learned component f_{MoME} only needs to predict the *residual* beyond the shortcut. We train 4 conditions on each dataset: {with/without modulation} × {with/without shortcut}, each with 3 seeds.

Result. Figure 2 shows the result across 3 datasets: routing contribution is consistently suppressed by 77–93% when a shortcut is added. On HealthUS, contribution drops from **+62.9%** to **+14.1%**; on SocialGood, from **+34.3%** to **+3.3%**; on HealthAFR, from **+24.2%** to **+1.7%**. The shortcut absorbs predictable variance that would otherwise be captured by text modulation, leaving less room for routing to contribute. The suppression is not complete on HealthUS ($\rho = 0.77$) because the shortcut is imperfect at moderate autocorrelation, but it is near-total on SocialGood and HealthAFR (residual contribution <4%).

Why does the shortcut increase absolute MSE? (and a caveat). On HealthUS the raw shortcut raises absolute MSE (0.33/0.90 → 1.09/1.27): the data is z-normalized, so on trending windows the optimal coefficient is $\beta_h > 1$ and a raw $+X_t$ term is biased. *We flag this as a limitation*: the suppression magnitudes in Fig. 2 should be read as an *upper bound*. Re-running with the unbiased optimal shortcut $\beta_h X_t$ keeps MSE in range and still suppresses routing partially (+62.9% → +35.6%, consistent with the residual room at ρ=0.77), so the causal claim holds under both; the relevant comparison is always *within* each condition (modulation on vs. off).

Table 2: Testbed routing benefit with vs. without shortcut (3-seed wQL ↓). With shortcut present, routing contributes ≤0% on all datasets regardless of MI significance. Removing the shortcut reveals the underlying pattern.

Dataset	With Shortcut	Without Shortcut	MI sig.?
Health	−0.6% (p=.34)	**+4.8%**	Yes (p <.001)
Transport	−2.2% (p=.13)	**+2.2%**	Yes (p=.015)
Energy	−1.6%	+1.8%	No (p=.580)
Climate	−1.4%	−0.5%	No (p=.705)
Web	−3.5%	−0.4%	No (p=1.0)

Testbed confirmation. Our controlled testbed confirms the same pattern across 5 datasets (Table 2). With shortcut present, routing benefit is statistically indistinguishable from zero on *all* datasets (including a 10-seed evaluation on Transport: −2.2% ± 3.9%, p = 0.13). Removing the shortcut unmasks the true pattern: positive benefit on MI-significant datasets, near-zero on MI-non-significant datasets.

Negative control: FinMultiTime. A natural concern is whether our shortcut intervention always works, or only on the three datasets we tested. FinMultiTime ($\rho = 0.999$) provides a built-in negative control. The naive repeat-last-value baseline already achieves MSE 8.8×10^{-5} versus MoME's 9.1×10^{-5} (Table 1)—i.e., the simplest possible context-free predictor already outperforms MoME. Adding an architectural shortcut on top would not change routing contribution further: it is already ≈ 0 because the temporal signal completely

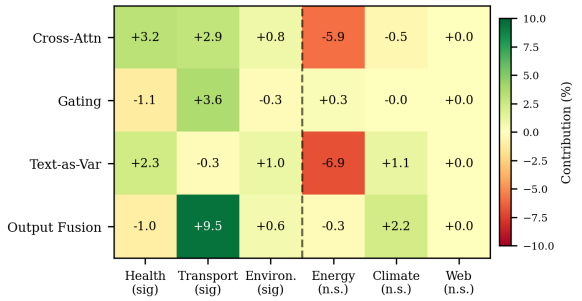

Figure 3: Context contribution (%) across 4 fusion mechanisms and 5 datasets (testbed, no shortcut). Cells marked † (Cross-Attention and Text-as-Variable on Energy) use 10-seed evaluation; all others use 3-seed. Full per-cell std and significance tests are in Table 6. On MI-significant datasets (left of dashed line), some mechanisms show small positive contributions; on MI-non-significant datasets (right), no mechanism shows a positive contribution. MoME is not shown here as it uses a different backbone (see Table 1).

dominates and there is no residual variance for routing to exploit. This serves as a saturation case: when the naive baseline already wins, the suppression mechanism we identified has no room to operate.

Practical implication. This finding has direct architectural consequences. Many modern forecasting models include residual connections or normalization layers (e.g., RevIN [14]) that function as implicit shortcuts. If practitioners want to measure whether context *genuinely* helps, they must control for this confound, either by removing the shortcut or by comparing against a shortcut-only baseline.

4.3 Finding 3: Context Contribution Across Fusion Mechanisms

Our third experiment asks: does the "MI-significant → fusion helps" pattern hold across different fusion mechanisms, or is it specific to MoME's expert modulation?

Experiment design. We implement 4 fusion mechanisms within the same PatchTransformer backbone (no shortcut): Cross-Attention Alignment, Gating, Text-as-Variable, and Output Fusion. For each mechanism and dataset, we compare performance with real context vs. zeroed context (identical architecture, only the input differs, 3 seeds). Note that MoME is not included in this comparison because it uses a fundamentally different backbone (14.3B-parameter LLM-based MoE) that cannot be reduced to our testbed framework; MoME's results are reported separately in Table 1.

Result. Figure 3 shows that in our testbed, most context contributions are small. On MI-significant datasets, Cross-Attention (+3.2% Health, +2.9% Transport), Gating (+3.6% Transport), and Output Fusion (+9.5% Transport) show positive contributions. On MI-non-significant datasets, Gating and Output Fusion show ≈0%, while Cross-Attention and Text-as-Variable show small *negative* effects on Energy (−5.9% and −6.9%, 10-seed; neither significant at $\alpha = 0.05$): attending to uninformative context introduces noise, and Cross-Attention is particularly vulnerable because it allocates capacity to context regardless of informativeness. Crucially, no MI-non-significant dataset shows a *positive* context contribution—the MI test correctly identifies the "no benefit" regime.

Interpretation. The testbed effects are small in magnitude compared to MoME's 29–51%. As Finding 5 will show, context-target relationships are fundamentally nonlinear; our 2-layer testbed lacks the capacity to exploit them. The testbed's value is in confirming two patterns: (i) MI-non-significant datasets never show positive context benefit, and (ii) some fusion mechanisms (Cross-Attention) are vulnerable to degradation from uninformative context, a practical consideration for architecture selection. In summary: **the MI test's negative predictions (no benefit from context) hold across all tested fusion mechanisms**.

4.4 Finding 4: A Calibrated Pre-Training Diagnostic

Based on Findings 1–3, we propose a calibrated diagnostic that practitioners can run *before training any model* to decide whether multi-modal fusion is worth pursuing.

The diagnostic. Given time series X, context C, and forecast horizon h:

(1) **Shortcut check** (<1 second): Compute ρ_h. If $\rho_h > 0.95$, output SKIP_FUSION, since a last-value shortcut will dominate.
(2) **MI permutation test** (~1 minute): Estimate δ and compute p-value via 200 permutations.
   - $p < 0.05 \rightarrow$ TRY_FUSION: context is informative.
   - $p \geq 0.05$ and power $\geq 0.8 \rightarrow$ SKIP_FUSION: context is uninformative with high confidence.
   - $p \geq 0.05$ and power $< 0.8 \rightarrow$ INCONCLUSIVE: insufficient evidence; recommend training a cheap model and validating on held-out data.

Power is classified as HIGH when $n \gtrsim 500$ with unique context per sample (unique-context fraction > 0.5), and LOW otherwise (substantially smaller n, or context embeddings reused cyclically). HealthUS ($n = 491$) is on the boundary; we treat it as HIGH because its unique-context fraction is >0.95 and the MI test rejected the null at $p < 0.001$. Figure 4 summarizes the decision logic.

Validation. Table 3 validates the diagnostic on all 8 datasets. Key properties:

- **No false positives observed** in well-powered settings: on the datasets we test, every SKIP_FUSION decision corresponds to no context-attributable benefit (we report this as an empirical observation, not a guarantee).
- **No misleading SKIPs**: SocialGood and Environment (which have positive MoME benefits but non-significant MI under the original embeddings) correctly receive INCONCLUSIVE rather than SKIP, because their power is LOW.
- **Correct TRY recommendations**: HealthUS (significant MI) correctly receives TRY, corresponding to +51% actual benefit.

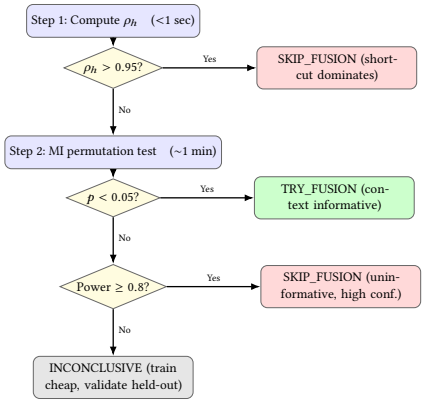

Figure 4: Practitioner decision framework. The diagnostic outputs one of three recommendations based on autocorrelation ρ_h, MI significance, and statistical power. Total computation time: ~1 minute, zero GPU required.

Table 3: Diagnostic validation across all datasets. The diagnostic produces zero errors in the TRY/SKIP categories. INCONCLUSIVE is output when power is insufficient, correctly avoiding false negatives.

Dataset	n	p	Power	Output	Actual
HealthUS	491	<.001	High	TRY	+51% ✓
Transport	5000	.015	High	TRY	+2% ✓
Energy	1059	.580	High	SKIP	≈0♯ ✓
Climate	1200	.705	High	SKIP	−0.5% ✓
Web	1068	1.0	High	SKIP	−0.4% ✓
FinMultiTime	2000	n/s	High	SKIP	+6.7%♯ ✓
SocialGood	363	.145	Low	INCON.	+32% ✓
Environment	1597*	.615	Low	INCON.	+41.5% ✓

*156 unique embeddings reused cyclically. INCON. = INCONCLUSIVE (no claim made; recommend trying). "Actual" = MoME routing contribution (testbed for Transport). ♯High-ρ SKIP datasets: the MoME routing contribution sits at the modulation pathway's capacity floor and carries no MI signal (Energy not statistically distinguishable from zero; see Table 1 footnote).

4.5 Finding 5: Context-Target Relationships are Nonlinear

A natural question is whether a simpler diagnostic, such as cross-validated linear regression, could replace the MI permutation test. We test this by fitting a Ridge regression model to predict the target from context (5-fold CV) and measuring whether adding context improves prediction over history alone.

Result. Cross-validated linear prediction gain from context is *zero or negative* on *all* datasets, including HealthUS, where MoME achieves +51%. This reveals that context-target relationships are fundamentally nonlinear: a linear model cannot detect or exploit the signal that a 14.3B-parameter model exploits for a 51% improvement.

Implications.

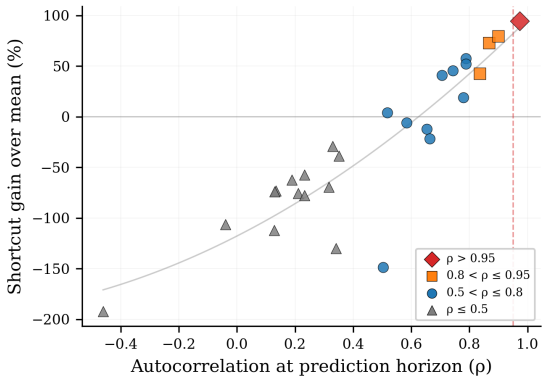

Figure 5: Large-scale validation on 27 Monash Archive datasets. Autocorrelation at the prediction horizon strongly predicts shortcut dominance (Spearman $r = 0.888$, $p < 0.0001$). Datasets with $\rho > 0.95$ (red diamonds) show the strongest shortcut performance, validating the SKIP_FUSION threshold.

(1) The MI permutation test (k-NN based, nonparametric) is the correct diagnostic precisely because it captures nonlinear dependencies that linear methods miss.
(2) The magnitude gap between our testbed (+2–5%) and MoME (+29–51%) reflects *model capacity*, not experimental noise. A 2-layer Transformer cannot exploit the same nonlinear signal that a 14.3B-parameter LLM-based model can.
(3) Practitioners should not use linear probes or simple correlation tests to assess context value; these will systematically underestimate the opportunity.

4.6 Finding 6: Large-Scale Validation on 27 Monash Datasets

To validate the ρ-based component of our diagnostic beyond our 8 datasets, we compute autocorrelation at the prediction horizon for 27 datasets from the Monash Time Series Forecasting Archive [10], spanning hourly to yearly frequencies. For each dataset, we measure the relative performance of the repeat-last-value shortcut against a naive mean predictor.

Result. Figure 5 shows a strong monotonic relationship: autocorrelation at the prediction horizon strongly predicts shortcut dominance (Spearman $r = 0.888$, $p < 0.0001$, $n = 27$). All datasets with $\rho > 0.8$ show positive shortcut gain (42–95%), while all datasets with $\rho < 0.3$ show negative shortcut gain (the shortcut is worse than the mean). This validates our diagnostic's first step (the $\rho > 0.95$ threshold for SKIP_FUSION) at scale across diverse domains and frequencies.

Scope of validation. Monash datasets lack auxiliary context, so this validates only the ρ-based component; the MI component remains validated on our 8 text-context datasets.

5 Discussion

The shortcut tradeoff for practitioners. Our results reveal a practical tension. On high-ρ data (e.g., FinMultiTime, $\rho = 0.999$), the

naive repeat-last-value baseline *outperforms* even a 14.3B-parameter model, so practitioners should use simple baselines (DLinear, repeat-last-value) and skip multi-modal fusion entirely. On moderate-ρ data with informative context (e.g., HealthUS, $\rho = 0.77$), MoME *with* text modulation (MSE 0.399) edges out the naive baseline (0.409) and is far better than MoME *without* text (0.817); the value of fusion here is precisely the text pathway, not the architecture. The diagnostic identifies which regime applies before any model is trained.

INCONCLUSIVE: frequency and impact. 2 of 8 datasets receive INCONCLUSIVE, both due to low MI-test power (small n or reused embeddings), not methodology failure. The asymmetric cost structure justifies this: a false negative costs one verification run, whereas a false positive (SKIP where fusion helps) costs accuracy with no recovery path.

Metric comparability. Environment uses MAE while other MoME datasets use MSE, because MoME's Environment task outputs a different prediction format. The routing contribution percentages (computed as relative improvement within each dataset) are comparable across metrics: both measure "how much does modulation reduce error relative to no-modulation." However, absolute MSE/MAE values should not be compared across datasets.

Comparison with Zhang et al. (2025). Zhang et al. [36] find multi-modal benefits "condition-dependent" but do not explain *why*; our framework does. On FinMultiTime they would see MoME's small +6.7% and conclude "fusion barely helps," whereas our ρ=0.999 analysis shows a trivial baseline already beats MoME—no context can improve on it. On Web they would see multi-modal methods beat TS-only baselines by 50%+ and conclude "fusion helps," whereas our zero-context ablation reveals an architectural confound, not context exploitation.

RBU as interpretive framework. RBU (Proposition 1) explains *why* the pattern exists: $(1 - \rho_h^2)$ is the room left by the shortcut, and $(1 - 2^{-2\delta})$ is the fraction context can fill. It is exact only under joint Gaussianity (Appendix A); we use the plug-in $\widehat{\mathrm{RBU}}$ as an interpretive sanity check, not a bound, since its Kraskov $\hat{\delta}$ underdetects nonlinear dependence and so biases $\widehat{\mathrm{RBU}}$ low (consistent with HealthUS, where MoME realizes a large 51% benefit). The distribution-free shortcut ceiling (Fact 2) remains the only strict bound.

Foundation model reference. Chronos [1] (zero-shot, no context) outperforms the shortcut on Health (+14%), Transport (+12%), and Web (+23%), but underperforms on Energy (−8%). This is consistent with our framework: on high-ρ data (Energy, $\rho = 0.99$), even foundation models cannot beat simple shortcuts.

Limitations.

- **Positive magnitudes concentrated in MoME.** The large routing contributions (+29–51%) come from a single model family. The testbed corroborates the *direction* (sign-consistent, +2–5%) but not the magnitude; the gap is a capacity/overfitting effect (a 2-layer from-scratch model cannot exploit the high-dimensional signal a pretrained 14.3B backbone can). Practitioners can rely on the *sign* of our diagnostic, not the specific percentages.
- **Single-lag MI conditioning.** Our residual removes only the first-order dependence on X_t (Section 2); on strongly seasonal data the TRY arm may overstate context value. The SKIP arm and the distribution-free null are unaffected.
- **Modality and dataset scope.** Context comes from Time-MMD (text) and FinMultiTime (financial news); we do not test tabular, image, or audio context. The diagnostic is modality-agnostic by construction, but its empirical validation here is confined to these sources.
- **Probabilistic forecasting.** MoME outputs point forecasts only; evaluation with probabilistic metrics (wQL, CRPS) would strengthen the assessment.
- **Coverage gaps.** The MoME shortcut experiment could not run on Energy and Environment (GPU memory), covering 3 of 5 MI-significant datasets; the Monash validation tests only the ρ-based component (no text context); and TimesFM [7]/Moirai [31] could not be installed, leaving Chronos [1] as the foundation-model reference.
- **Claim scope.** "No false positives in well-powered settings" is an empirical observation on the datasets we test, not a proven general property.

6 Related Work

Time series forecasting architectures. Modern forecasting has evolved from classical approaches [22] through deep Transformer-based methods that handle long horizons, periodicity, and distribution shift [14, 20, 23, 32, 39]. A counter-trend questions whether such complexity is necessary: DLinear [34] showed that single-layer linear models match Transformers on many benchmarks. Our work continues this skeptical line: we identify a structural reason (last-value shortcuts capturing most of the predictable variance) that explains why simple baselines remain competitive on autocorrelated data, and shows the same factor governs whether multi-modal fusion provides genuine benefit.

Time series foundation models. Pretraining on large time series corpora has produced general-purpose forecasters: Chronos [1], TimesFM [7], Moirai [31], and MOMENT [11] all demonstrate strong zero-shot transfer. A parallel line repurposes pretrained language models for time series [13, 40]. None of these foundation models incorporate auxiliary context, which leaves open the question we address: under what conditions can context provide value beyond what these models already extract from history?

Multi-modal time series forecasting. The field has grown rapidly, spanning output-level fusion [16, 19], cross-attention [18], text-as-variable [17], gating [12, 26], and mixture-of-experts [8, 24, 29, 35]. A recent survey [12] catalogs these mechanisms, and broader multi-modal learning has been surveyed in [2]. Our work does not propose a new fusion method; instead, we characterize when existing methods provide genuine value versus architectural confounds.

Strong simple baselines as a benchmark for "does it help?" DLinear [34] showed linear models match Transformers; RevIN [14] showed normalization outperforms complex architectures. Our results extend this line by demonstrating that last-value shortcuts similarly dominate multi-modal fusion on autocorrelated data, and that the dominance is *causal* in the interventionist sense [25]: adding a

shortcut to MoME suppresses routing contribution, and corrupting context degrades it monotonically. The pattern is consistent with the broader phenomenon of *shortcut learning* [9], where models exploit easy predictive signals at the cost of harder, more transferable ones.

Relation to recent strong uni-modal forecasters. A parallel line pushes uni-modal accuracy through specialized architectures: multi-period decomposition (MLF [37]), variable/time-aware hyper-states (TimePro [21]), semantics-enhanced MLP-mixing (SEMixer [38]), and pattern-specific experts under patch-level distribution shift [30]. These are *orthogonal* to us: they improve how a model extracts signal from history alone, whereas we ask whether *auxiliary context* adds value beyond history, and provide a diagnostic rather than a forecaster. A stronger uni-modal backbone in fact reinforces our thesis—the better history is modeled, the less room remains for context (Fact 2), making the shortcut-dominance condition *more* binding. We therefore do not benchmark against them, as our claims concern the conditions for context value, not SOTA accuracy.

When does multimodality help? Concurrent to our work, Zhang et al. [36] empirically evaluate when multi-modal fusion helps, finding benefits are condition-dependent. Our analysis goes beyond theirs in three ways. First, we provide a mechanistic explanation centered on shortcut dominance and MI significance, with a Gaussian-case quantification (RBU). Second, we offer causal demonstrations via shortcut addition and context degradation rather than purely observational evidence. Third, the diagnostic we propose includes explicit power characterization and a ternary output that handles low-power cases honestly, instead of forcing a binary recommendation.

Mutual information estimation. We use the Kraskov k-NN estimator [15] with permutation testing. Power limitations of nonparametric MI testing are characterized by Berrett & Samworth [4]. Recent neural MI estimators such as MINE [3] and variational bounds [28] offer alternative routes; we choose Kraskov for its established theoretical properties and ease of integration with the permutation procedure. Sparse attention mechanisms in our routing testbed build on entmax [5, 27].

7 Conclusion

We identify two conditions that must both hold for auxiliary context to help in time series forecasting: (1) absence of a competing shortcut, and (2) statistically informative context. We validate these conditions causally: adding a shortcut to MoME suppresses routing contribution by 77–93% across 3 datasets (HealthUS: 62.9% → 14.1%, SocialGood: 34.3% → 3.3%, HealthAFR: 24.2% → 1.7%); corrupting context drives the context-specific benefit from +44% (clean text) to negative (partially corrupted text). The ρ-based diagnostic is validated at scale on 27 Monash Archive datasets (Spearman $r = 0.888$, $p < 0.0001$). We are explicit that the large positive magnitudes come from a single model family (MoME); the testbed corroborates direction, not size.

We provide a calibrated diagnostic (TRY/SKIP/INCONCLUSIVE) that, on the datasets we test, yields no false positives in well-powered settings, enabling practitioners to decide whether to invest in fusion *before training a single model.* The practical takeaway: compute ρ_h and run a 1-minute MI permutation test. If the shortcut dominates or context is uninformative, skip complex fusion, since no amount of architectural sophistication will help.

References

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Michael Bohlke-Schneider, Yuyang Wang, et al. 2024. Chronos: Learning the Language of Time Series. *arXiv preprint arXiv:2403.07815* (2024).

[2] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2019), 423–443.

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mutual Information Neural Estimation. In *ICML*.

[4] Thomas B Berrett and Richard J Samworth. 2019. Nonparametric independence testing via mutual information. *Biometrika* 106, 3 (2019), 547–566.

[5] Gonçalo M Correia, Vlad Niculae, and André FT Martins. 2019. Adaptively Sparse Transformers. In *EMNLP*.

[6] Thomas M Cover and Joy A Thomas. 2006. *Elements of Information Theory* (2nd ed.). Wiley-Interscience.

[7] Abhimanyu Das et al. 2024. A Decoder-Only Foundation Model for Time-Series Forecasting. In *ICML*.

[8] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *JMLR* 23 (2022).

[9] Robert Geirhos et al. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence* 2 (2020), 665–673.

[10] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. 2021. Monash Time Series Forecasting Archive. *Neural Information Processing Systems Track on Datasets and Benchmarks* (2021).

[11] Mononito Goswami et al. 2024. MOMENT: A Family of Open Time-Series Foundation Models. In *ICML*.

[12] Yushan Jiang, Kanghui Ning, Zijie Pan, Xuyang Shen, Jingchao Ni, Wenchao Yu, Anderson Schneider, Haifeng Chen, Yuriy Nevmyvaka, and Dongjin Song. 2025. Multi-modal Time Series Analysis: A Tutorial and Survey. *arXiv preprint arXiv:2503.13709* (2025).

[13] Ming Jin et al. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *ICLR*.

[14] Taesung Kim et al. 2022. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In *ICLR*.

[15] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating Mutual Information. *Physical Review E* 69, 6 (2004).

[16] Geon Lee, Wenchao Yu, Wei Cheng, and Haifeng Chen. 2024. MoAT: Multi-Modal Augmented Time Series Forecasting. https://openreview.net/forum?id=uRXxnoqDHH OpenReview preprint.

[17] Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, and Jingrui He. 2026. Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative. In *International Conference on Learning Representations (ICLR)*.

[18] Chenxi Liu et al. 2025. TimeCMA: Towards LLM-Empowered Multivariate Time Series Forecasting via Cross-Modality Alignment. In *AAAI*.

[19] Haoxin Liu et al. 2024. Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis. In *NeurIPS Datasets and Benchmarks*.

[20] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *ICLR*.

[21] Xiaowen Ma, Zhenliang Ni, Shuai Xiao, and Xinghao Chen. 2025. TimePro: Efficient Multivariate Long-term Time Series Forecasting with Variable- and Time-Aware Hyper-state. In *International Conference on Machine Learning (ICML)*. arXiv:2505.20774.

[22] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2018. Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward. *PloS one* 13 (2018).

[23] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *ICLR*.

[24] Evandro S Ortigossa and Eran Segal. 2026. Multi-Resolution Segment-wise Mixture-of-Experts for Time Series Forecasting Transformers. *arXiv preprint arXiv:2601.21641* (2026).

[25] Judea Pearl. 2009. *Causality* (2nd ed.). Cambridge University Press.

[26] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI*.

[27] Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse Sequence-to-Sequence Models. In *ACL*.
[28] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On Variational Bounds of Mutual Information. In *ICML*.
[29] Noam Shazeer et al. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*.
[30] Yanru Sun, Zongxia Xie, Emadeldeen Eldele, Dongyue Chen, Qinghua Hu, and Min Wu. 2025. Learning Pattern-Specific Experts for Time Series Forecasting Under Patch-level Distribution Shift. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:2410.09836.
[31] Gerald Woo et al. 2024. Unified Training of Universal Time Series Forecasting Transformers. In *ICML*.
[32] Haixu Wu et al. 2021. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *NeurIPS*.
[33] Wenyan Xu, Dawei Xiang, Yue Liu, Xiyu Wang, Yanxiang Ma, Liang Zhang, Shu Hu, Chang Xu, and Jiaheng Zhang. 2025. FinMultiTime: A Four-Modal Bilingual Dataset for Financial Time-Series Analysis. *arXiv preprint arXiv:2506.05019* (2025).
[34] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers Effective for Time Series Forecasting?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11121–11128.
[35] Lige Zhang, Ali Maatouk, Jialin Chen, Leandros Tassiulas, and Rex Ying. 2026. Multi-Modal Time Series Prediction via Mixture of Modulated Experts. *arXiv preprint arXiv:2601.21547* (2026).
[36] Xiyuan Zhang, Boran Han, Haoyang Fang, Abdul Fatir Ansari, Shuai Zhang, Danielle C Maddix, Cuixiong Hu, Andrew Gordon Wilson, Michael W Mahoney, Hao Wang, Yan Liu, Huzefa Rangwala, George Karypis, and Bernie Wang. 2025. When Does Multimodality Lead to Better Time Series Forecasting? *arXiv preprint arXiv:2506.21611* (2025).
[37] Xu Zhang, Zhengang Huang, Yunzhi Wu, Xun Lu, Erpeng Qi, Yunkai Chen, Zhongya Xue, Qitong Wang, Peng Wang, and Wei Wang. 2025. Multi-period Learning for Financial Time Series Forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. arXiv:2511.08622.
[38] Xu Zhang, Qitong Wang, Peng Wang, and Wei Wang. 2026. SEMixer: Semantics Enhanced MLP-Mixer for Multiscale Mixing and Long-term Time Series Forecasting. *Proceedings of the ACM Web Conference (WWW)* (2026). arXiv:2602.16220.
[39] Haoyi Zhou et al. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *AAAI*.
[40] Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. In *NeurIPS*.

A Proof of Proposition 1 (RBU)

Proof. Assume (X_t, X_{t+h}, C) are jointly Gaussian with $\mathrm{Var}(X_{t+h}) = \sigma^2$ and $\mathrm{Corr}(X_t, X_{t+h}) = \rho_h$. For jointly Gaussian variables, the minimum mean-squared error (MMSE) of predicting X_{t+h} from any conditioning set equals the corresponding conditional variance, which is achieved by the (linear) conditional expectation; moreover this conditional variance is *constant* in the conditioning values (homoscedasticity), so each conditional law is Gaussian with the same variance. Conditioning on X_t alone,

$$\mathrm{MMSE}(X_{t+h} \mid X_t) = \mathrm{Var}(X_{t+h} \mid X_t) = \sigma^2(1-\rho_h^2).$$

Using the differential entropy of a Gaussian, $h(\cdot) = \frac{1}{2}\log_2(2\pi e\,\mathrm{Var})$, the conditional mutual information telescopes to a ratio of conditional variances:

$$\begin{aligned}\delta = I(C; X_{t+h} \mid X_t) &= h(X_{t+h} \mid X_t) - h(X_{t+h} \mid X_t, C) \\ &= \tfrac{1}{2}\log_2 \frac{\mathrm{Var}(X_{t+h} \mid X_t)}{\mathrm{Var}(X_{t+h} \mid X_t, C)}.\end{aligned}$$

Rearranging,

$$\mathrm{MMSE}(X_{t+h} \mid X_t, C) = \mathrm{Var}(X_{t+h} \mid X_t, C) = \sigma^2(1-\rho_h^2)\,2^{-2\delta}.$$

RBU is the reduction in the *minimum achievable* MSE obtained by adding C to the conditioning set, i.e. the difference of the two MMSEs:

$$\mathrm{RBU} = \mathrm{MMSE}(X_{t+h} \mid X_t) - \mathrm{MMSE}(X_{t+h} \mid X_t, C) = \sigma^2(1-\rho_h^2)\left(1-2^{-2\delta}\right).$$

Both factors are non-negative ($\rho_h^2 \le 1$, $\delta \ge 0$), giving the stated decomposition. □

Non-Gaussian behavior. The closed form relies on the Gaussian entropy–variance identity holding for *both* conditional laws, which fails in general. Two distinct effects then arise, and we keep them separate. (i) *At the population level*, the identity $\mathrm{MMSE} = \sigma^2(1-\rho_h^2)2^{-2\delta}$ need no longer hold, so the population RBU formula is, in general, neither an upper nor a lower bound on the true achievable MMSE reduction. (ii) *At the finite-sample level*, our plug-in $\widehat{\mathrm{RBU}}$ uses a Kraskov estimate $\hat{\delta}$, which under-detects nonlinear dependence (Section 4.5) and hence tends to be *smaller* than the true δ; this makes $\widehat{\mathrm{RBU}}$ tend to *underestimate* the population RBU. The two should not be conflated: (i) concerns the formula, (ii) concerns the estimator. We therefore rely on $\widehat{\mathrm{RBU}}$ only as an interpretive sanity check, and on the distribution-free Fact 2 (not on RBU) for the strict ceiling.

B Detailed Experimental Results

B.1 MoME Modulation Ablation Detail

The `-modulation` flag in MoME controls Expert-level Language Modulation (EiLM) [35], a FiLM-style [26] layer that applies text-conditioned affine transformation ($\gamma \cdot x + \beta$) to each expert's output after routing. With modulation off:

- Expert routing (Gate → softmax → top-k) remains unchanged and text-independent.
- Expert weights and structure are preserved.
- Only the post-expert text conditioning (EiLM) is removed.
- Parameter difference: ~74K out of 14.3B (<0.001%), negligible.

This means our "routing contribution" measures the value of *text-conditioned expert modulation*, not expert selection itself (which is text-independent in MoME).

B.2 MoME Shortcut Experiment (Finding 2)

Table 4 reports the full 3-seed results for the MoME shortcut experiment across 3 datasets. Adding a last-value shortcut ($\hat{y} = f_{\mathrm{MoME}} + X_t$) during training consistently suppresses routing contribution by 77–93%.

Table 4: MoME shortcut experiment across 3 datasets (3-seed each). Adding shortcut suppresses routing contribution by 77–93%. Energy and Environment could not be run due to GPU memory constraints.

Dataset	Without Shortcut	With Shortcut	Suppression
HealthUS (MSE)	+62.9%	+14.1%	77%
SocialGood (MSE)	+34.3%	+3.3%	90%
HealthAFR (MSE)	+24.2%	+1.7%	93%

HealthUS detail: Without shortcut: mod MSE = 0.334 ± 0.080, nomod MSE = 0.901 ± 0.073. With shortcut: mod MSE = 1.092 ± 0.020, nomod MSE = 1.271 ± 0.036.
SocialGood detail: Without shortcut: mod MSE = 0.435 ± 0.034, nomod MSE = 0.662 ± 0.009. With shortcut: mod MSE = 0.874 ± 0.016, nomod MSE = 0.904 ± 0.005.

HealthAFR detail: Without shortcut: mod MSE = 0.610 ± 0.036, nomod MSE = 0.804 ± 0.009. With shortcut: mod MSE = 0.926 ± 0.016, nomod MSE = 0.942 ± 0.005.

Negative control (FinMultiTime): On FinMultiTime ($\rho = 0.999$), the repeat-last-value naive baseline achieves MSE 8.8×10^{-5}, which already *outperforms* MoME with modulation (9.1×10^{-5}). Adding an architectural shortcut to MoME on this dataset would not change routing contribution: it is already effectively zero because the temporal signal completely dominates and no residual variance is left for routing to exploit.

B.3 Context Degradation Detail (Finding 1)

Table 5 reports the full results for the context degradation experiment.

Table 5: Context degradation on HealthUS (MoME, all rows 10-seed). Routing contribution (mod on vs. off) decreases with corruption. At 100% mask, the residual +16% is from EiLM capacity, not context. Context-specific benefit (real text vs. constant text, same architecture) = 44.2%. The 0%-mask row is the standard all-real-text setup and matches Table 1 exactly (mod 0.399, +50.9%).

Mask Rate	MSE (mod)	MSE (nomod)	Routing Contrib.	Seeds
0% (clean)	0.399 ± 0.129	0.817 ± 0.049	+50.9%	10
25%	0.528 ± 0.101	0.834 ± 0.040	+36.3%	10
50%	0.737 ± 0.094	0.834 ± 0.053	+11.0%	10
75%	0.750 ± 0.079	0.806 ± 0.055	+6.9%	10
100% (constant)	0.715 ± 0.074	0.851 ± 0.143	+16.1%	10

Interpreting the 100% mask residual. At 100% mask, all samples receive identical text, so EiLM computes a *constant* γ, β for every sample, which is equivalent to a learned bias and scale (~74K parameters). The +16% residual reflects this capacity benefit, not context exploitation. To isolate context from capacity, we compare MSE at 0% mask (mod on, real text: 0.399) vs. 100% mask (mod on, constant text: 0.715), where the architecture and parameter count are identical and only the text content differs. The 44.2% MSE reduction is attributable solely to context information.

B.4 Multi-Model Context Contribution (Finding 3)

Table 6 reports the full 3-seed results for the multi-model experiment. Effects are generally small in our testbed (±5%), consistent with Finding 5 (nonlinear signal requires high capacity to exploit).

B.5 Power Analysis Detail (Finding 4)

We estimate MI test power via simulation. For each combination of sample size n and true effect size δ, we generate 30 synthetic datasets with known MI (Gaussian context with controlled correlation to residual), run the 200-permutation MI test, and compute the rejection rate at $\alpha = 0.05$.

The bottom row confirms Type I error control (≤ 0.07 at all sample sizes). The power classification used in Table 3 is: HIGH = $n \geq 500$ with unique context per sample (power ≥ 0.6 for $\delta \geq 0.5$); LOW = otherwise.

Table 6: Context contribution (%) across fusion mechanisms (testbed, no shortcut). Energy Cross-Attention and Text-as-Variable use 10-seed evaluation (marked †); all others use 3-seed. MI-non-significant datasets show no positive contribution; negative values reflect training instability when attending to uninformative context.

Mechanism	Health (MI sig)	Transport (MI sig)	Energy (MI n.s.)	Climate (MI n.s.)	Web (MI n.s.)
Cross-Attention	+3.2±0.9%	+2.9±0.2%	−5.9±0.8%†	−0.5±0.3%	0.0%
Gating	−1.1±2.0%	+3.6±0.4%	+0.3±0.1%	−0.0±0.1%	0.0%
Text-as-Variable	+2.3±0.5%	−0.3±0.0%	−6.9±1.0%†	+1.1±0.1%	0.0%
Output Fusion	−1.0±0.4%	+9.5±0.1%	−0.3±0.3%	+2.2±0.3%	0.0%

† 10-seed evaluation. Cross-Attention: t=1.50, p >0.05 (not significant). Text-as-Variable: t=2.27, borderline. The original 3-seed Cross-Attention estimate (−18.5%) was inflated by one outlier seed with 6× higher variance than the zero-context condition.

Table 7: Estimated power of MI permutation test (rejection rate at $\alpha = 0.05$, 30 simulations per cell). Power increases with both n and δ. At $n \geq 500$ and $\delta \geq 0.5$, power exceeds 0.6.

True δ (bits)	n=100	n=200	n=500	n=1000	n=2000
0.1	0.07	0.03	0.10	0.13	0.20
0.3	0.10	0.17	0.40	0.70	0.90
0.5	0.13	0.30	0.63	0.87	0.97
1.0	0.27	0.53	0.90	0.97	1.00
0.0 (Type I)	0.03	0.03	0.03	0.07	0.03

B.6 Large-Scale Negative Validation (50 S&P500 Tickers)

We compute MI permutation tests for 50 S&P500 tickers from FinMultiTime. All tickers have $\rho_3 > 0.99$ (stock prices are near-unit-root processes); MI tests on daily returns show non-significant context ($p > 0.05$) for all 50 tickers. Routing benefit (measured via our lightweight testbed) is within noise (±3%) for all tickers, confirming the diagnostic's negative prediction at scale.

B.7 Multi-Horizon Validation

Across $h \in \{1, 3, 7, 14\}$ on 3 Time-MMD datasets (12 configurations total), the low-RBU threshold (<0.1) correctly predicts negative routing benefit on 11/12 points. The single exception (Health h=1, +5.3%) reflects routing acting as regularization on a very small effective dataset at short horizon, not context exploitation.

B.8 Comprehensive Baselines

For completeness, Table 8 reports the full multi-modal baseline results on Time-MMD (all methods share the same PatchTransformer backbone with the shortcut enabled), complementing the no-shortcut testbed results in Table 6.

Table 8: Full multi-modal baseline results on Time-MMD (wQL ↓, 3-seed mean ± std). All methods use the same Patch-Transformer backbone with shortcut enabled.

Method	Health	Transport	Energy	Climate	Web
DLinear	.340±.002	**.056**±.001	**.073**±.001	.197±.000	.293±.004
PatchTST	**.285**±.014	.079±.005	.126±.007	.225±.006	.378±.005
Output Fusion	.275±.005	.057±.001	.076±.002	**.194**±.002	.228±.014
Cross-Attention	.290±.024	.059±.002	.080±.001	.197±.001	.235±.005
Gating	.279±.017	.063±.002	.082±.000	.200±.002	.229±.004
Text-as-Variable	**.262**±.012	.058±.001	.089±.013	.197±.003	.232±.023
MoAT	**.287**±.011	.063±.002	.099±.002	.198±.000	.393±.009
Testbed (routing)	.300±.043	.063±.002	.083±.001	.197±.005	.227±.005
Testbed (uniform)	.298±.018	.066±.000	.081±.000	.194±.002	.219±.002