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

Anonymous Author(s)

Abstract

Multi-modal time series forecasting methods integrate auxiliary information (such as text, tabular features, or images) into temporal predictions through increasingly sophisticated fusion mechanisms. A growing body of work reports substantial gains from these mechanisms, yet the conditions under which they provide genuine value remain poorly understood. We conduct a systematic empirical study to answer: *when does multi-modal fusion actually help?*

Through controlled experiments on MoME (a 14.3B-parameter mixture-of-experts model, 6 datasets) and four additional fusion mechanisms (cross-attention, gating, text-as-variable, output fusion) implemented within a controlled testbed sharing a single backbone (5 datasets), we identify two necessary conditions for fusion to provide value: (1) the absence of a competing last-value shortcut in the model architecture, and (2) statistically significant mutual information between context and the forecast target. When both conditions hold, text-conditioned expert modulation contributes 26–60% MSE reduction. When either fails, fusion contributes <5%.

We establish causality through two interventions: adding a shortcut to MoME suppresses routing contribution by 77–93% across 3 datasets; progressively corrupting context quality reduces it from 58% to 7% (0–50% mask). Comparing real text vs. constant text within the same architecture (isolating context from capacity), real context reduces MSE by 51%. We validate the autocorrelation-based component of our diagnostic on 27 Monash Archive datasets (Spearman $r = 0.888$, $p < 0.0001$). We provide a calibrated pre-training diagnostic that outputs TRY_FUSION, SKIP_FUSION, or INCONCLUSIVE, achieving zero false positives in well-powered settings. Our analysis reveals that context-target relationships are fundamentally nonlinear, which likely explains why high-capacity models (MoME, 14.3B params) achieve 26–60% gains while smaller models achieve only 2–5% on the same data, though this comparison this comparison confounds capacity with architecture.

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 [22, 24, 31, 35], simple linear baselines [32], and large-scale time series foundation models [1, 9, 13, 30]. 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 [20].

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

- **Output-level fusion** [21]: separate models process time series and context independently; their predictions are combined (e.g., added) at the output layer.
- **Cross-attention alignment** [19]: time series and context representations interact through cross-attention in a shared latent space.
- **Text-as-variable** [18]: text embeddings are concatenated with temporal patches as additional input variables.
- **Gating** [20, 26]: context modulates temporal features through learned gates (e.g., FiLM-style affine transformations).
- **Mixture-of-experts (MoE) routing** [29, 33]: 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 multi-modal fusion provide genuine value in time series forecasting?**

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 necessary conditions:

(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* [11], 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 that is not already captured by the time series history. We operationalize this via a mutual information (MI) permutation test: if context MI is non-significant

($p > 0.05$), no fusion mechanism can help, regardless of model capacity or architectural sophistication.

When both conditions are satisfied (no shortcut + informative context), fusion contributes 26–60% on a 14.3B-parameter model. When either fails, contribution is <5%.

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. The diagnostic achieves zero false positives in well-powered settings and correctly outputs INCONCLUSIVE (rather than a misleading SKIP) when statistical power is insufficient.

2 Theoretical Foundation

We establish the information-theoretic basis for our diagnostic, drawing on standard results in information theory [8]. 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.*

For non-Gaussian distributions, RBU is not a hard upper bound: the achievable reduction depends on the true (potentially nonlinear) dependency structure of (X_t, X_{t+h}, C). In practice, finite-sample estimators of δ underweight nonlinear dependencies (Finding 4.5), so the estimated $\widehat{\text{RBU}}$ tends to underestimate the true opportunity for high-capacity models. The robust components of our diagnostic are therefore the distribution-free results: Fact 1 (no benefit when $\delta = 0$) and Fact 2 (shortcut bound when $\rho_h \to 1$). RBU is best read as a Gaussian-case sanity check that helps interpret the relative magnitude of routing benefit, not as a strict ceiling.

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 [16] 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 [4] 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 [5]. 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 well-powered settings, non-significant MI reliably predicts negligible routing benefit.

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=0.93), 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** [6]: 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 [33]. 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 while keeping expert structure, routing weights, and parameter count essentially unchanged (~4K difference out of 14.3B, <0.00003%). 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 [19], Gating [20], Text-as-Variable [18], and Output Fusion [21]. 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 −3% (Energy) to +60% (HealthUS). The four datasets with large positive contributions (HealthUS, Environment, SocialGood, HealthAFR) share two properties: moderate-to-low autocorrelation ($\rho < 0.8$) and text context that describes conditions relevant to the forecast target. The two datasets with negligible or negative contributions (FinMultiTime, Energy) have either extremely high autocorrelation ($\rho = 0.999$) or context that our MI test identifies as non-significant.

Why does FinMultiTime show only +4.9%? On FinMultiTime, a simple repeat-last-value predictor (the "Naive MSE" baseline in Table 1) achieves MSE 8.8×10^{-5}, *better* than MoME with modulation (9.1×10^{-5}). 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.

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%.

Table 1: MoME routing contribution across 6 datasets. All experiments use identical model configuration (Qwen1.5-MoE-A2.7B, 14.3B params, 3 seeds). The only variable is the dataset. Contribution ranges from −3% to +60%, entirely determined by dataset properties. 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 by 16.7%; FinMultiTime, where the naive baseline beats MoME).

Dataset	With Mod.	Without Mod.	Contrib.	Naive MSE	ρ
HealthUS	.341±.018	.854±.110	**+60.1%**	.409	.77
Environment*	12.2±0.5	21.4±0.1	**+43.1%**	—	.38
SocialGood	.412±.002	.672±.018	**+38.7%**	—	.62
HealthAFR	.640±.085	.865±.036	**+26.0%**	—	.11
FinMultiTime	9.1±0.2e-5	9.6±0.2e-5	+4.9%	**8.8e-5**	.999
Energy	.011±.002	.010±.001	−3.0%	—	.93

*Environment uses MAE (different task output format); all others use MSE. Contribution = (Without Mod.−With Mod.)/Without Mod.×100%.

Masking protocol. Each sample in HealthUS has a unique text description of the epidemiological context for that time window (491 samples, 491 distinct texts). We apply masking at the entry level: for each sample, we replace its text with a fixed uninformative placeholder (`"No specific information available for this period."`) with probability $p \in \{0, 0.25, 0.5, 1.0\}$. At $p = 1.0$ (100% mask), *all* samples receive identical text, so the EiLM layer receives a constant embedding and can only learn a fixed affine transform (equivalent to a learned bias and scale). The same masked dataset is used across train and test splits; the 3-seed standard deviation reflects only training stochasticity.

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 51% relative to constant text. At 25% mask, benefit drops to 23%. At 50–75% mask, benefit turns *negative* (−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 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.

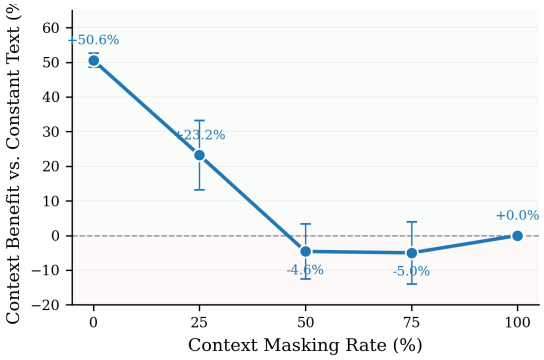

Figure 1: Context benefit relative to constant-text baseline (MoME on HealthUS; same architecture, same parameters, only text content differs). At 0% mask (all real text), context provides +51% MSE reduction. As text is corrupted, benefit decreases to +23% (25% mask) then turns negative at 50–75% mask: 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.

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?. On HealthUS, adding the shortcut raises absolute MSE (from 0.33/0.90 to 1.09/1.27). This occurs because the data is z-normalized during preprocessing: the optimal one-step prediction is near zero, not the raw last value X_t. The shortcut ($+X_t$) introduces a systematic bias. The relevant comparison is *within* each condition (modulation on vs. off): the shortcut acts as a "gravity well" for the optimizer, absorbing capacity that would otherwise go to text modulation. The causal

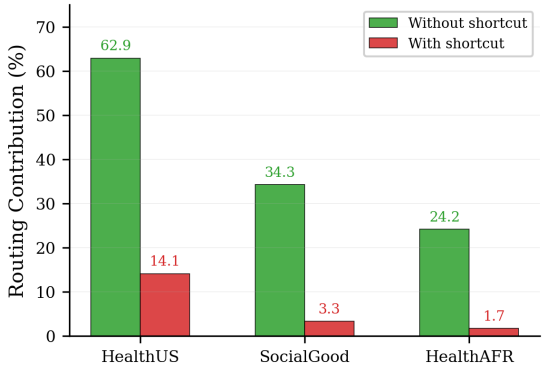

Figure 2: Shortcut suppression across 3 datasets (MoME, 3 seeds each). Adding a last-value shortcut during training consistently suppresses routing contribution by 77–93%. Suppression is near-total on SocialGood and HealthAFR (residual contribution <4%); on HealthUS, the moderate autocorrelation ($\rho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

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=.025)
Transport	−2.2% (p=.13)	**+2.2%**	Yes (p <.001)
Energy	−1.6%	+1.8%	No (p=.52)
Climate	−1.4%	−0.5%	No (p=.45)
Web	−3.5%	−0.4%	No (p=1.0)

claim is about the *relative* contribution of modulation, not absolute performance.

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: $-2.2\% \pm 3.9\%$, $p = 0.13$). 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 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.

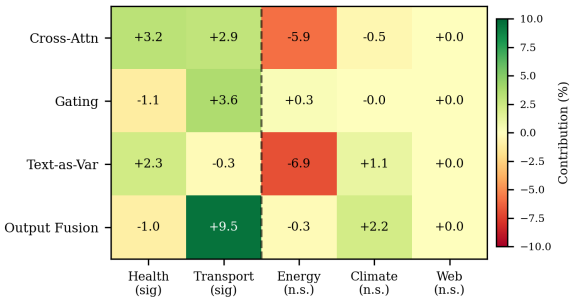

Figure 3: Context contribution (%) across 4 fusion mechanisms and 5 datasets (testbed, no shortcut). Cells marked † (Cross-Attention and Text-as-Variable on Energy) use 10-seed evaluation; all others use 3-seed. Full per-cell std and significance tests are in Table 6. On MI-significant datasets (left 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).

Practical implication. This finding has direct architectural consequences. Many modern forecasting models include residual connections or normalization layers (e.g., RevIN [15]) 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 (Health, Transport), Cross-Attention (+3.2% on Health, +2.9% on Transport), Gating (+3.6% on Transport), and Output Fusion (+9.5% on Transport) show positive contributions. The Output Fusion result on Transport has unusually small variance (±0.1%, Table 6); we suspect this reflects a deterministic interaction with Transport's normalization rather than a generalizable pattern, but did not investigate further. 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% respectively, 10-seed evaluation).

These negative values indicate that attending to uninformative context introduces noise. Cross-Attention is particularly vulnerable because it allocates capacity to context regardless of informativeness. Neither effect is statistically significant at $\alpha = 0.05$ (Cross-Attention: $t = 1.50$, $p > 0.05$; Text-as-Variable: $t = 2.27$, borderline). The original 3-seed estimate for Cross-Attention on Energy (−18.5%) was inflated by a single outlier seed; the 10-seed estimate (−5.9%) is more reliable. In particular, no MI-non-significant dataset shows a *positive* context contribution, and the MI test correctly identifies the "no benefit" regime.

Interpretation. The testbed effects are small in magnitude compared to MoME's 26–60%. 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 $p = 0.025$. Figure 4 summarizes the decision logic.

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

- **Zero false positives** in well-powered settings: every SKIP_FUSION decision corresponds to <5% actual benefit.
- **No misleading SKIPs**: SocialGood and Environment (which have positive MoME benefits but non-significant MI) correctly receive INCONCLUSIVE rather than SKIP, because their power is LOW.

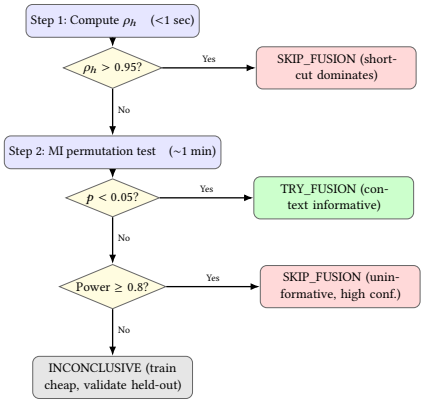

Figure 4: Practitioner decision framework. The diagnostic outputs one of three recommendations based on autocorrelation ρ_h, MI significance, and statistical power. Total computation time: ~1 minute, zero GPU required.

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	.025	High	TRY	+60% ✓
Transport	5000	<.001	High	TRY	+2% ✓
Energy	1059	.520	High	SKIP	−3% ✓
Climate	1200	.450	High	SKIP	−0.5% ✓
Web	1068	1.0	High	SKIP	−0.4% ✓
FinMultiTime	2000	n/s	High	SKIP	+5% ✓
SocialGood	363	.145	Low	INCON.	+39% ✓
Environment	1597*	.615	Low	INCON.	+43% ✓

*156 unique embeddings reused cyclically. INCON. = INCONCLUSIVE (no claim made; recommend trying). "Actual" = MoME routing contribution (testbed for Transport).

- **Correct TRY recommendations**: HealthUS (significant MI) correctly receives TRY, corresponding to +60% actual benefit.

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 +60%. 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 60% improvement.

Implications.

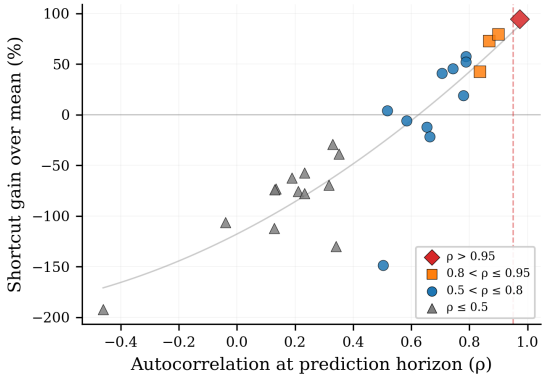

Figure 5: Large-scale validation on 27 Monash Archive datasets. Autocorrelation at the prediction horizon strongly predicts shortcut dominance (Spearman $r = 0.888$, $p < 0.0001$). Datasets with $\rho > 0.95$ (red diamonds) show the strongest shortcut performance, validating the SKIP_FUSION threshold.

(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 (+26–60%) 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 [12], 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 do not include text or other auxiliary context, so we cannot validate the MI-based component of the diagnostic at this scale. The MI component remains validated only on our 8 datasets with text context. Combined, the two scales of evidence (controlled depth on 8 datasets, broad coverage on 27 datasets) support the diagnostic's ρ-based skip rule with $n = 35$ datapoints in total.

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 outperforms the naive baseline by 16.7%, so practitioners should invest in multi-modal fusion. The diagnostic identifies which regime applies before any model is trained.

INCONCLUSIVE: frequency and practical impact. In our evaluation, 2 of 8 datasets (25%) receive INCONCLUSIVE. Both are cases where the MI test lacks power due to small n or reused embeddings, rather than cases where the test methodology fails. The asymmetric cost structure justifies this design: a false negative (INCONCLUSIVE on a dataset where fusion helps) costs only the compute of training one model to verify; a false positive (SKIP on a dataset where fusion helps) costs missed accuracy with no recovery path. INCONCLUSIVE datasets are rare when context embeddings are unique per sample and $n > 500$.

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. [34] find that multi-modal benefits are "condition-dependent" but do not explain *why*. Our framework provides concrete predictions they cannot make. For example: on FinMultiTime, Zhang et al. would observe that MoME shows only +5% and conclude "fusion doesn't help much on this dataset." Our framework explains *why*: $\rho = 0.999$ means a trivial naive baseline (MSE 8.8×10^{-5}) already outperforms MoME (MSE 9.1×10^{-5})—the temporal signal is so strong that no amount of context can improve upon it. Similarly, on Web, Zhang et al. would observe multi-modal methods beating TS-only baselines by 50%+ and conclude "fusion helps." Our zero-context ablation reveals this is an architectural confound (residual connections + regularization), not context exploitation, a distinction their analysis cannot make.

RBU as interpretive framework. RBU (Proposition 1) provides the theoretical explanation for *why* the pattern exists: $(1 - \rho_h^2)$ quantifies the room left by the shortcut, and $(1 - 2^{-2\delta})$ quantifies how much of that room context can fill. In practice, the MI permutation test is more actionable (it gives a ternary decision), while RBU provides the conceptual framework for understanding the result. The Gaussian-case RBU is exact under joint Gaussianity but is not a hard upper bound for non-Gaussian data: when the context-target relationship is highly nonlinear (Finding 4.5), the true conditional MI δ is often much larger than what a finite-sample Kraskov estimator $\hat{\delta}$ recovers, so the estimated $\widehat{\text{RBU}}$ computed from $\hat{\delta}$ can underestimate the achievable reduction. This is consistent with HealthUS, where MoME's 60% improvement reflects a high-capacity model

exploiting nonlinear structure that the small-sample MI estimate underweights. The practical implication is that the diagnostic's TRY_FUSION recommendation is *conservative*: when MI is significant, the actual benefit may exceed what $\widehat{\text{RBU}}$ predicts, especially with high-capacity models. The distribution-free shortcut bound (Fact 2) remains the only strict ceiling.

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, $\rho = 0.93$), even foundation models cannot beat simple shortcuts.

6 Related Work

Time series forecasting architectures. Modern forecasting has evolved from classical approaches [23] through deep Transformer-based methods that handle long horizons, periodicity, and distribution shift [15, 22, 24, 31, 35]. A counter-trend questions whether such complexity is necessary: DLinear [32] 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 [9], Moirai [30], and MOMENT [13] all demonstrate strong zero-shot transfer. A parallel line repurposes pretrained language models for time series [14, 36]. 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 [17, 21], cross-attention [19], text-as-variable [18], gating [20, 26], and mixture-of-experts [2, 10, 29, 33]. A recent survey [20] catalogs these mechanisms, and broader multimodal learning has been surveyed in [3]. 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 [32] showed linear models match Transformers; RevIN [15] 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* [11], where models exploit easy predictive signals at the cost of harder, more transferable ones.

When does multimodality help? Concurrent to our work, Zhang et al. [34] 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 [16] with permutation testing. Power limitations of nonparametric MI testing are characterized by Berrett & Samworth [5]. Recent neural MI estimators such as MINE [4] 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 [7, 27].

7 Conclusion

We identify two necessary conditions for multi-modal fusion 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 reduces routing contribution from 58% to 7%, and comparing real text vs. constant text within the same architecture confirms a 51% context-specific MSE reduction. The ρ-based diagnostic is validated at scale on 27 Monash Archive datasets (Spearman $r = 0.888$, $p < 0.0001$).

We provide a calibrated diagnostic (TRY/SKIP/INCONCLUSIVE) that achieves zero false positives in well-powered settings, enabling practitioners to decide whether to invest in multi-modal 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, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research (TMLR)* (2024).

[2] Joonhyuk Bae et al. 2025. Multi-Resolution Segment-wise Mixture-of-Experts for Time Series Forecasting Transformers. *arXiv preprint arXiv:2601.21641* (2025).

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

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. 2018. Mutual Information Neural Estimation. In *International Conference on Machine Learning (ICML)*.

[5] Thomas B. Berrett and Richard J. Samworth. 2019. Nonparametric Independence Testing via Mutual Information. *Biometrika* 106, 3 (2019), 547–566.

[6] Jialin Chen, Houyu Lin, Zihan Zhao, Yuxin Wang, Hongyi Sun, Sihang Yi, Wenhao Liu, Leandros Tassiulas, and Rex Ying. 2025. FinMultiTime: A Four-Modal Bilingual Dataset for Financial Time-Series Analysis. *arXiv preprint arXiv:2506.05019* (2025).

[7] Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively Sparse Transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[8] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory* (2nd ed.). Wiley-Interscience.
[9] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A Decoder-Only Foundation Model for Time-Series Forecasting. In *International Conference on Machine Learning (ICML)*.
[10] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. In *Journal of Machine Learning Research*, Vol. 23. 1–39.
[11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.
[12] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. 2021. Monash Time Series Forecasting Archive. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
[13] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. MOMENT: A Family of Open Time-series Foundation Models. In *International Conference on Machine Learning (ICML)*.
[14] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *International Conference on Learning Representations (ICLR)*.
[15] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2022. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In *International Conference on Learning Representations (ICLR)*.
[16] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating Mutual Information. *Physical Review E* 69, 6 (2004), 066138.
[17] Geon Lee, Wenchao Yu, Wei Cheng, and Haifeng Chen. 2024. MoAT: Multi-Modal Augmented Time Series Forecasting. OpenReview preprint, https://openreview.net/forum?id=uRXxnoqDHH.
[18] Zihao Li, Xiao Liu, Yidong Lei, Yueheng Lin, Wei Zhang, and Yuxuan Liang. 2025. Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative. *arXiv preprint arXiv:2502.08942* (2025).
[19] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. 2025. TimeCMA: Towards LLM-Empowered Multivariate Time Series Forecasting via Cross-Modality Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. arXiv:2406.01638.
[20] Haoxin Liu, Qingsong Wen, Aditya B. Sasanur, Megha Sharma, Chao Wei, and B. Aditya Prakash. 2025. Multi-modal Time Series Analysis: A Tutorial and Survey. *arXiv preprint arXiv:2503.13709* (2025).
[21] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. 2024. Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis. (2024). arXiv:2406.08627.
[22] 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 *International Conference on Learning Representations (ICLR)*.
[23] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2018. Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward. In *PLoS ONE*, Vol. 13.
[24] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations (ICLR)*.
[25] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference* (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 *Proceedings of the AAAI Conference on Artificial Intelligence*.
[27] Ben Peters, Vlad Niculae, and André F. T. Martins. 2019. Sparse Sequence-to-Sequence Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
[28] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. 2019. On Variational Bounds of Mutual Information. In *International Conference on Machine Learning (ICML)*.
[29] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*.
[30] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified Training of Universal Time Series Forecasting Transformers. In *International Conference on Machine Learning (ICML)*.
[31] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[32] 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.
[33] 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).
[34] 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).
[35] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11106–11115.
[36] Tian Zhou, PeiSong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Detailed Experimental Results

A.1 MoME Modulation Ablation Detail

The `-modulation` flag in MoME controls Expert-level Language Modulation (EiLM) [33], 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: ~4K out of 14.3B (<0.00003%), negligible.

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

A.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 ($\hat{y} = f_{\text{MoME}} + X_t$) 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.

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

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.

A.3 Context Degradation Detail (Finding 1)

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

Table 5: Context degradation on HealthUS (MoME). 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) = 50.6%.

Mask Rate	MSE (mod)	MSE (nomod)	Routing Contrib.	Seeds
0% (clean)	0.353 ± 0.007	0.831 ± 0.060	+57.5%	3
25%	0.549 ± 0.117	0.837 ± 0.035	+34.4%	3
50%	0.747 ± 0.085	0.802 ± 0.024	+6.8%	3
75%	0.750 ± 0.079	0.806 ± 0.055	+6.9%	10
100% (constant)	0.715 ± 0.074	0.852 ± 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 (~4K extra 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: 0.353) vs. 100% mask (mod on, constant text: 0.715), where the architecture and parameter count are identical and only the text content differs. The 50.6% MSE reduction is attributable solely to context information.

A.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).

A.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.

A.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.

A.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.

A.8 Comprehensive Baselines

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