OSVIF: ORDER-SPECTRAL VARIABLE INTERACTION FRAMEWORK FOR GENERAL MULTIVARIATE TIME SERIES ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-variable dependencies in multivariate time series involve variable combinations of different sizes. These combinations can provide distinct information for characterizing a target variable and contribute jointly through cross-scale synergies. The roles of both combination information and these synergies vary across target variables. However, existing methods often mix information from combinations of different sizes or use constructed group structures to represent only a subset of combinations separately. This organization makes it difficult to preserve distinct representations of combination information by size, limiting the modeling of this information and its cross-scale synergies, as well as its selective use for different target variables. Consequently, cross-variable dependencies remain insufficiently characterized, limiting performance on downstream time series analysis tasks. To address these limitations, we propose OSVIF (Order-Spectral Variable Interaction Framework), a general framework for multivariate time series analysis. For each target variable, OSVIF covers all variable combinations and uses two spectra to link the organization of combination information by size with the modeling of cross-scale synergies. The natural order spectrum covers all source-variable combinations, preserves distinct representations by size, and applies target conditioning to form order-specific states. The natural coalition spectrum covers all combinations formed from these order-specific states and the state of the target variable itself, modeling how information from combinations of different sizes contributes synergistically to the characterization of the target variable. Both spectra use generating polynomials to cover their respective combination spaces in full without enumerating individual combinations, keeping computational costs manageable. By combining full combination coverage with distinct representations by size, OSVIF enables the selective use of combination information for each target variable and captures cross-scale synergies. It provides a unified variable interaction backbone for five tasks: long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection. Experiments on 35 benchmark datasets show that OSVIF achieves strong performance across all five tasks, demonstrating its effectiveness and generality.

1 INTRODUCTION

Variables in multivariate time series often exhibit cross-variable dependencies (Wu et al., 2020; Xue et al., 2026). Modeling these dependencies directly affects performance on downstream tasks such as forecasting, imputation, classification, and anomaly detection. In many practical settings, characterizing a target variable requires information from both individual variables and combinations of multiple variables. Combinations of different sizes capture complementary aspects of multivariate dynamics, and the relevance of a given combination varies across target variables. We define combination size as the number of variables in a combination. Combinations of size two, for example, comprise all pairs of variables. For each target variable, we refer to all remaining variables as source variables. In Figure 1(a), the horizontal axis denotes the combination size m, that is, the number of source variables in a combination. The figure provides empirical support for these observations. The heatmap shows the relative weights of information from combinations of different sizes for each

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

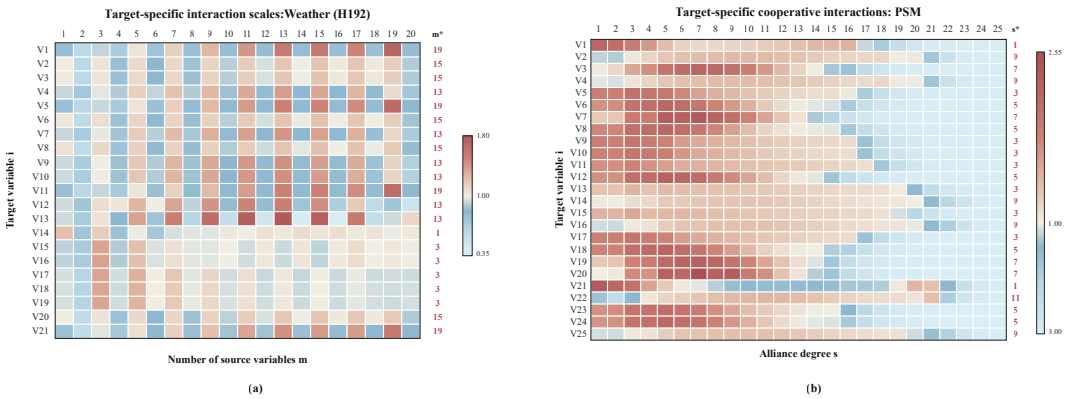

Figure 1: Target-dependent heterogeneity in the contributions of different combination sizes and their cross-size synergies.

target variable on the Weather dataset (Max Planck Institute for Biogeochemistry, n.d.). For a given target, these weights vary across sizes, and their distributions differ across targets. These patterns suggest that target variables rely on information from different combination sizes to varying degrees, motivating the selective use of this information for each target. Modeling cross-variable dependencies therefore calls for a target-centered approach that systematically organizes and distinguishes the contributions of combinations by size and adapts their use to each target.

However, existing approaches to organizing variable combinations provide limited support for distinguishing information by size and using it selectively for each target. Methods based on pairwise interactions can incorporate combination information (Zhang & Yan, 2023; Liu et al., 2024; Wu et al., 2020), but they often mix information from different combination sizes within a unified representation, making the contributions associated with each size difficult to distinguish for a given target variable. Methods that model variable groups represent some combinations separately (Tian et al., 2025; Xiao et al., 2025; Xue et al., 2026), but their coverage is constrained by the constructed group structures. Information from relevant combinations that are not separately represented must be captured through subsequent propagation and fusion, which also mix this information into a unified representation. As a result, information from different combination sizes remains difficult to preserve separately and use selectively across targets. This limitation can leave cross-variable dependencies insufficiently characterized and limit performance on downstream time series analysis tasks.

Distinguishing and selectively using information from different combination sizes is not sufficient to fully characterize cross-variable dependencies for a given target. Information captured by combinations of one size often needs to be interpreted in the context of information from other sizes. Characterizing a target variable therefore requires both distinguishing combination information by size and accounting for synergies between combinations of different sizes. The roles of these synergies also vary across target variables. We define coalition order s as the number of natural-order states participating in a coalition. In Figure 1(b), the horizontal axis denotes coalition order s, and the heatmap shows the relative contributions of cross-scale synergies to each target variable on the PSM dataset (Abdulaal et al., 2021). For a given target, these synergies contribute to varying degrees, and their contribution distributions differ across targets. A more complete characterization of cross-variable dependencies therefore requires a target-centered approach that distinguishes and selectively uses combination information by size while modeling its cross-scale synergies.

Existing studies model such synergies through hierarchical relationships among groups (He et al., 2022; Wu et al., 2025a) or using feature transformations and information exchange between groups (Ying et al., 2025; Ren et al., 2026). However, the combinations that participate directly in joint modeling remain constrained by the constructed group structures. When combinations relevant to a target variable span a wide range of sizes, these structures can admit only some of them into direct joint modeling, leaving synergies between combinations of different sizes insufficiently characterized. This limitation can hinder the use of information from different combination sizes to assess the state of the target variable and how it changes, thereby affecting downstream performance.

Together, these observations motivate a target-centered approach that distinguishes the roles of variable combinations of different sizes and models their cross-scale synergies, providing an effective basis for long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection. We propose OSVIF (Order-Spectral Variable Interaction Framework), a general framework for multivariate time series analysis that uses two spectra to model the roles of source variable combinations of different sizes and their cross-scale synergies for each target variable.

The first layer constructs the natural order spectrum. OSVIF independently encodes each source variable and uses a generating polynomial for source combinations to cover all source variable combinations and organize their information by size. It then applies target conditioning to form a state at each natural order, enabling selective use of information from different sizes for each target variable. This process does not depend on a particular group structure or require individual combinations to be enumerated. Two complementary views, Identity and Cross, further capture within-coordinate and cross-coordinate interactions, respectively, enriching the representation of interactions within combinations at each natural order.

The second layer constructs the natural coalition spectrum. It treats the states at all natural orders and the state of the target variable itself as members, then uses a generating polynomial for state combinations to cover all member combinations. These combinations are grouped into coalition orders by the number of participating members to model synergies across combination sizes. Synergy modeling thus extends beyond combinations connected by a particular structure.

The two spectra bring the selective use of information from different combination sizes and the modeling of cross-scale synergies together in a unified representation backbone. This backbone provides a common basis for modeling cross-variable dependencies across five tasks: long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection. Our main contributions are as follows.

1. We introduce a target-centered perspective on cross-variable dependency modeling that distinguishes and selectively uses information from source combinations of different sizes while modeling their cross-scale synergies. This perspective supports the characterization of the state of a target variable and how it changes, providing a more complete representation of cross-variable dependencies for diverse time series analysis tasks.
2. We design OSVIF, whose natural order spectrum covers all source combinations, organizes their information by size, and forms states conditioned on the target variable. The natural coalition spectrum then combines these states with the state of the target variable itself to model cross-scale synergies. Both spectra use generating polynomials to cover their respective combination spaces in full without relying on predefined group structures or enumerating individual combinations, providing a single interaction backbone shared across all five tasks.
3. Experiments on 35 benchmark datasets show that OSVIF achieves strong performance in long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection using a unified representation backbone. These results demonstrate the effectiveness of the framework in supporting a broad range of time series analysis tasks through the selective use of combination information for each target variable and the modeling of cross-scale synergies.

2 Related Work

Multivariate time series often exhibit dependencies across variables, which existing methods capture through different strategies. Some process each variable separately using a channel-independent approach (Nie et al., 2023), whereas others integrate information across variables through attention or graph propagation (Zhang & Yan, 2023; Liu et al., 2024; Wu et al., 2020). Although the latter approaches incorporate information from variable combinations, they often mix information from combinations of different sizes within a shared representation, making it difficult to selectively use this information for each target variable.

To address this mixing, group-based methods construct separate representations for selected variable combinations. These methods represent combinations as hyperedges and, in some cases, further distinguish the roles of variables within each group (Tian et al., 2025; Xiao et al., 2025; Xue et al.,

2026). However, the combinations they explicitly represent remain limited by the constructed group structures. Information from relevant combinations that lack separate representations must still be captured through subsequent propagation and fusion. These operations mix such information within aggregated representations, making it difficult to maintain distinct representations for these combinations.

The effect of a variable combination on a target variable can depend on the information provided by other combinations, motivating the modeling of synergy among combinations. Existing studies jointly model variable combinations through hierarchical group relations (He et al., 2022; Wu et al., 2025a) or using feature transformations and information exchange between groups (Ying et al., 2025; Ren et al., 2026). However, the combinations directly involved in joint modeling remain limited by the group structures. Relevant combinations of different sizes can therefore remain insufficiently covered, limiting the characterization of their synergy.

Broadening the coverage of variable combinations and modeling their synergy require addressing the computational challenge posed by the rapidly growing number of possible combinations. One line of research exploits the algebraic structure of combination terms to avoid explicit enumeration through polynomial recurrences or dynamic programming (Duvenaud et al., 2011; Blondel et al., 2016). Another controls computational cost by restricting the set of combinations or compressing intermediate representations (Murphy et al., 2019; Lian et al., 2018). In particular, algebraic recurrences provide the computational foundation for efficiently constructing representations that cover all combinations and are organized by combination size.

Together, these studies establish foundations for representing variable combinations, modeling them jointly, and computing their information efficiently. However, existing approaches do not simultaneously provide full combination coverage, distinct organization by combination size, target-specific selection, and cross-scale synergy modeling. OSVIF addresses this gap with two generating-polynomial spectra. A detailed discussion of the relevant literature is provided in Appendix A.

3 METHOD

3.1 OVERALL FRAMEWORK

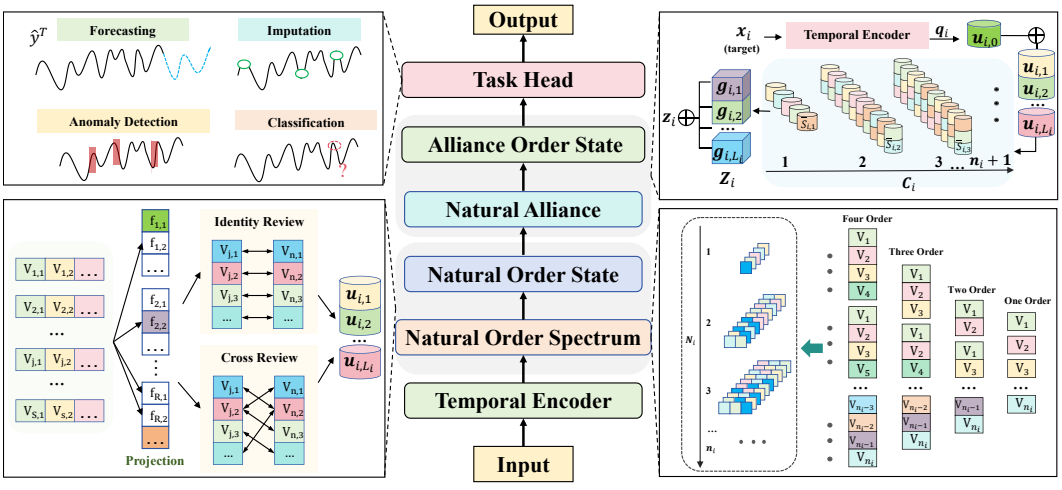

Figure 2: Overview of the OSVIF framework. The four task-head families cover five evaluation tasks because the forecasting head is used for both long-term and short-term forecasting.

Let T be the input sequence length and V the number of variables. Let $\mathbf{X} \in \mathbb{R}^{T\times V}$ be a multivariate time series and $\mathbf{O} \in \{0,1\}^{T\times V}$ its observation mask, where $O_{t,j} = 1$ indicates that variable j is observed at time t. We omit the batch dimension. For each target variable $i \in [V] := \{1,\dots,V\}$, the source-variable set is $\mathcal{A}_i = [V] \setminus \{i\}$, with $n_i = |\mathcal{A}_i| = V - 1$.

OSVIF constructs two target-centered spectra. The natural order spectrum organizes evidence from all source-variable combinations by the number of participating sources. Target conditioning converts each order into a natural order state. The target's direct state and all natural order states are then treated as members of a second combinatorial space. The natural coalition spectrum organizes all combinations of these members by coalition size and thereby represents their cross-scale synergies. Coalition-order states are shared across four head families: forecasting, imputation, classification, and anomaly detection. The forecasting head serves both long-term and short-term forecasting, giving five evaluation tasks in total. Figure 2 summarizes the pipeline:

$$(\mathbf{X}, \mathbf{O}) \to \mathcal{N}_i \to \mathbf{u}_{i,m} \to \mathcal{C}_i \to \mathbf{g}_{i,s} \to (\mathcal{Z}_i, \mathbf{z}_i) \to \widehat{\mathbf{y}}^{(\tau)}. \tag{1}$$

3.2 Natural Order Spectrum

Independent source representations. Combination order is meaningful only if variables remain unmixed before combination. Let d_e denote the encoder output dimension and d the common combination-space dimension. OSVIF encodes each source $j \in \mathcal{A}_i$ independently as $\mathbf{e}_j = \mathrm{Enc}(\mathbf{X}_{:,j}, \mathbf{O}_{:,j}) \in \mathbb{R}^{d_e}$, and maps it to a shared d-dimensional space as $\mathbf{v}_j = \eta(\mathbf{e}_j) \in \mathbb{R}^d$.

Generating polynomial. OSVIF represents the complete source-combination space with

$$P_i(z) = \bigodot_{j \in \mathcal{A}_i} (\mathbf{1}_d + \mathbf{v}_j z) = \sum_{m=0}^{n_i} \mathbf{E}_{i,m} z^m, \tag{2}$$

$$\mathbf{E}_{i,m} = \sum_{\substack{S \subseteq \mathcal{A}_i \\ |S|=m}} \bigodot_{j \in S} \mathbf{v}_j, \tag{3}$$

where z is a formal order marker and $\odot$ denotes the Hadamard product. Thus, $\mathbf{E}_{i,m}$ aggregates every combination containing exactly m sources without assigning an order to the members of a combination. The nonempty coefficients are normalized by their number of terms:

$$\overline{\mathbf{E}}_{i,m} = \frac{\mathbf{E}_{i,m}}{\binom{n_i}{m}}, \qquad \mathcal{N}_i = \{\overline{\mathbf{E}}_{i,m}\}_{m=1}^{n_i}. \tag{4}$$

The subset expansion defines the semantics of each coefficient; OSVIF computes all target-specific coefficients without explicit subset enumeration using the shared complement product tree in Appendix B.1.

3.3 Target-Conditional Natural-Order States

Hadamard aggregation retains interactions among matching representation coordinates. To complement these interactions with cross-coordinate structure, OSVIF learns R orthogonal directions $\{\mathbf{f}_r\}_{r=1}^R$, with $\mathbf{f}_r^\top \mathbf{f}_{r'} = \delta_{rr'}$, and projects each source as $p_{j,r} = \langle \mathbf{v}_j, \mathbf{f}_r \rangle$. For order m, define

$$\overline{e}_{i,m,r} = \frac{1}{\binom{n_i}{m}} \sum_{\substack{S \subseteq \mathcal{A}_i \\ |S|=m}} \prod_{j \in S} p_{j,r}, \qquad \mathbf{I}_{i,m} := \overline{\mathbf{E}}_{i,m}, \tag{5}$$

$$d_{i,m,r} = \left\langle \overline{\mathbf{E}}_{i,m}, \mathbf{f}_r^{\odot m} \right\rangle, \qquad c_{i,m,r} = \overline{e}_{i,m,r} - d_{i,m,r}, \qquad \mathbf{C}_{i,m} = \sum_{r=1}^{R} c_{i,m,r} \mathbf{f}_r. \tag{6}$$

$\mathbf{I}_{i,m}$ is the Identity view of within-coordinate interactions, whereas $\mathbf{C}_{i,m}$ is the Cross view formed from the residual cross-coordinate contribution. Appendix B.2 derives this decomposition. In particular, $\mathbf{C}_{i,1} = \mathbf{0}$.

Let $\mathbf{q}_i \in \mathbb{R}^{d_q}$ be the target condition produced by the temporal encoder. The two views are fused into an order-specific, target-conditional state:

$$\mathbf{u}_{i,m} = \Phi(\mathbf{I}_{i,m}, \mathbf{C}_{i,m}, \mathbf{q}_i), \qquad m = 1, \ldots, n_i. \tag{7}$$

The target's independently encoded representation is $\mathbf{h}_{i,0} = \mathrm{Enc}(\mathbf{X}_{:,i}, \mathbf{O}_{:,i}) \in \mathbb{R}^{d_e}$. It is mapped through the Identity branch of the same target-conditioned fusion to obtain the direct state $\mathbf{u}_{i,0} = \mathbf{h}_{i,0}^I$. The zero-preserving normalization, order encoding, and gated fusion are specified in Appendix B.3.

3.4 Natural Coalition Spectrum and Coalition-Order States

For target i, the target direct state and natural order states form the member set $\mathcal{M}_i = \{0, 1, \ldots, n_i\}$, whose size is $L_i = |\mathcal{M}_i| = n_i + 1$. Each member is mapped into a common coalition space as $\mathbf{a}_{i,m} = \phi_A(\mathbf{u}_{i,m}, \mathbf{q}_i) \in \mathbb{R}^{d_a}$ for $m \in \mathcal{M}_i$, where d_a is the coalition-space dimension. OSVIF then applies a second generating polynomial,

$$Q_i(z) = \bigodot_{m \in \mathcal{M}_i} (\mathbf{1}_{d_a} + \mathbf{a}_{i,m} z) = \sum_{s=0}^{L_i} \mathbf{S}_{i,s} z^s, \tag{8}$$

$$\mathbf{S}_{i,s} = \sum_{\substack{G \subseteq \mathcal{M}_i \\ |G| = s}} \bigodot_{m \in G} \mathbf{a}_{i,m}. \tag{9}$$

Hence, $\mathbf{S}_{i,s}$ aggregates all synergies involving exactly s members. Organizing these terms by coalition order preserves how many order-specific states participate in each cross-scale synergy instead of collapsing all such effects into one representation. We use square-root normalization across coalition sizes and define

$$\overline{\mathbf{S}}_{i,s} = \frac{\mathbf{S}_{i,s}}{\sqrt{\binom{L_i}{s}}}, \qquad \mathcal{C}_i = \{\overline{\mathbf{S}}_{i,s}\}_{s=1}^{L_i}. \tag{10}$$

Each coalition order is read under the target condition:

$$\mathbf{g}_{i,s} = \mathrm{Read}(\overline{\mathbf{S}}_{i,s}, \mathbf{q}_i), \qquad \mathcal{Z}_i = \{\mathbf{g}_{i,s}\}_{s=1}^{L_i}. \tag{11}$$

The product-tree computation is given in Appendix B.4; the coalition mapping, contextual encodings, and zero-anchored readout are given in Appendix B.5.

3.5 Task Adaptation and Learning Objective

OSVIF preserves the order-resolved representation $\mathcal{Z}_i$ and forms the target-level aggregate $\mathbf{z}_i = \sum_{s=1}^{L_i} \mathbf{g}_{i,s}$. Let τ index one of the five evaluation tasks. A lightweight head reads the shared representations:

$$\widehat{\mathbf{y}}^{(\tau)} = \mathrm{Head}_\tau\left(\{\mathcal{Z}_i, \mathbf{z}_i\}_{i=1}^{V}\right), \qquad \mathcal{L}_\tau = \ell_\tau\left(\widehat{\mathbf{y}}^{(\tau)}, \mathbf{y}^{(\tau)}\right). \tag{12}$$

Here, $\widehat{\mathbf{y}}^{(\tau)}$ is the prediction, $\mathbf{y}^{(\tau)}$ is the ground-truth output, and ℓ_τ is the task-specific loss. The backbone and task head are trained jointly. The concrete heads and objectives for forecasting, imputation, classification, and anomaly detection are detailed in Appendix B.6.

4 Experiments

We extensively evaluate OSVIF on five mainstream time series tasks, including long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection. Across all tasks, OSVIF is compared with state-of-the-art baselines, and the results demonstrate its strong capability as a general purpose time series model. Due to space limitations, we report a summary of the main results here. Detailed experimental configurations and implementation details are provided in Appendix C.

4.1 Long-Term Forecasting

Setups. Long-term time series forecasting plays a critical role in many real-world applications. To verify the performance of our method on this task, we evaluate it over 10 real-world benchmark datasets, including four ETT datasets (ETTh1, ETTh2, ETTm1, and ETTm2) (Zhou et al., 2021), Electricity (Trindade, 2015), Traffic (Chen et al., 2001), Weather (Max Planck Institute for Biogeochemistry, n.d.), ILI (Centers for Disease Control and Prevention, n.d.), as well as Exchange-Rate and Solar-Energy(Lai et al., 2018). For the ILI dataset, we set the input look-back length to 36 and adopt prediction lengths of 24, 36, 48, and 60. For all remaining datasets, the input look-back length is 96, with prediction lengths of 96, 192, 336, and 720. For each dataset, we average the metrics across these four prediction lengths and report the resulting MSE and MAE.

Table 1: Long-term forecasting. Average MSE and MAE over four horizons; ETT(Avg) averages the four ETT subsets. Full results: Appendix Table 14. **Red**: best, Blue: second best.

Models	OSVIF (Ours)		TimeMixer (2024)		UniTS-ST (2024)		iTransformer (2024)		TimesNet (2023)		PatchTST (2023)		Crossformer (2023)		TiDE (2023)		DLinear (2023)		SCINet (2022a)		FEDformer (2022)		Stationary (2022c)		Autoformer (2021)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	0.203	**0.291**	0.250	0.344	0.331	0.403	0.242	0.327	0.199	0.298	0.216	0.318	0.244	0.334	0.251	0.344	0.212	0.300	0.268	0.365	0.214	0.327	**0.193**	0.296	0.227	0.338
ETT(Avg)	**0.389**	0.406	0.419	0.422	0.415	0.419	0.391	**0.405**	0.411	0.417	0.401	0.409	0.685	0.578	0.482	0.470	0.442	0.444	0.689	0.597	0.408	0.428	0.471	0.464	0.465	0.459
Exchange	0.382	0.412	0.462	0.451	0.418	0.440	0.368	0.409	0.445	0.459	0.367	**0.404**	0.940	0.707	0.370	0.413	**0.354**	0.414	0.750	0.626	0.519	0.429	0.461	0.461	0.613	0.539
Traffic	**0.528**	0.345	0.747	0.446	0.903	0.539	0.711	0.450	0.627	0.334	0.529	0.341	0.550	**0.304**	0.760	0.473	0.625	0.383	0.804	0.509	0.610	0.376	0.624	0.340	0.628	0.379
Weather	**0.252**	**0.277**	0.268	0.292	0.259	0.284	0.272	0.292	0.264	0.289	0.265	0.285	0.259	0.315	0.271	0.320	0.265	0.317	0.292	0.363	0.309	0.360	0.288	0.314	0.338	0.382
Solar-Energy	**0.236**	**0.266**	0.299	0.318	0.346	0.349	0.325	0.332	0.291	0.317	0.287	0.333	0.641	0.639	0.347	0.417	0.330	0.401	0.282	0.375	0.291	0.381	0.261	0.381	0.586	0.557
ILI	2.335	**0.929**	3.527	1.341	4.212	1.504	3.121	1.246	**2.323**	1.008	2.332	1.037	5.244	1.628	2.394	1.065	3.715	1.380	2.393	1.066	4.093	1.481	3.945	1.417	3.971	1.471

Table 2: Short-term forecasting. Average MAE, MAPE, and RMSE over PEMS03, PEMS04, PEMS07, and PEMS08. Full results: Appendix Table 15. **Red**: best, Blue: second best.

Models	OSVIF (Ours)	TimeMixer (2024)	iTransformer (2024)	TiDE (2023)	Crossformer (2023)	PatchTST (2023)	TimesNet (2023)	UniTS-ST (2024)	MICN (2023b)	DLinear (2023)	FEDformer (2022)	Stationary (2022c)	Autoformer (2021)
MAE	**20.09**	20.99	22.45	21.86	21.84	25.33	24.76	30.79	20.76	23.31	23.50	21.32	22.62
MAPE	**12.63**	14.35	15.07	13.80	15.51	17.72	17.46	20.98	14.38	14.68	15.01	14.09	14.89
RMSE	**31.59**	32.52	34.99	34.42	33.58	38.11	37.80	46.20	32.18	37.32	36.78	36.20	34.49

Results. Table 1 shows that OSVIF achieves the best result on **8** of the 14 dataset metric pairs. It obtains the best MSE on the ETT average, Traffic, and Solar-Energy datasets, and the best MAE on Electricity, Weather, Solar-Energy, and ILI. The largest gains are observed on Solar-Energy, where OSVIF improves the MSE by **9.6%** and the MAE by **16.1%** over the strongest competing method. On Weather, it further improves the MSE by **2.7%** and the MAE by **2.5%**. OSVIF also achieves the lowest MSE on the ETT average and Traffic, as well as the second lowest MAE on the ETT average. Although it is less competitive on Exchange and on the ILI MSE, the consistently strong results across the remaining datasets demonstrate the effectiveness of the proposed representation for diverse multivariate time series, particularly those with complex cross variable dynamics.

4.2 Short-Term Forecasting

Setups. For short-term forecasting, we use four traffic flow datasets, PEMS03, PEMS04, PEMS07, and PEMS08 from the PeMS traffic system (Chen et al., 2001). The input length is fixed to 96 and the prediction length is set to 12. Since zero and low-flow values in real traffic data can affect the stability of MAPE, we use MAE and RMSE as the main evaluation metrics and MAPE as a supplementary metric.

Results. Table 2 shows that OSVIF outperforms the compared baselines on short-term forecasting, achieving the best average MAE, MAPE, and RMSE across the four PeMS datasets. Compared with the strongest competing results, OSVIF reduces MAE and RMSE by **3.2%** and **1.8%**, respectively, and reduces MAPE by **8.5%**. OSVIF performs well on these datasets by organizing higher-order variable interactions and their cross-scale synergies for each prediction target.

4.3 Imputation

Setups. To evaluate the ability of OSVIF to recover temporal structure under random missingness, we use six datasets, including four ETT datasets (Zhou et al., 2021) (ETTh1, ETTh2, ETTm1, and ETTm2), Electricity (Trindade, 2015), and Weather (Max Planck Institute for Biogeochemistry, n.d.). The input length is set to 96, and 12.5%, 25%, 37.5%, and 50% of time points are randomly masked. The model uses only the unmasked positions as input, and MSE and MAE are computed only on the artificially masked positions.

Results. The imputation results in Table 3 shows that OSVIF consistently outperforms the compared baselines, achieving the best MAE on all six datasets and the best or tied-best MSE throughout. Compared with the corresponding second-best results, it reduces error by **8.7%** on average over the entries with strict improvements. Notably, on Electricity, OSVIF reduces MSE and MAE by **9.1%** and **8.4%**, respectively, while improving MAE by **21.8%** on Weather. These results demonstrate its robust ability to recover missing values across diverse temporal patterns. Detailed results under

Table 3: Imputation. Average MSE and MAE over masking ratios 12.5%, 25%, 37.5%, and 50%. Full results: Appendix Table 16 (two parts). **Red**: best, Blue: second best.

Models	OSVIF (Ours)		TimeMixer (2024)		UniTS-ST (2024)		TimesNet (2023)		iTransformer (2024)		PatchTST (2023)		ETSformer (2022)		LightTS (2022b)		DLinear (2023)		FEDformer (2022)		Stationary (2022c)		Autoformer (2021)		Pyraformer (2022b)		Informer (2021)		LogTrans (2019)		Reformer (2020)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	**0.025**	**0.102**	0.071	0.181	0.609	0.552	0.027	0.107	0.105	0.226	0.046	0.137	0.120	0.253	0.104	0.218	0.093	0.206	0.061	0.177	0.036	0.126	0.051	0.151	0.717	0.570	0.071	0.188	0.050	0.154	0.055	0.165
ETTm2	**0.020**	**0.078**	0.042	0.133	0.436	0.477	0.022	0.088	0.171	0.260	0.029	0.101	0.208	0.328	0.046	0.151	0.096	0.208	0.101	0.215	0.026	0.099	0.029	0.105	0.465	0.508	0.156	0.292	0.119	0.246	0.158	0.280
ETTh1	**0.080**	**0.184**	0.211	0.320	0.705	0.591	**0.080**	0.191	0.216	0.329	0.122	0.228	0.202	0.329	0.284	0.374	0.201	0.306	0.116	0.246	0.094	0.202	0.103	0.214	0.843	0.662	0.161	0.279	0.219	0.332	0.122	0.245
ETTh2	**0.048**	**0.136**	0.101	0.218	0.487	0.499	0.050	0.147	0.246	0.325	0.065	0.162	0.367	0.436	0.120	0.251	0.142	0.260	0.163	0.279	0.053	0.152	0.056	0.156	1.078	0.791	0.337	0.452	0.186	0.318	0.234	0.352
Electricity	**0.070**	**0.174**	0.189	0.302	0.669	0.652	0.095	0.212	0.118	0.246	0.077	0.190	0.214	0.339	0.131	0.262	0.132	0.261	0.130	0.260	0.100	0.218	0.101	0.225	0.296	0.382	0.222	0.328	0.174	0.303	0.200	0.314
Weather	**0.031**	**0.043**	0.060	0.105	0.176	0.278	**0.031**	0.058	0.072	0.144	0.033	0.055	0.076	0.171	0.056	0.117	0.053	0.110	0.099	0.203	0.032	0.059	**0.031**	0.057	0.152	0.234	0.045	0.105	0.039	0.076	0.038	0.087

different masking ratios are provided in Tables 16 of the Appendix, which is split across two parts for presentation.

4.4 Classification

Setups. To evaluate whether the learned representations can discriminate multivariate temporal patterns at different levels of granularity, we use 10 datasets from the UEA Time Series Classification Archive (Bagnall et al., 2018), covering diverse domains such as speech, EEG, traffic, and human activity. All methods are evaluated under the same data splits and protocol, with classification accuracy as the evaluation metric.

Table 4: Classification accuracy (%) averaged over 10 UEA datasets (Bagnall et al., 2018). Full results: Appendix Table 17. **Red**: best, Blue: second best.

Category	Classical methods			RNN			TCN	Transformers			
Method	DTW (1994)	XGBoost (2016)	ROCKET (2020)	LSTM (1997)	LSTNet (2018)	LSSL (2021)	TCN (2018)	Transformer (2017)	Reformer (2020)	Informer (2021)	Pyraformer (2022b)
Average Accuracy	67.0	66.0	72.5	48.6	71.8	70.9	70.3	71.9	71.5	72.1	70.8
Category	Transformers						MLP		CNN	General-purpose models	
Method	Autoformer (2021)	Non-stationary Transformer (2022c)	FEDformer (2022)	ETSformer (2022)	Flowformer (2022)	iTransformer (2024)	DLinear (2023)	LightTS (2022b)	TiDE (2023)	TimesNet (2023)	OSVIF (Ours)
Average Accuracy	71.1	72.7	70.7	71.0	73.0	70.5	67.5	70.4	69.5	73.6	**73.7**

Results. Table 4 shows that OSVIF achieves the highest average accuracy of **73.7%** across the 10 UEA datasets, slightly outperforming TimesNet (73.6%) and Flowformer (73.0%). Detailed per-dataset results are provided in Table 17 of the Appendix. OSVIF obtains the best accuracy on SelfRegulationSCP2 (58.9%) and achieves high accuracies on SpokenArabicDigits (98.8%), JapaneseVowels (97.3%), and SelfRegulationSCP1 (91.1%). The highest overall average suggests that OSVIF provides effective representations for multivariate time series classification.

4.5 Anomaly Detection

Setups. For anomaly detection, we use five widely adopted multivariate anomaly detection benchmarks: SMD (Su et al., 2019), MSL and SMAP (Hundman et al., 2018), SWaT (Mathur & Tippenhauer, 2016), and PSM (Abdulaal et al., 2021). We report AUROC and AUPRC because both are threshold independent: AUROC measures the overall ranking ability to separate anomalous from normal observations, whereas AUPRC places greater emphasis on the precision and recall of the anomaly class and is more informative under severe class imbalance.

Results. Table 5 shows that OSVIF achieves the highest average AUROC and AUPRC, reaching 0.6413 and 0.3023, respectively. Compared with the strongest baseline, CATCH, OSVIF improves the two metrics by 0.0461 and 0.0883. The leading average scores demonstrate the effectiveness of OSVIF for threshold free multivariate anomaly detection.

4.6 Ablations

To analyze the contribution of the main structural components in OSVIF, we conduct controlled ablations on representative tasks from three task families: long-term forecasting, imputation, and anomaly detection. These tasks cover regression-based prediction, missing-value recovery, and threshold-free anomaly ranking, providing complementary views of how the proposed interaction modules affect different output interfaces. We ablate four components: higher-order source interactions, the Cross interaction view, the natural coalition mechanism, and task-conditioned readout. All ablated variants keep the same data split, task setting, and evaluation protocol as the full model.

Table 5: Anomaly detection. Average AUROC and AUPRC are reported. Full results: Appendix Table 18. **Red**: best, Blue: second best.

Method	Avg. AUROC	Avg. AUPRC	Method	Avg. AUROC	Avg. AUPRC
OCSVM (2001)	0.5590	0.1894	**TimesNet (2023)**	0.5346	0.1798
IForest (2008)	0.5126	0.1570	**PatchTST (2023)**	0.5298	0.1764
LSTM-VAE (2020)	0.5035	0.1144	**DualTF (2024)**	0.5704	0.1802
TranAD (2022)	0.5309	0.2039	**iTransformer (2024)**	0.5198	0.1756
Anomaly Transformer (2022)	0.4892	0.1404	**UniTS-ST (2024)**	0.5392	0.1857
TFAD (2022a)	0.5000	0.1356	**TimeMixer (2024)**	0.5407	0.1805
DCdetector (2023)	0.5066	0.1368	**CATCH (2025b)**	0.5952	0.2140
D3R (2023a)	0.5550	0.1589	**OSVIF (Ours)**	**0.6413**	**0.3023**
MEMTO (2023)	0.5560	0.1511			

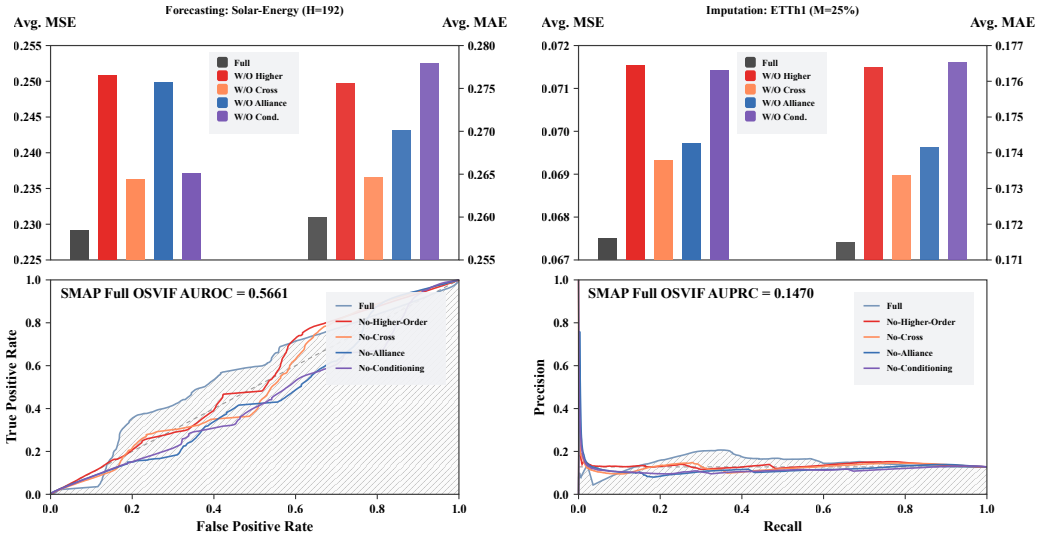

Figure 3: Ablation analysis across forecasting, imputation, and anomaly detection tasks.

Figure 3 shows the results. For long-term forecasting on Solar-Energy with $H = 192$ and imputation on ETTh1 with $M = 25\%$, removing any component increases MSE and MAE relative to the full model. For anomaly detection on SMAP, the full model also encloses larger ROC and PR areas than the ablated variants, indicating better threshold-free ranking performance. The degradation patterns differ across tasks and metrics, suggesting that the relative importance of each component depends on the target task and data characteristics. Nevertheless, the consistent drop after removing higher-order interactions, Cross-view modeling, coalition modeling, or conditioning supports the view that OSVIF's performance comes from their joint effect, with no single module being dispensable.

5 Conclusion

In this paper, we introduced OSVIF, a target-centered framework for general multivariate time series analysis. OSVIF organizes source-variable combinations into a natural order spectrum and further captures cross-order synergies through a natural coalition spectrum. Both spectra are implemented by generating-polynomial constructions, which cover the relevant combination spaces without explicit subset enumeration or predefined group structures. With target conditioning, Identity and Cross interaction views, and lightweight task-specific heads, the same backbone supports long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection. Extensive experiments across these task families show that OSVIF delivers strong and consistent results, while ablation and representation analyses verify the benefits of higher-order interactions, cross-view modeling, coalition modeling, and task conditioning. These findings suggest that preserving order-specific information and modeling its synergies provide an effective principle for multivariate time series modeling. Future work will improve efficiency for higher-dimensional variables and longer contexts, extend OSVIF toward pretraining and transfer.

AI USE STATEMENT

In this work, we use AI tools to assist with language polishing and translation. The tools were not used to generate experimental data or independently determine scientific claims. All AI-assisted content and edits were reviewed and verified by us. We take full responsibility for the final content of this work, including any text or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

For reproducibility, please refer to the material.

REFERENCES

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2485–2494, 2021.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, pp. 359–370, 1994.

Mathieu Blondel, Akinori Fujino, Naonori Ueda, and Masakazu Ishihata. Higher-order factorization machines. *Advances in Neural Information Processing Systems*, 29, 2016.

Centers for Disease Control and Prevention. Influenza-like illness (ILI) data, n.d. URL `https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`.

Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. Freeway performance measurement system: mining loop detector data. *Transportation Research Record*, 1748 (1):96–102, 2001.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with TiDE: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, 2006.

Angus Dempster, François Petitjean, and Geoffrey I Webb. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020.

David K Duvenaud, Hannes Nickisch, and Carl Rasmussen. Additive Gaussian processes. *Advances in Neural Information Processing Systems*, 24, 2011.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, 34:572–585, 2021.

Hui He, Qi Zhang, Simeng Bai, Kun Yi, and Zhendong Niu. CATN: Cross attentive tree-aware network for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4030–4038, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 387–395, 2018.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104, 2018.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in Neural Information Processing Systems*, 32, 2019.

Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. xDeepFM: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1754–1763, 2018.

Shuyu Lin, Ronald Clark, Robert Birke, Sandro Schonborn, Niki Trigoni, and Stephen Roberts. Anomaly detection for time series using VAE-LSTM hybrid model. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4327–4331. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053558.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the 2008 IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. SCINet: Time series modeling and forecasting with sample convolution and interaction. *Advances in Neural Information Processing Systems*, 35:5816–5828, 2022a.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2022b.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35:9881–9893, 2022c.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Aditya P Mathur and Nils Ole Tippenhauer. SWaT: A water treatment testbed for research and training on ICS security. In *2016 International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater)*, pp. 31–36. IEEE, 2016.

Max Planck Institute for Biogeochemistry. Weather, n.d. URL `https://www.bgc-jena.mpg.de/wetter/`.

Ryan L Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In *International Conference on Learning Representations*, 2019.

Youngeun Nam, Susik Yoon, Yooju Shin, Minyoung Bae, Hwanjun Song, Jae Gil Lee, and Byung Suk Lee. Breaking the time-frequency granularity discrepancy in time-series anomaly detection. In *Proceedings of the ACM Web Conference 2024*, pp. 4204–4215, 2024.

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

Chuanru Ren, Yao Lu, Tianjin Huang, Haowen Zheng, Hengde Zhu, Yunyin Li, Hengxiao Li, and Lu Liu. TimeCAP: A channel-aware pre-training framework for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25108–25116, 2026.

Bernhard Scholkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. MEMTO: Memory-guided transformer for multivariate time series anomaly detection. *arXiv preprint arXiv:2312.02530*, 2023.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2828–2837, 2019.

Changyuan Tian, Zhicong Lu, Zequn Zhang, Heming Yang, Wei Cao, Zhi Guo, Xian Sun, and Li Jin. HyperMixer: Specializable hypergraph channel mixing for long-term multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20885–20893, 2025.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6): 1201–1214, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Chengsen Wang, Zirui Zhuang, Qi Qi, Jingyu Wang, Xingyu Wang, Haifeng Sun, and Jianxin Liao. Drift does not matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. doi: 10.52202/075280-0473.

Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. MICN: Multi-scale local and global context modeling for long-term series forecasting. In *International Conference on Learning Representations*, 2023b.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. ETSformer: Exponential smoothing transformers for time-series forecasting. *arXiv preprint arXiv:2202.01381*, 2022.

Binqing Wu, Zongjiang Shang, Jianlong Huang, and Ling Chen. MillGNN: Learning multi-scale lead-lag dependencies for multi-variate time series forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 3344–3354, 2025a.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. In *International Conference on Machine Learning*, pp. 24226–24242, 2022.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *International Conference on Learning Representations*, 2025b.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 753–763, 2020.

Ziyi Xiao, Cong Luo, Jiajia Hu, Guodong Sa, and Yueyang Wang. Exploring dual-view graph structures: Contrastive learning with graph and hypergraph for multivariate time series classification. *Neural Networks*, 192:107859, 2025.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Rundong Xue, Hao Hu, Zhitao Zeng, Xiangmin Han, Zhiqiang Tian, Shaoyi Du, and Yue Gao. Role hypergraph contrastive learning for multivariate time-series analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27468–27476, 2026.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3033–3045, 2023.

Zenan Ying, Zhi Zheng, Huijun Hou, Tong Xu, Qi Liu, Jinke Wang, and Wei Chen. SGN: Shifted window-based hierarchical variable grouping for multivariate time series classification. *Advances in Neural Information Processing Systems*, 38:119820–119851, 2025.

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

Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. TFAD: A decomposition time series anomaly detection architecture with time-frequency analysis. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 2497–2507, 2022a.

Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented MLP structures. *arXiv preprint arXiv:2207.01186*, 2022b.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pp. 27268–27286. PMLR, 2022.

A Detailed Related Work

Multivariate time series often exhibit cross-variable dependencies, which existing methods handle using different strategies. One line of work adopts a channel-independent approach. For example, PatchTST (Nie et al., 2023) processes each channel separately using shared parameters. Another line integrates information across variables through attention or graph propagation. Crossformer (Zhang & Yan, 2023) uses learnable routers to aggregate and distribute information during cross-variable attention. iTransformer (Liu et al., 2024) encodes the historical sequence of each variable as a token and models cross-variable dependencies through self-attention among these tokens. MTGNN (Wu et al., 2020) learns directed relations between variables and integrates information from related variables through multi-hop graph propagation. Through aggregation and nonlinear transformations, these interaction-based methods can incorporate information from combinations of multiple variables into variable representations. However, they mix information from combinations of different sizes within a single representation for each variable, without preserving separate representations by combination size. When combinations of different sizes play different roles for a target variable, this organization makes it difficult to selectively use information at each size for that target. This limitation may hinder the modeling of cross-variable dependencies and affect performance on downstream time series analysis tasks.

To separate information from different combinations, group-based methods construct distinct representations for selected variable combinations. HyperMixer (Tian et al., 2025) learns associations between variables and hyperedges to organize variables into groups, with each hyperedge representation encoding information from the corresponding combination. DVG-CL (Xiao et al., 2025) uses both graph and hypergraph views. The graph view represents variables as nodes connected by pairwise relations, whereas the hypergraph view groups neighboring variables into hyperedges to represent their combinations. RHCL (Xue et al., 2026) constructs hyperedges from related variables based on sparse representation relations within each time window. It then forms role-specific combination representations according to the cooperative and suppressive roles of variables within each hyperedge. These methods enrich the representation of cross-variable dependencies by organizing variable combinations through hyperedges. However, the combinations represented separately remain limited by the constructed group structures. Information from relevant combinations that lack separate representations must still be captured through subsequent propagation and fusion, which mix this information into aggregated variable representations and make it difficult to preserve distinctly. The coverage of the group structures therefore constrains the extent to which relevant combination information can be captured and selectively used for a target variable.

The effect of a variable combination on a target variable may depend on the information provided by other combinations, motivating the modeling of synergy among combinations. Building on combination representations, existing studies explore joint modeling across combinations. One line of work relies on hierarchical group relations. CATN (He et al., 2022) uses pairwise cross-attention to mutually adjust combination representations at selected tree levels and then fuses the resulting information. MillGNN (Wu et al., 2025a) aggregates information from combinations with lead–lag relations at each grouping level, then updates variable representations using messages from multiple levels. In these methods, both the combinations directly involved in joint modeling and the information flow among them are constrained by the constructed hierarchies. Another line of work jointly models combinations using feature transformations and information exchange between groups. SGN (Ying et al., 2025) applies pointwise convolutions across groups to jointly transform representations of different variable combinations. TimeCAP (Ren et al., 2026) uses variable representations as cross-attention queries to jointly retrieve information from router representations that encode variable combinations. Although these methods jointly use information from multiple combinations, the combinations available as distinct units for joint modeling remain limited by the constructed group structures. If relevant combinations of different sizes are not directly included, their synergy may be insufficiently captured. This limitation can hinder the use of information across sizes to assess the state of a target variable and changes in that state.

Broadening the coverage of variable combinations and modeling their synergy require addressing the computational challenge posed by the rapidly growing number of possible combinations. Related research, including work on general feature interaction learning, offers several approaches to controlling this cost. One line of work exploits the algebraic structure of combination terms to avoid explicit enumeration. Additive Gaussian processes (Duvenaud et al., 2011) construct interac-

tion kernels with orders defined by the number of features in each combination and evaluate these kernels efficiently through recurrences for elementary symmetric polynomials. Higher-order factorization machines (Blondel et al., 2016) represent feature combinations using factorized parameters and compute interaction terms of each order through ANOVA kernels and dynamic programming. Another line of work controls computational cost by restricting combinations or compressing intermediate representations. k-ary Janossy Pooling (Murphy et al., 2019) limits the number of members in each combination and applies symmetrization to obtain representations invariant to permutations of those members. xDeepFM (Lian et al., 2018) uses a compressed interaction network that multiplies representations from the preceding layer with the base input representations and compresses the results to progressively construct higher-order feature interactions. These approaches reduce the cost of combination modeling through algebraic computation, restrictions on combinations, or representation compression. In particular, algebraic recurrences provide the computational foundation for our method to cover all combinations without explicit enumeration while efficiently constructing representations organized by combination size.

Together, these studies inform the representation of variable combinations, joint modeling across combinations, and efficient computation. To address the mixing of combination information and structural constraints on combination coverage, OSVIF introduces a natural order spectrum that covers all combinations of source variables and preserves separate representations by combination size. It forms target-conditioned states that enable selective use of information at each size for a given target variable. To extend the scope of synergy modeling beyond these structural constraints, OSVIF further introduces a natural coalition spectrum that models all combinations drawn from the order-specific states and the state of the target variable itself. This spectrum enables a more complete characterization of how information across sizes jointly affects the target variable. Both spectra use generating polynomials to avoid explicit enumeration, controlling computational cost while fully covering their respective combination spaces. Together, they provide a general representation framework for five time series analysis tasks: long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection.

B Method Details

This appendix details how the two spectra realize the target-centered modeling objectives in Section 3. The natural order pipeline retains source-variable combination information by size and conditions each state on the target variable. The natural coalition pipeline combines these states with the target direct state to represent cross-scale synergies by coalition order. Product trees cover both combination spaces without explicit enumeration.

B.1 Exact Computation of the Natural Order Spectrum

Section 3.2 defines $P_i(z) = \prod_{j\in\mathcal{A}_i}^{\odot}(\mathbf{1}_d + \mathbf{v}_j z)$. Computing this product independently for every target would repeat almost the same multiplication V times because $\mathcal{A}_i = [V] \setminus \{i\}$. OSVIF instead uses a shared complement product tree to obtain all leave-one-out polynomials with one upward and one downward pass.

Each leaf stores $(\mathbf{1}_d + \mathbf{v}_j z)$. During the upward pass, adjacent subtree polynomials are merged by exact coordinate-wise polynomial convolution, truncated only at the largest degree valid for the current subtree. The root therefore stores $\prod_{j=1}^{V}(\mathbf{1}_d + \mathbf{v}_j z)$. During the downward pass, each node receives the product of all factors outside its subtree. At leaf i, this exterior product is exactly $P_i(z)$. Dividing its coefficients by $\binom{n_i}{m}$ yields $\overline{\mathbf{E}}_{i,m}$ for all targets without using the target index as an outer loop. This preserves the coefficient semantics and full combination coverage while avoiding repeated computation across targets.

B.2 Derivation of the Identity and Cross Views

For a source combination $S \subseteq \mathcal{A}_i$, $|S| = m$, its product after projection onto $\mathbf{f}_r$ is

$$\prod_{j\in S} p_{j,r} = \prod_{j\in S}\left(\sum_{k_j=1}^{d} v_{j,k_j} f_{r,k_j}\right) = \sum_{\{k_j\}_{j\in S}} \prod_{j\in S} v_{j,k_j} f_{r,k_j}. \tag{13}$$

Terms with $k_j = k$ for every $j \in S$ describe within-coordinate interactions. All remaining terms, for which at least two sources select different coordinates, describe cross-coordinate interactions. Averaging over the $\binom{n_i}{m}$ combinations preserves this additive separation:

$$\overline{e}_{i,m,r} = d_{i,m,r} + c_{i,m,r}. \tag{14}$$

The Hadamard coefficient $\overline{\mathbf{E}}_{i,m}$ contains precisely the within-coordinate products. Its contribution after projection is

$$d_{i,m,r} = \langle \overline{\mathbf{E}}_{i,m}, \mathbf{f}_r^{\odot m} \rangle = \sum_{k=1}^{d} (\overline{\mathbf{E}}_{i,m})_k f_{r,k}^m. \tag{15}$$

Subtracting this quantity from the full projected interaction isolates $c_{i,m,r}$, which is mapped back through $\mathbf{f}_r$ and summed across directions to construct $\mathbf{C}_{i,m}$. When $m = 1$, every term contains only one coordinate choice, so $c_{i,1,r} = 0$. Identity and Cross are complementary structural views; they are not asserted to be a strict additive decomposition in the original coordinate space. Keeping the two views distinct within each natural order avoids mixing sizes and allows target conditioning to use their contributions differently.

B.3 Target-Conditioned Fusion

Section 3.3 defines $\mathbf{u}_{i,m} = \Phi(\mathbf{I}_{i,m}, \mathbf{C}_{i,m}, \mathbf{q}_i)$. For $w \in \{I, C\}$, first apply zero-preserving RMS normalization:

$$\widetilde{\mathbf{X}}^I_{i,m} = \frac{\mathbf{I}_{i,m}}{\sqrt{\mathrm{mean}(\mathbf{I}^{\odot 2}_{i,m}) + \epsilon}}, \qquad \widetilde{\mathbf{X}}^C_{i,m} = \frac{\mathbf{C}_{i,m}}{\sqrt{\mathrm{mean}(\mathbf{C}^{\odot 2}_{i,m}) + \epsilon}}, \tag{16}$$

where $\epsilon = 10^{-6}$. This operation maps an all-zero view to an all-zero view.

To share the fusion network across orders, let $\xi = m/\max(n_i, 1)$ and define the continuous order encoding

$$\boldsymbol{\rho}_m = [\xi, \sin\xi, \cos\xi, \sin(2\xi)]. \tag{17}$$

For view identifier $\mathbf{c}_w$, the zero-anchored hidden mapping is

$$\mathbf{h}^w_{i,m} = \sigma\Big(W_h[\mathbf{q}_i, \widetilde{\mathbf{X}}^w_{i,m}, \boldsymbol{\rho}_m, \mathbf{c}_w] + b_h\Big) - \sigma(W_h[\mathbf{q}_i, \mathbf{0}, \boldsymbol{\rho}_m, \mathbf{c}_w] + b_h)\,, \tag{18}$$

where σ is GELU. Subtracting the zero-input response ensures that absent evidence produces no hidden response.

For $m \geq 2$, a softmax gate combines the two views:

$$(\alpha^I_{i,m}, \alpha^C_{i,m}) = \mathrm{Softmax}\left(g(\mathbf{q}_i, \boldsymbol{\rho}_m, \mathbf{h}^I_{i,m}, \mathbf{h}^C_{i,m})\right), \tag{19}$$

$$\mathbf{u}_{i,m} = W_J\left[\alpha^I_{i,m}\mathbf{h}^I_{i,m}; \alpha^C_{i,m}\mathbf{h}^C_{i,m}\right]. \tag{20}$$

At $m = 1$, the Cross view is zero and $\mathbf{u}_{i,1} = \mathbf{h}^I_{i,1}$. For the target direct state, an adapter first maps $\mathbf{h}_{i,0}$ to the same view space as $\mathbf{I}_{i,0} = \mathrm{Adapt}(\mathbf{h}_{i,0})$. The fusion then uses $\boldsymbol{\rho}_0 = \mathbf{0}$ and the Identity branch, giving $\mathbf{u}_{i,0} = \mathbf{h}^I_{i,0}$. Conditioning on $\mathbf{q}_i$ allows the same order information to be used differently across targets. The order encoding preserves combination size, and zero anchoring prevents absent views from producing spurious evidence.

B.4 Exact Computation of the Natural Coalition Spectrum

Section 3.4 defines

$$Q_i(z) = \prod_{m \in \mathcal{M}_i}^{\odot} (\mathbf{1}_{d_a} + \mathbf{a}_{i,m} z). \tag{21}$$

Each member $m \in \mathcal{M}_i$ forms a leaf polynomial $(\mathbf{1}_{d_a} + \mathbf{a}_{i,m}z)$. Adjacent subtree polynomials are merged by exact coordinate-wise convolution, retaining every degree valid for the subtree. A single upward pass therefore returns all coefficients $\mathbf{S}_{i,0}, \ldots, \mathbf{S}_{i,L_i}$ at the root. Unlike the natural order spectrum, no leave-one-out operation is required. The nonempty coefficients are divided by $\sqrt{\binom{L_i}{s}}$ to obtain $\overline{\mathbf{S}}_{i,s}$. The product tree therefore covers the full synergy space without enumerating the 2^{L_i} member combinations or imposing a predefined group structure.

B.5 Coalition Mapping and Readout

The coalition layer uses the four-dimensional trigonometric encoding

$$\psi(v) = [\sin(\pi v), \cos(\pi v), \sin(2\pi v), \cos(2\pi v)]. \tag{22}$$

This encoding is distinct from $\boldsymbol{\rho}_m$: it supplies the coalition layer with member-order, coalition-order, and valid-order-count contexts.

The complete coalition mapping is

$$\mathbf{a}_{i,m} = \phi_A\left(\mathbf{u}_{i,m}, \mathbf{q}_i, \psi(m/\max(n_i, 1))\right) \in \mathbb{R}^{d_a}. \tag{23}$$

It is implemented with the same zero-anchoring principle as the natural-order fusion: the network response at $\mathbf{u}_{i,m} = \mathbf{0}$ is subtracted so that an absent member cannot create coalition evidence.

For readout, define

$$\boldsymbol{\delta}_s = \psi(s/L_i), \qquad \boldsymbol{\kappa}_i = \psi(L_i/L_{\max}), \tag{24}$$

where $L_{\max}$ is the maximum number of members. The complete readout is

$$\mathbf{g}_{i,s} = \mathrm{Read}\left(\overline{\mathbf{S}}_{i,s}, \mathbf{q}_i, \boldsymbol{\delta}_s, \boldsymbol{\kappa}_i\right). \tag{25}$$

$\boldsymbol{\delta}_s$ identifies the current coalition order, whereas $\boldsymbol{\kappa}_i$ records how many coalition orders are valid for the target. The readout is zero-anchored with respect to $\overline{\mathbf{S}}_{i,s}$, so zero synergy evidence produces $\mathbf{g}_{i,s} = \mathbf{0}$.

B.6 Task-Specific Heads and Objectives

The generic head in Section 3.5 is instantiated as follows. Batch index b is restored in the losses below.

Forecasting. The forecasting head applies a shared bias-free linear map to $\mathbf{z}_i$, with output dimension equal to the prediction horizon H. Inputs are normalized per variable by a reversible mean–standard-deviation transformation, and predictions are mapped back to the original scale. The objective is mean squared error over valid forecast positions:

$$\mathcal{L}_{\text{for}} = \frac{1}{|\Omega|} \sum_{(b,t,i)\in\Omega} (\widehat{Y}_{b,t,i} - Y_{b,t,i})^2. \tag{26}$$

Imputation. The imputation head maps $\mathbf{z}_i$ linearly to the reconstruction at each target time. During training, observed entries are randomly masked and the loss is evaluated only on these artificial omissions:

$$\mathcal{L}_{\text{imp}} = \frac{1}{|\Omega_{\text{mask}}|} \sum_{(b,t,i)\in\Omega_{\text{mask}}} (\widehat{X}_{b,t,i} - X_{b,t,i})^2. \tag{27}$$

Classification. The classification head uses both $\mathbf{z}_i$ and $\mathcal{Z}_i$. A backbone output layer maps coalition-order states to order-wise evidence, sums over orders, and averages over targets to obtain token-level logits. Two residual terms refine these logits: target-query pooling adaptively aggregates targets from $\mathcal{Z}_i$, and a GRU models the temporal order of the aggregated token states. The sample logits are the sum of the time-averaged token logits and both residuals. Training uses

$$\mathcal{L}_{\text{cls}} = \mathrm{CrossEntropy}(\widehat{\mathbf{y}}, y). \tag{28}$$

Anomaly detection. The anomaly head has Value and Relation branches. The Value branch predicts the current observation from $\mathbf{z}_i$. The Relation branch obtains an order-wise logit $r_{b,i,s}$ from each $\mathbf{g}_{i,s}$. Let $\mu_{b,i,s} \in \{0, 1\}$ indicate whether coalition order s is valid for target i in sample b. The Relation branch aggregates valid orders as

$$r_{b,i} = \frac{\sum_s \mu_{b,i,s} r_{b,i,s}}{\sqrt{\max(1, \sum_s \mu_{b,i,s})}}. \tag{29}$$

The Value branch uses the Huber loss. The Relation branch contrasts matched samples r^+ with samples r^- formed by shifting the batch cyclically by half a batch. Let Ω_{value} be the set of valid observation indices (b, t, i), and let $\Omega_{\text{rel}} = \{(b, i) : \sum_s \mu_{b,i,s} > 0\}$ be the set of valid relation indices:

$$\mathcal{L}_{\text{value}} = \text{mean}_{(b,t,i)\in\Omega_{\text{value}}} \text{Huber}(\widehat{X}_{b,t,i}, X_{b,t,i}), \tag{30}$$

$$\mathcal{L}_{\text{rel}} = \frac{1}{2}\left[\text{mean}_{(b,i)\in\Omega_{\text{rel}}} \text{softplus}(-r^+_{b,i}) + \text{mean}_{(b,i)\in\Omega_{\text{rel}}} \text{softplus}(r^-_{b,i})\right], \tag{31}$$

$$\mathcal{L}_{\text{ano}} = \frac{1}{2}(\mathcal{L}_{\text{value}} + \mathcal{L}_{\text{rel}}). \tag{32}$$

C Implementation Details

Dataset details. Table 6 summarizes the datasets used by OSVIF. T, H, and M denote the input sequence length, prediction length, and random masking ratio, respectively.

Table 6: Dataset descriptions. The dataset size is organized in (Train, Validation, Test).

Tasks	Dataset	Dim	Series Length	Data Size	Information (Frequency)
Long-term forecasting	ETTh1	7	T=96; H∈{96,192,336,720}	(8640, 2880, 2880)	Electricity transformer (Hourly)
	ETTh2			(8640, 2880, 2880)	Electricity transformer (Hourly)
	ETTm1	7	T=96; H∈{96,192,336,720}	(34560, 11520, 11520)	Electricity transformer (15 mins)
	ETTm2			(34560, 11520, 11520)	Electricity transformer (15 mins)
	Electricity	321		(18412, 2631, 5261)	Electricity load (Hourly)
	Traffic	862		(12280, 1755, 3509)	Road occupancy (Hourly)
	Weather	21	T=96; H∈{96,192,336,720}	(36887, 5269, 10540)	Weather (10 mins)
	Exchange	8		(5311, 759, 1518)	Exchange rate (Daily)
	Solar-Energy	137		(36792, 5256, 10512)	Solar power (10 mins)
	ILI	7	T=36; H∈{24,36,48,60}	(676, 96, 194)	Influenza-like illness (Weekly)
Short-term forecasting	PEMS03	358	T=96; H=12	(15724, 5242, 5242)	Traffic flow (5 mins)
	PEMS04	307		(10195, 3398, 3399)	Traffic flow (5 mins)
	PEMS07	883		(16934, 5645, 5645)	Traffic flow (5 mins)
	PEMS08	170		(10713, 3571, 3572)	Traffic flow (5 mins)
Imputation	ETTh1	7	T=96; M∈{12.5%,25%,37.5%,50%}	(8640, 2880, 2880)	Electricity transformer (Hourly)
	ETTh2			(8640, 2880, 2880)	Electricity transformer (Hourly)
	ETTm1			(34560, 11520, 11520)	Electricity transformer (15 mins)
	ETTm2			(34560, 11520, 11520)	Electricity transformer (15 mins)
	Electricity	321		(18412, 2631, 5261)	Electricity load (Hourly)
	Weather	21		(36887, 5269, 10540)	Weather (10 mins)
Classification (UEA)	EthanolConcentration	3	1751	(209, 52, 263)	Spectroscopy (4 classes)
	FaceDetection	144	62	(4712, 1178, 3524)	MEG face detection (2 classes)
	Handwriting	3	152	(118, 32, 850)	Handwriting trajectory (26 classes)
	Heartbeat	61	405	(164, 40, 205)	Heart sound record (2 classes)
	JapaneseVowels	12	29	(216, 54, 370)	Speech record (9 classes)
	PEMS-SF	963	144	(214, 53, 173)	Traffic pattern (7 classes)
	SelfRegulationSCP1	6	896	(214, 54, 293)	EEG self-regulation (2 classes)
	SelfRegulationSCP2	7	1152	(160, 40, 180)	EEG self-regulation (2 classes)
	SpokenArabicDigits	13	93	(5279, 1320, 2199)	Speech digits (10 classes)
	UWaveGestureLibrary	3	315	(96, 24, 320)	Motion gesture (8 classes)
Anomaly detection	MSL	55	T=100	(46653, 11664, 73729)	Spacecraft telemetry
	PSM	25		(105984, 26497, 87841)	Server monitoring
	SMAP			(108146, 27037, 427617)	Spacecraft telemetry
	SMD	38		(566724, 141681, 708420)	Server machines
	SWaT	51		(396000, 99000, 449919)	Industrial water treatment

Experimental details. Table 7 summarizes the OSVIF settings needed to reproduce the main experiments, including the shared model capacity, optimizer, learning rate, loss family, effective batch size, maximum training budget, and early-stopping patience. The loss entries are compact task-level labels; the corresponding heads and objective formulas are defined in Appendix B.6.

All deep learning models are implemented in PyTorch and trained on six NVIDIA A40 48G GPUs. Training configurations, data splits, model capacities, and evaluation rules are fixed before testing, and the test set is not used for model configuration or hyperparameter tuning. Across tasks, we use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, without AMSGrad. The default weight decay is 10^{-4}, the global gradient norm is clipped at 5.0. `Microbatch` only splits physical batches under memory constraints and does not change the logical effective batch size reported in Table 7.

Model capacity. M denotes the maximum natural source order, R denotes the number of cross-hidden-coordinate directions in the Cross branch, and K denotes the number of coalition groups used by the natural coalition readout. In the main experiments, no upper limit is imposed on the natural source order, so $M = V - 1$ and all source variables are used for each target. All tasks use $R = 16,\ K = 8$ to balance expressiveness and computational cost and to avoid dataset-specific

Table 7: Experiment configurations across tasks.

Tasks / Configurations	Model Hyper-parameter			Training Process					
	Natural order limit M	Cross rank R	Alliance groups K	Optimizer	LR	Loss	Effective batch	Max epochs	Patience
Long-term forecasting	$V-1$	16	8	AdamW	1×10^{-3}	MSE ($\mathcal{L}_{\text{for}}$)	128 auto microbatch	120	15
Short-term forecasting	$V-1$	16	8	AdamW	1×10^{-3}	MSE ($\mathcal{L}_{\text{for}}$)	128 auto microbatch	120	15
Imputation	$V-1$	16	8	AdamW	1×10^{-3}	masked MSE ($\mathcal{L}_{\text{imp}}$)	128 auto microbatch	120	15
Classification	$V-1$	16	8	AdamW	1×10^{-3}	cross-entropy ($\mathcal{L}_{\text{cls}}$)	128 auto microbatch	120	15
Anomaly detection	$V-1$	16	8	AdamW	3×10^{-4}	Value + relation ($\mathcal{L}_{\text{ano}}$)	128 auto microbatch	120	15

** LR denotes the initial learning rate.*

capacity tuning that would affect cross-task comparability. This configuration is fixed before testing and it does not represent the individually optimal setting for every dataset.

Baseline details. OSVIF is evaluated as a unified cross-variable interaction backbone for five tasks: long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection. The baseline description follows the task structure of the experiments and records the variable-information pathway used by each comparison method. The comparison includes channel-independent temporal models, dense or implicit feature-mixing models, multiscale and decomposition models, explicit cross-variable attention models, task-specific models, and task-general time-series backbones. Published results are used under matching data splits, input and output configurations, masking ratios, evaluation metrics, and result-aggregation procedures. Locally reproduced methods use the same preprocessing, training, validation, test splits, and metric-computation rules as the OSVIF experiments.

Long-term forecasting. The long-term forecasting table contains channel-independent, temporal-backbone, explicit cross-variable, and task-general methods. PatchTST applies patch-based temporal encoding to each channel with shared parameters, and DLinear fits a linear temporal mapping for each channel (Nie et al., 2023; Zeng et al., 2023). TiDE uses an MLP encoder–decoder with dense feature projections for historical observations and covariates (Das et al., 2023). Autoformer combines series decomposition with autocorrelation-based temporal aggregation; FEDformer combines decomposition with frequency-enhanced attention; Non-stationary Transformer uses stationarization together with de-stationary attention; and SCINet uses recursive multi-scale convolution and interaction blocks (Wu et al., 2021; Zhou et al., 2022; Liu et al., 2022c;a). TimesNet maps period-specific temporal variations to two-dimensional representations and processes them with shared convolutional transformations, while TimeMixer mixes trend and seasonal representations across temporal scales (Wu et al., 2023; Wang et al., 2024). Crossformer uses dimension–segment-wise embeddings and router-mediated cross-dimension attention, and iTransformer represents each variable history as a variate token followed by attention across variate tokens (Zhang & Yan, 2023; Liu et al., 2024). UniTS-ST supplies a task-general backbone with attention across time and variable dimensions and task-adaptive interfaces (Gao et al., 2024). These methods provide distinct routes from multivariate inputs to target predictions, including separate channel processing, shared temporal representations, explicit variable attention, and unified task adaptation.

Short-term forecasting. The short-term forecasting comparison uses the same principal variable-modeling families under the PeMS traffic protocol. TimeMixer, TimesNet, Autoformer, FEDformer, and Non-stationary Transformer provide decomposition-, period-, multiscale-, and attention-based temporal representations (Wang et al., 2024; Wu et al., 2023; 2021; Zhou et al., 2022; Liu et al., 2022c). PatchTST and DLinear provide channel-independent references, while TiDE provides dense MLP feature projections (Nie et al., 2023; Zeng et al., 2023; Das et al., 2023). Crossformer and iTransformer explicitly exchange information across dimensions through router-mediated cross-dimension attention and variate-token attention, respectively (Zhang & Yan, 2023; Liu et al., 2024). MICN uses local and global convolutional branches with multi-scale temporal decomposition, and UniTS-ST uses a shared attention backbone with task-specific interfaces (Wang et al., 2023b; Gao et al., 2024). The short-term setting therefore evaluates variable processing together with local traffic dynamics and short prediction horizons.

Imputation. The imputation comparison evaluates multivariate reconstruction under the same input length, masking ratios, masked-position metrics, and data splits. TimeMixer, TimesNet, Autoformer,

FEDformer, and Non-stationary Transformer encode temporal structure through multiscale mixing, period-aware transformations, decomposition, frequency-enhanced attention, and stationarity-aware attention, respectively (Wang et al., 2024; Wu et al., 2023; 2021; Zhou et al., 2022; Liu et al., 2022c). PatchTST and DLinear process channels independently with shared parameters, while iTransformer and UniTS-ST use variable tokens or joint time–variable attention for multivariate representation learning (Nie et al., 2023; Zeng et al., 2023; Liu et al., 2024; Gao et al., 2024). ETSformer represents temporal dynamics with exponential-smoothing components and attention; LightTS uses lightweight sampling and MLP transformations; and the attention baselines Pyraformer, Informer, LogTrans, and Reformer use pyramidal, probabilistically sparse, logarithmically sparse, and locality-sensitive attention patterns, respectively (Woo et al., 2022; Zhang et al., 2022b; Liu et al., 2022b; Zhou et al., 2021; Li et al., 2019; Kitaev et al., 2020). The imputation group consequently covers channel-wise reconstruction, shared temporal feature extraction, explicit variable interaction, and sparse or hierarchical temporal attention.

Classification. The classification table includes methods with different input representations and variable-processing paths. DTW compares multivariate sequences through distance alignment, XGBoost operates on engineered or flattened features through tree ensembles, and ROCKET uses randomized convolutional features followed by a linear classifier (Berndt & Clifford, 1994; Chen & Guestrin, 2016; Dempster et al., 2020). LSTM, LSTNet, and LSSL encode sequential dependencies with recurrent or state-space operations, and TCN uses dilated temporal convolutions (Hochreiter & Schmidhuber, 1997; Lai et al., 2018; Gu et al., 2021; Bai et al., 2018). Transformer, Reformer, Informer, and Pyraformer use self-attention with dense, locality-sensitive, probabilistically sparse, and pyramidal temporal patterns (Vaswani et al., 2017; Kitaev et al., 2020; Zhou et al., 2021; Liu et al., 2022b). Autoformer, Non-stationary Transformer, FEDformer, ETSformer, and Flowformer introduce decomposition, stationarity adjustment, frequency enhancement, exponential-smoothing components, or flow-based attention into the temporal representation (Wu et al., 2021; Liu et al., 2022c; Zhou et al., 2022; Woo et al., 2022; Wu et al., 2022). iTransformer performs attention over variate tokens; DLinear and LightTS use linear or lightweight MLP transformations; TiDE uses an MLP encoder–decoder; and TimesNet uses period-aware two-dimensional temporal transformations (Liu et al., 2024; Zeng et al., 2023; Zhang et al., 2022b; Das et al., 2023; Wu et al., 2023). OSVIF uses the same target-centered variable-interaction backbone as in the other tasks and attaches the classification head defined in Appendix B.6. The comparison therefore includes classical multivariate representations, recurrent and convolutional encoders, temporal-attention variants, lightweight linear or MLP models, and general-purpose time-series backbones.

Anomaly detection. The anomaly-detection table combines classical detectors, reconstruction and representation models, and general-purpose time-series backbones. OCSVM and IForest score multivariate observations through kernel-boundary and isolation-based representations (Scholkopf et al., 2001; Liu et al., 2008). LSTM-VAE models temporal windows with recurrent variational encoding and reconstruction, TranAD uses a Transformer-based reconstruction and adversarial training procedure, and Anomaly Transformer uses association discrepancies between temporal positions (Lin et al., 2020; Tuli et al., 2022; Xu et al., 2022). TFAD uses temporal feature extraction for anomaly scoring; DCdetector uses dual-view contrastive representations; D3R and MEMTO use reconstruction or memory-guided representations; DualTF combines time- and frequency-domain representations; and CATCH uses frequency patching with a channel-fusion module for patch-wise channel correlations (Zhang et al., 2022a; Yang et al., 2023; Wang et al., 2023a; Song et al., 2023; Nam et al., 2024; Wu et al., 2025b). TimesNet, PatchTST, iTransformer, UniTS-ST, and TimeMixer transfer their temporal and variable-processing backbones to anomaly scoring (Wu et al., 2023; Nie et al., 2023; Liu et al., 2024; Gao et al., 2024; Wang et al., 2024). All methods use the five anomaly-detection datasets, common data splits and preprocessing, continuous score aggregation, and the same AUROC/AUPRC implementations. The anomaly score is oriented so that larger values indicate greater anomaly likelihood; task-specific detectors retain their native scoring interfaces, and the general-purpose backbones use the task adaptation described in the experimental configuration.

Across the five tasks, OSVIF keeps the natural order spectrum, natural coalition spectrum, and target-conditioned interaction construction in the shared backbone, with a task-specific output head for each evaluation interface. The resulting comparison places OSVIF alongside channel-independent processing, implicit feature mixing, temporal decomposition, multiscale modeling, explicit cross-variable attention, sparse attention, reconstruction, and task-general adaptation under task-matched evaluation protocols.

Metric details. We use MSE and MAE for long-term forecasting and imputation following standard forecasting/imputation protocols (Zhou et al., 2021; Wu et al., 2023); imputation metrics are computed only on artificially masked positions Ω_m. For short-term PeMS forecasting, we follow the MAE/MAPE/RMSE protocol used by SCINet (Liu et al., 2022a); metrics are computed after inverse normalization, and MAPE excludes near-zero ground-truth values with $\Omega_\delta = \{i \in \Omega : |y_i| > \delta\}$, where $\delta = 1.0$. Classification uses accuracy under the UEA protocol (Bagnall et al., 2018). Anomaly detection uses threshold-free AUROC and AUPRC, standard ranking metrics for ROC and precision–recall analysis (Fawcett, 2006; Davis & Goadrich, 2006). Let y_i and $\hat{y}_i$ be the ground-truth and predicted values over evaluation positions Ω, and let c_i, $\hat{c}_i$ denote ground-truth and predicted classes for N test samples. The metrics are

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{|\Omega|}\sum_{i\in\Omega}(y_i-\hat{y}_i)^2, \qquad \mathrm{MAE} = \frac{1}{|\Omega|}\sum_{i\in\Omega}|y_i-\hat{y}_i|, \\ \mathrm{RMSE} &= \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega}(y_i-\hat{y}_i)^2}, \quad \mathrm{MAPE} = \frac{100}{|\Omega_\delta|}\sum_{i\in\Omega_\delta}\left|\frac{y_i-\hat{y}_i}{y_i}\right|, \\ &\mathrm{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{c}_i = c_i). \end{aligned} \tag{33}$$

For anomaly detection, let s_i be a continuous anomaly score and $a_i \in \{0,1\}$ be the binary label. Traversing distinct score thresholds yields operating points $(\mathrm{FPR}_k, \mathrm{TPR}_k, \mathrm{Precision}_k, \mathrm{Recall}_k)$. We compute

$$\begin{aligned} \mathrm{AUROC} &= \sum_k \frac{\mathrm{TPR}_k + \mathrm{TPR}_{k-1}}{2}\big(\mathrm{FPR}_k - \mathrm{FPR}_{k-1}\big), \\ \mathrm{AUPRC} &= \sum_k (\mathrm{Recall}_k - \mathrm{Recall}_{k-1})\mathrm{Precision}_k. \end{aligned} \tag{34}$$

Lower values are better for MSE, MAE, RMSE, and MAPE, while higher values are better for Accuracy, AUROC, and AUPRC.

D Hyperparameter Sensitivity Analysis

To examine whether the fixed-capacity setting is overly sensitive to local hyperparameter perturbations, Figure 4 reports sensitivity results for the maximum natural source order M, the Cross rank R, and the number of coalition groups K. The experiments cover two representative task families, long-term forecasting and imputation. Each point is the mean MSE across seeds, and the error bar denotes the standard deviation across seeds. Although the main experiments use the uncapped setting $M = V - 1$, the sensitivity analysis explicitly truncates the natural source order and evaluates $M \in \{1, 2, 3, 4, 5\}$. For R and K, the main experimental settings $R = 16$ and $K = 8$ are used as reference endpoints to compare smaller capacities with the unified capacity budget.

The results show clear task dependence for all three hyperparameters. In forecasting, a smaller K or a moderate R can obtain MSE close to, or slightly better than, the full-capacity setting on ETTm1-H96. In imputation, the differences among M, R, and K are generally smaller, but lower or intermediate capacity does not consistently dominate. Based on these results, the main experiments use $R = 16,\ K = 8$ as a unified cross-task capacity setting, preserving comparability across tasks and datasets while providing sufficient capacity for interaction modeling. The sensitivity results further suggest that OSVIF's behavior arises from the joint effect of natural-order organization, dual-view interaction, coalition readout, and task-conditioned readout.

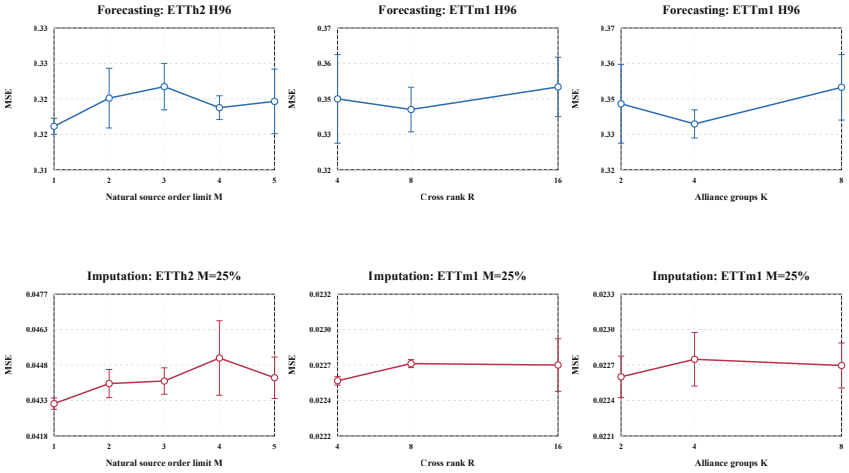

Figure 4: Hyperparameter sensitivity analysis of OSVIF under the fixed-capacity setting. The top row shows forecasting tasks and the bottom row shows imputation tasks; columns correspond to the maximum natural source order M, Cross rank R, and number of coalition groups K.

E Representation Analysis

We analyze OSVIF from four perspectives: whether removing key components causes measurable performance loss; whether task conditions and different coalition degrees participate in the output construction of trained models; whether the allocation between the Identity and Cross views changes across tasks; and whether the intermediate quantities used for interpretation satisfy the invariants implied by the model definition. Let f denote the full model and $f^{(-c)}$ denote the model after intervening on component or readout path c, with corresponding errors $\mathcal{L}(f)$ and $\mathcal{L}(f^{(-c)})$. The structural effect is uniformly defined as

$$\Delta_c = 100 \times \frac{\mathcal{L}(f^{(-c)}) - \mathcal{L}(f)}{\mathcal{L}(f)}. \tag{35}$$

We organize the analysis around four questions. Component ablations measure how the proposed structures affect test error. Condition perturbations examine whether the prediction function depends on the task condition rather than merely receiving it as an input. Coalition-degree statistics describe how output evidence is allocated across low- and high-degree synergies. Identity/Cross and invariant analyses further characterize whether the quantities used in the interpretation respect the structural definitions of OSVIF.

E.1 Stage-Wise Evidence Distribution

Figure 5 compares the representation distributions of three stages in the OSVIF forward path: first-order source-pair outputs, valid natural-order states, and valid coalition outputs. For each window-target unit (b, i) at stage g, let $a^{(g)}_{b,i,j}$ be the derived nonnegative amplitude, where j indexes the candidate members at that stage. The two-dimensional statistics are defined as

$$\mu^{(g)}_{b,i} = \frac{1}{n^{(g)}_{b,i}} \sum_j \log\left(1 + a^{(g)}_{b,i,j}\right), \qquad \sigma^{(g)}_{b,i} = \mathrm{Std}_j\left[\log\left(1 + a^{(g)}_{b,i,j}\right)\right]. \tag{36}$$

The bar height is the percentage of window-target units whose (μ, σ) falls into the corresponding two-dimensional bin. The source stage uses the first six derived first-order source pairs for each tar-

get, while the natural-order and coalition stages use all valid terms. Each stage histogram is independently normalized to 100%. Figure 5 shows stage-level changes in the evidence distribution across source terms, natural-order states, and coalition outputs. These changes provide direct evidence that the quantities analyzed in later tables correspond to distinct stages of the OSVIF hierarchy.

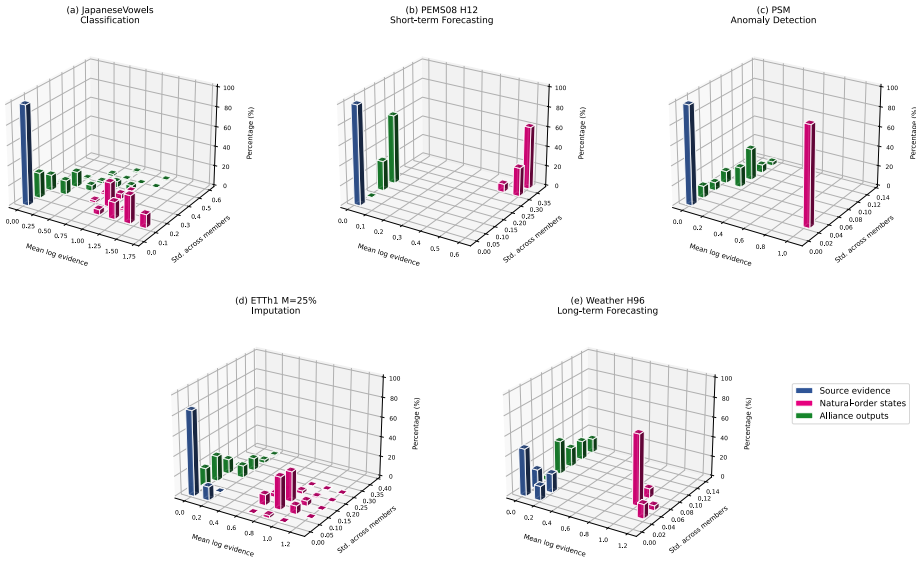

Figure 5: Three-dimensional distributions of OSVIF forward-stage evidence across five task types. Panels: (a) JapaneseVowels classification, (b) PEMS08 H12 short-term forecasting, (c) PSM anomaly detection, (d) ETTh1 M=25% imputation, and (e) Weather H96 long-term forecasting.

E.2 Component and Condition Interventions

Table 8 summarizes controlled ablations of key OSVIF components on the full test set. The four variants restrict the natural source order to first order (No-Higher-Order), disable the Cross branch (No-Cross), replace conditioned coalition readout with additive aggregation (No-Alliance), and disable all task-condition injection (No-Conditioning), respectively. Except for the specified switch, the task setting, data split, optimization process, and evaluation metric are kept unchanged. Long-term forecasting uses Solar-Energy H192, and imputation uses ETTh1 M=25%. Positive values indicate that the ablated model has higher MSE than the corresponding full model.

Table 8: Relative MSE changes after ablating key components on the full test set.

Task	Dataset / Condition	No-Higher-Order	No-Cross	No-Alliance (additive)	No-Conditioning
Long-term forecasting	Solar-Energy / H192	+9.45%	+3.11%	+9.03%	+3.48%
Imputation	ETTh1 / M=25%	+5.98%	+2.72%	+3.28%	+5.83%

All four ablations yield positive Δ_c under both conditions, but their effect sizes differ: on Solar-Energy H192, restricting higher-order sources and using additive coalition readout increase MSE by 9.45% and 9.03%, respectively; on ETTh1 imputation, the four increases range from 2.72% to 5.98%. Thus, in the tested settings, each removed component corresponds to a measurable loss, and the magnitude of that loss depends on the task condition. Table 9 focuses on condition dependence. `Correct` uses the original target condition. `Constant` replaces all target conditions by their target-wise mean. `Shuffled-target` reorders target conditions in a fixed reverse order. `Zero` sets the condition vector to zero. Relative changes are measured against the corresponding correct-condition prediction, so the table reflects how sensitive each task is to the semantic alignment of its condition vector.

All non-correct conditions produce positive error increases, and `Shuffled-target` causes larger degradation than `Constant` under all four conditions. This ordering directly shows that target-

Table 9: Local MSE changes caused by task-condition perturbations.

Task	Dataset / Condition	Constant	Shuffled-target	Zero
Long-term forecasting	ETTm2 / H96	+23.55%	+44.86%	+38.21%
Long-term forecasting	Weather / H192	+136.94%	+361.19%	+140.64%
Imputation	ETTh1 / M=25%	+432.02%	+917.54%	+612.39%
Imputation	ETTh2 / M=25%	+70.29%	+144.22%	+209.04%

wise condition alignment affects the prediction beyond the condition mean. Among the four rows, ETTh1 M=25% and Weather H192 have the largest perturbation effects, so their outputs are the most condition-sensitive in this comparison.

E.3 Coalition-Degree Analysis

OSVIF first organizes source interactions into natural-order states and then combines them with the target state through coalition readout. To distinguish whether a coalition degree is merely available from whether it contributes to the prediction, we measure the Euclidean norm of the output from each coalition degree and its normalized proportion. Let $\Delta y_{b,i,s}$ be the output contribution of target i at coalition degree s for sample b. The aggregated contribution proportion over samples and targets is defined as

$$q_s = \frac{\sum_{b,i} \|\Delta y_{b,i,s}\|_2}{\sum_{s'} \sum_{b,i} \|\Delta y_{b,i,s'}\|_2 + \epsilon}. \tag{37}$$

Across the same five task types as Figure 5, Table 10 summarizes the first-degree coalition proportion, the cumulative proportion of the first three coalition degrees, and the number of effective coalition degrees, where the latter is defined as $|\{s : \exists(b,i), v_{b,i,s} = 1\}|$.

Table 10: Proportions of coalition output contributions.

Task	Dataset / Condition	Degree-1 Proportion	Top-3 Cumulative Proportion	Effective Coalition Degrees
Long-term forecasting	Weather / H96	0.449	0.854	21
Short-term forecasting	PEMS08 / H12	0.451	0.874	51
Imputation	ETTh1 / M=25%	0.824	0.994	7
Classification	JapaneseVowels	0.201	0.473	12
Anomaly detection	PSM	0.077	0.223	25

The five tasks exhibit different degree allocations. Weather H96 and PEMS08 H12 have similar first-degree coalition proportions, and their cumulative contributions from the first three degrees both exceed 85%. ETTh1 imputation is even more concentrated in the first three degrees. JapaneseVowels and PSM, by contrast, place 52.7% and 77.7% of their output norm, respectively, in coalition degrees four and above, yielding more dispersed degree spectra. Thus, the same coalition readout forms task-dependent output decompositions rather than collapsing all coalition orders into a uniform pattern.

To examine the complementarity of different coalition degrees, Table 11 compares the complete coalition sum with two degree-restricted readouts: `Degree-1`, which keeps only $\Delta y_{b,i,1}$, and `Higher-only`, which keeps only $\sum_{s\geq 2} \Delta y_{b,i,s}$. Relative changes are measured with respect to the complete sum $\sum_s \Delta y_{b,i,s}$.

Table 11: Local MSE changes caused by truncating coalition outputs.

Task	Dataset / Condition	Degree-1	Higher-only
Long-term forecasting	ETTm2 / H96	+9.40%	+25.96%
Long-term forecasting	Weather / H192	+72.82%	+39.62%
Imputation	ETTh1 / M=25%	+60.32%	+1013.12%
Imputation	ETTh2 / M=25%	+22.29%	+354.73%

Under all four conditions, both degree-restricted readouts increase MSE, indicating that the first-degree and higher-degree outputs are complementary rather than interchangeable. In the two imputation rows, `Higher-only` is much worse than `Degree-1`, with increases of 1013.12% versus 60.32% on ETTh1 and 354.73% versus 22.29% on ETTh2. On Weather H192, the ordering is

reversed, with `Degree-1` increasing MSE more than `Higher-only`. These task-dependent orderings are the basis for treating coalition-degree analysis as part of the method analysis rather than as a generic visualization.

E.4 Identity and Cross Analysis

The Identity view characterizes coordinate-wise products over the same hidden coordinates, while the Cross view is formed by off-diagonal interactions in low-rank projections. For natural orders $m \geq 2$, the dual-view selector produces Softmax weights satisfying $w^I_{b,i,m} + w^X_{b,i,m} = 1$. Let $\mathcal{V}_m$ be the set of valid window-target units at order m. We first define the within-order average as

$$\bar{w}^X_m = \frac{1}{|\mathcal{V}_m|} \sum_{(b,i)\in\mathcal{V}_m} w^X_{b,i,m}, \tag{38}$$

We then report $|\mathcal{M}|^{-1} \sum_{m\in\mathcal{M}} \bar{w}^X_m$, together with $\min_m \bar{w}^X_m$ and $\max_m \bar{w}^X_m$, over the valid natural-order set $\mathcal{M}$. First-order terms do not contain the Cross branch and are therefore excluded from this statistic.

Table 12: Cross-order statistics of Cross fusion weights.

Task	Dataset / Condition	Mean Cross Weight	Minimum	Maximum
Long-term forecasting	Weather / H96	0.475	0.456	0.491
Short-term forecasting	PEMS08 / H12	0.152	0.073	0.242
Imputation	ETTh1 / M=25%	0.732	0.614	0.785
Classification	JapaneseVowels	0.482	0.465	0.498
Anomaly detection	PSM	0.460	0.458	0.466

The Cross weights vary substantially with the task condition: the cross-order mean reaches 0.732 for ETTh1 imputation, is 0.152 for PEMS08 H12, and is close to 0.5 for the other three conditions. Together with the No-Cross ablation in Table 8, this gives two direct observations: Cross receives nonzero allocation across all five analyzed tasks, and removing Cross increases MSE in the two ablation settings. The Identity/Cross analysis therefore links the learned allocation weights with a corresponding component-removal effect.

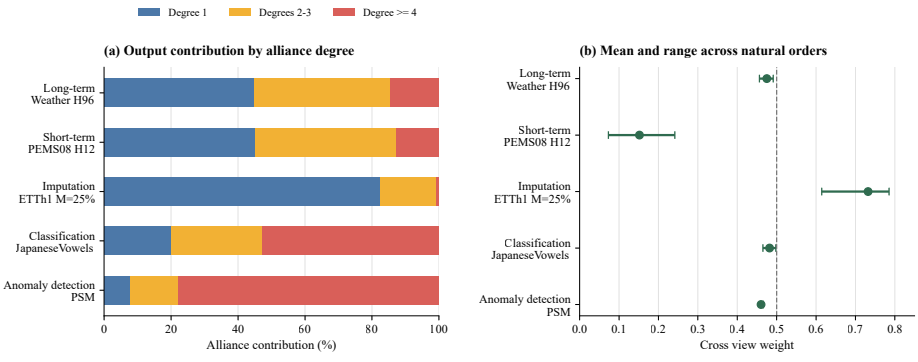

Figure 6: Coalition contribution and Cross-view allocation across five task types. The left panel separates coalition output-norm proportions into degree 1, degrees 2–3, and degrees 4 and above. In the right panel, dots denote the cross-order mean of Cross weights for orders 2 and above, and horizontal lines show the minimum-to-maximum range of means across natural orders rather than confidence intervals. The dashed line indicates equal weighting of Identity and Cross.

E.5 ILI-H60 Case Study

Beyond aggregate statistics, we examine whether the higher-order outputs of the model align with an external correlation structure at the variable level on ILI-H60. For 128 consecutive windows,

we flatten the future interval and compute the variable correlation matrix R^{future}. If the number of variables is V, the absolute correlation centrality of variable i is defined as

$$c_i = \frac{1}{V-1} \sum_{j \neq i} \left| R_{ij}^{\text{future}} \right|. \tag{39}$$

From the coalition-degree output spectrum, we take the mean absolute value over windows and output dimensions and sum coalition degrees two and above to obtain variable-level higher-order evidence e_i^{HO}. We then compute the Pearson correlation and Spearman rank correlation between $\{c_i\}$ and $\{e_i^{\text{HO}}\}$ over the $V = 7$ variables. This comparison uses the future correlation matrix as an external descriptor.

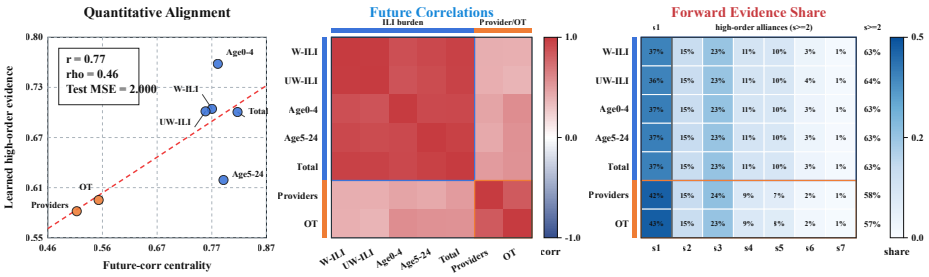

Figure 7: Coalition evidence derived from the forward process on ILI-H60.

In Figure 7, the Pearson coefficient between future correlation centrality and model higher-order evidence is 0.766, and the Spearman coefficient is 0.464. The average higher-order coalition proportion over the seven variables is 61.5%. These positive correlation values establish a variable-level association between the model's higher-order evidence and an external future-correlation descriptor.

E.6 Structural Invariants

Finally, we examine the structural invariants implied by the OSVIF definition. For natural order $m \geq 2$, let $C_{m,r}$ denote the weighted Cross component in the r-th low-rank direction. The Cross term is defined as

$$C_m = \sum_{r=1}^{R} C_{m,r}. \tag{40}$$

For each natural order, define $e_{\text{abs},m} = \|C_m - \sum_r C_{m,r}\|_\infty$ and $e_{\text{rel},m} = e_{\text{abs},m}/(\|C_m\|_\infty + \epsilon)$. The table reports the maximum values across orders. The remaining invariants correspond to three structural requirements: the target variable is excluded from its own source set; if P is a variable permutation, the output satisfies $f(Px) = Pf(x)$; and the valid-order sets satisfy $\mathcal{V}_{m+1} \subseteq \mathcal{V}_m$. These properties connect the analyzed quantities to target exclusion, source permutation equivariance, valid-order nesting, and Cross-direction synthesis.

Table 13: Structural invariant checks.

Invariant	Observed Value	Criterion	Result
Target-self exclusion	0 violations	0	Pass
Source-permutation equivariance error	4.77e-7	$\leq$ 1e-5	Pass
Valid-order nesting	0 violations	0	Pass
Cross direction-sum absolute error	0.0	$\leq$ 1e-5	Pass
Cross direction-sum relative error	0.0	$\leq$ 1e-5	Pass
Cross directions	16	Equal to effective rank	Pass

The zero violation counts and the equivariance error within numerical tolerance show that the analyzed quantities preserve the intended structure of OSVIF. Combining these invariants with the preceding ablations and allocation statistics gives three evidence-based conclusions: component removal increases error in the tested settings; coalition degrees and Identity/Cross views have different allocation profiles across task families; and the quantities used for interpretation satisfy the target-centered and order-aware structural constraints defined by the model.

F Full Experimental Results

Due to space limitations in the main text, the complete experimental results for the five task types are summarized below: Table 14 reports long-term forecasting results, Table 15 reports short-term forecasting results, Table 16 reports imputation results in two parts, Table 17 reports classification results, and Table 18 reports anomaly-detection results. These tables provide the per-dataset basis for the cross-task performance claims in the main text and show where the gains change with dataset, horizon, masking ratio, or metric. Red bold and blue underlined values indicate the best and second best results, respectively.

Table 14: Long-term forecasting results. Except for ILI, whose input length is $T = 36$, all datasets use $T = 96$. Prediction lengths H are shown in the table. **Red**: best, Blue: second best.

Models		OSVIF (Ours)		TimeMixer (2024)		UniTS-ST (2024)		iTransformer (2024)		TimesNet (2023)		PatchTST (2023)		Crossformer (2023)		TiDE (2023)		DLinear (2023)		SCINet (2022a)		FEDformer (2022)		Stationary (2022c)		Autoformer (2021)	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.391**	**0.411**	0.430	0.443	0.489	0.465	0.405	0.418	0.427	0.436	0.460	0.447	0.423	0.448	0.479	0.464	0.397	0.412	0.654	0.599	0.395	0.424	0.513	0.491	0.449	0.459
	192	**0.444**	0.442	0.482	0.466	0.523	0.486	0.457	0.448	0.477	0.462	0.477	**0.429**	0.471	0.474	0.525	0.492	0.446	0.441	0.719	0.631	0.469	0.470	0.534	0.504	0.500	0.482
	336	0.491	0.470	0.521	0.484	0.557	0.502	0.501	0.472	0.530	0.494	0.546	0.496	0.570	0.546	0.565	0.515	**0.489**	**0.467**	0.778	0.659	0.530	0.499	0.588	0.535	0.521	0.496
	720	**0.506**	**0.492**	0.607	0.557	0.605	0.546	0.543	0.516	0.548	0.508	0.544	0.517	0.653	0.621	0.594	0.558	0.513	0.510	0.836	0.699	0.598	0.544	0.643	0.616	0.514	0.512
	Avg	**0.458**	**0.453**	0.510	0.487	0.544	0.500	0.476	0.464	0.496	0.475	0.516	0.484	0.529	0.522	0.541	0.507	0.461	0.457	0.747	0.647	0.498	0.484	0.570	0.537	0.496	0.487
ETTh2	96	0.320	0.368	0.335	0.366	0.333	0.375	**0.299**	**0.349**	0.334	0.375	0.308	0.355	0.745	0.584	0.400	0.440	0.340	0.394	0.707	0.621	0.358	0.397	0.476	0.458	0.346	0.388
	192	0.401	0.415	0.475	0.444	0.409	0.420	**0.382**	**0.400**	0.429	0.432	0.393	0.405	0.877	0.656	0.528	0.509	0.482	0.479	0.860	0.689	0.429	0.439	0.512	0.493	0.456	0.452
	336	0.441	0.448	0.506	0.470	0.453	0.456	**0.425**	**0.434**	0.452	0.450	0.427	0.436	1.043	0.731	0.643	0.571	0.591	0.541	1.000	0.744	0.496	0.487	0.552	0.551	0.482	0.486
	720	0.449	0.461	0.447	0.456	0.460	0.466	**0.431**	**0.448**	0.478	0.477	0.436	0.450	1.104	0.763	0.874	0.679	0.839	0.661	1.249	0.838	0.463	0.474	0.562	0.560	0.515	0.511
	Avg	0.403	0.423	0.441	0.434	0.414	0.429	**0.384**	**0.408**	0.423	0.434	0.391	0.411	0.942	0.684	0.611	0.550	0.563	0.519	0.954	0.723	0.437	0.449	0.526	0.516	0.450	0.459
ETTm1	96	0.340	0.375	**0.330**	**0.374**	0.350	0.379	0.346	0.379	0.354	0.386	0.352	**0.374**	0.404	0.426	0.364	0.387	0.346	**0.374**	0.418	0.438	0.379	0.419	0.386	0.398	0.505	0.475
	192	0.385	0.400	0.382	0.397	0.395	0.401	0.386	0.395	0.394	0.406	**0.374**	**0.387**	0.450	0.451	0.398	0.404	0.382	0.391	0.439	0.450	0.426	0.441	0.459	0.444	0.553	0.496
	336	**0.415**	0.421	**0.415**	0.419	0.423	0.417	0.423	0.419	0.429	0.428	0.421	**0.414**	0.532	0.515	0.428	0.425	**0.415**	0.415	0.490	0.485	0.445	0.459	0.495	0.464	0.621	0.537
	720	0.477	0.457	0.506	0.475	0.486	0.451	0.495	0.459	0.510	0.469	**0.462**	**0.449**	0.666	0.589	0.487	0.461	0.473	0.451	0.595	0.550	0.543	0.490	0.585	0.516	0.671	0.561
	Avg	**0.404**	0.413	0.408	0.416	0.414	0.412	0.413	0.413	0.422	0.422	0.406	**0.407**	0.513	0.495	0.419	0.419	**0.404**	0.408	0.485	0.481	0.448	0.452	0.481	0.456	0.588	0.517
ETTm2	96	0.183	**0.267**	0.193	0.278	**0.182**	0.270	0.188	0.274	0.189	0.271	0.183	0.270	0.287	0.366	0.207	0.305	0.193	0.293	0.286	0.377	0.203	0.287	0.192	0.274	0.255	0.339
	192	**0.246**	**0.307**	0.272	0.324	0.252	0.311	0.253	0.313	0.258	0.310	0.255	0.314	0.414	0.492	0.290	0.364	0.284	0.361	0.399	0.445	0.269	0.328	0.280	0.339	0.281	0.340
	336	0.315	0.354	0.363	0.375	0.315	0.353	0.313	0.351	0.324	0.349	**0.309**	**0.347**	0.597	0.542	0.377	0.422	0.382	0.429	0.637	0.591	0.325	0.366	0.334	0.361	0.339	0.372
	720	0.419	0.415	0.442	0.423	**0.410**	0.405	0.413	0.406	0.435	0.413	0.412	**0.404**	1.730	1.042	0.558	0.524	0.558	0.525	0.960	0.735	0.421	0.415	0.417	0.413	0.433	0.432
	Avg	0.291	0.335	0.317	0.350	**0.290**	0.335	0.292	0.336	0.302	0.336	**0.290**	**0.334**	0.757	0.610	0.358	0.404	0.354	0.402	0.954	0.723	0.305	0.349	0.306	0.347	0.327	0.371
Weather	96	**0.170**	**0.211**	0.175	0.225	0.181	0.226	0.191	0.231	**0.170**	0.220	0.186	0.227	0.195	0.271	0.202	0.261	0.195	0.252	0.221	0.306	0.217	0.296	0.173	0.223	0.266	0.336
	192	0.217	**0.256**	0.233	0.270	0.222	0.260	0.238	0.270	0.226	0.266	0.234	0.265	**0.209**	0.277	0.242	0.298	0.237	0.295	0.261	0.340	0.276	0.336	0.245	0.285	0.307	0.367
	336	0.274	**0.297**	0.286	0.309	0.280	0.301	0.291	0.308	0.285	0.306	0.284	0.301	**0.273**	0.332	0.287	0.335	0.282	0.331	0.309	0.378	0.339	0.380	0.321	0.338	0.359	0.395
	720	0.349	**0.346**	0.376	0.363	0.354	0.349	0.368	0.358	0.376	0.364	0.356	0.349	0.379	0.401	0.351	0.386	**0.345**	0.382	0.377	0.427	0.403	0.428	0.414	0.410	0.419	0.428
	Avg	**0.252**	**0.277**	0.268	0.292	0.259	0.284	0.272	0.292	0.264	0.289	0.265	0.285	0.264	0.320	0.271	0.320	0.265	0.315	0.292	0.363	0.309	0.360	0.288	0.314	0.338	0.382
ECL	96	0.172	**0.265**	0.227	0.322	0.310	0.384	0.213	0.299	**0.168**	0.272	0.190	0.296	0.219	0.314	0.237	0.329	0.210	0.302	0.247	0.345	0.193	0.308	0.169	0.273	0.201	0.317
	192	**0.180**	**0.272**	0.231	0.329	0.312	0.389	0.222	0.311	0.188	0.289	0.199	0.304	0.231	0.322	0.236	0.330	0.210	0.305	0.257	0.355	0.201	0.315	0.182	0.286	0.222	0.334
	336	0.205	**0.293**	0.254	0.352	0.330	0.404	0.243	0.331	0.208	0.307	0.217	0.319	0.246	0.337	0.249	0.344	0.223	0.319	0.269	0.369	0.214	0.329	**0.200**	0.304	0.231	0.443
	720	0.256	0.335	0.288	0.373	0.372	0.434	0.291	0.368	0.233	0.325	0.258	0.352	0.280	0.363	0.284	0.373	0.258	0.350	0.299	0.390	0.246	0.355	**0.222**	**0.321**	0.254	0.361
	Avg	0.203	**0.291**	0.250	0.344	0.331	0.403	0.242	0.327	0.199	0.298	0.216	0.318	0.244	0.334	0.251	0.344	0.225	0.319	0.268	0.365	0.214	0.327	**0.193**	0.296	0.227	0.338
Traffic	96	**0.499**	0.328	0.702	0.436	0.884	0.527	0.670	0.428	0.596	**0.317**	0.526	0.347	0.644	0.429	0.805	0.493	0.650	0.396	0.788	0.499	0.587	0.366	0.612	0.338	0.613	0.388
	192	**0.506**	0.337	0.742	0.440	0.877	0.531	0.677	0.436	0.615	0.335	0.522	**0.332**	0.665	0.431	0.756	0.474	0.598	0.370	0.789	0.505	0.604	0.373	0.613	0.340	0.616	0.382
	336	0.539	0.352	0.749	0.444	0.904	0.541	0.714	0.453	0.629	0.333	**0.517**	0.334	0.674	0.420	0.762	0.477	0.605	0.373	0.797	0.508	0.621	0.383	0.618	**0.328**	0.622	0.337
	720	0.568	0.361	0.794	0.463	0.949	0.558	0.784	0.484	0.669	**0.352**	**0.552**	**0.352**	0.683	0.424	0.719	0.449	0.645	0.394	0.841	0.523	0.626	0.382	0.653	0.355	0.660	0.408
	Avg	**0.528**	0.345	0.747	0.446	0.903	0.539	0.711	0.450	0.627	**0.334**	0.529	0.341	0.667	0.426	0.760	0.473	0.625	0.383	0.804	0.509	0.610	0.376	0.624	0.340	0.628	0.379
Exchange	96	**0.084**	**0.203**	0.097	0.219	0.107	0.230	0.089	0.209	0.118	0.248	0.088	0.205	0.256	0.367	0.094	0.218	0.088	0.218	0.267	0.396	0.148	0.278	0.111	0.237	0.197	0.323
	192	**0.173**	**0.295**	0.202	0.319	0.222	0.339	0.181	0.303	0.218	0.335	0.176	0.299	0.470	0.509	0.184	0.307	0.176	0.315	0.351	0.459	0.271	0.315	0.219	0.335	0.300	0.369
	336	0.343	0.425	0.375	0.442	0.376	0.445	0.334	0.420	0.399	0.468	**0.301**	**0.397**	1.268	0.883	0.349	0.431	0.313	0.427	1.324	0.853	0.460	0.427	0.421	0.476	0.509	0.524
	720	0.929	0.726	1.175	0.825	0.965	0.748	0.868	0.704	1.044	0.786	0.901	0.714	1.767	1.068	0.852	0.698	**0.839**	**0.695**	1.058	0.797	1.195	**0.695**	1.092	0.769	1.447	0.941
	Avg	0.382	0.412	0.462	0.451	0.418	0.440	0.368	0.409	0.445	0.459	0.367	**0.404**	0.940	0.707	0.370	0.413	**0.354**	0.414	0.750	0.626	0.519	0.429	0.461	0.454	0.613	0.539
Solar-Energy	96	**0.191**	**0.234**	0.238	0.290	0.278	0.313	0.270	0.300	0.247	0.295	0.265	0.323	0.232	0.302	0.312	0.399	0.290	0.378	0.237	0.344	0.286	0.341	0.321	0.380	0.456	0.446
	192	**0.229**	**0.260**	0.290	0.312	0.328	0.344	0.311	0.325	0.285	0.314	0.288	0.332	0.371	0.410	0.339	0.416	0.320	0.398	0.280	0.380	0.291	0.337	0.346	0.369	0.588	0.561
	336	**0.266**	**0.289**	0.331	0.333	0.395	0.378	0.355	0.348	0.313	0.328	0.301	0.339	0.495	0.515	0.368	0.430	0.353	0.415	0.304	0.389	0.354	0.416	0.357	0.387	0.595	0.588
	720	**0.259**	**0.282**	0.337	0.339	0.381	0.363	0.363	0.352	0.319	0.333	0.295	0.336	0.526	0.542	0.370	0.425	0.357	0.413	0.308	0.388	0.380	0.437	0.375	0.424	0.733	0.633
	Avg	**0.236**	**0.266**	0.299	0.318	0.346	0.349	0.325	0.332	0.291	0.317	0.287	0.333	0.406	0.442	0.347	0.417	0.330	0.401	0.282	0.375	0.328	0.383	0.350	0.390	0.586	0.557
ILI	24	2.662	**0.944**	3.523	1.367	4.230	1.529	3.086	1.235	2.669	1.018	**2.435**	1.059	5.216	1.622	2.518	1.090	3.627	1.382	2.492	1.079	4.130	1.506	3.983	1.442	4.040	1.499
	36	2.596	1.000	3.503	1.352	4.185	1.501	3.185	1.258	**2.219**	**0.994**	2.349	1.039	4.987	1.590	2.501	1.094	3.637	1.368	2.364	1.057	4.056	1.484	3.931	1.421	3.964	1.470
	48	**2.082**	**0.871**	3.501	1.322	4.217	1.490	3.350	1.299	2.249	1.001	2.296	1.025	5.319	1.638	2.315	1.043	3.690	1.371	2.350	1.057	4.036	1.461	3.887	1.396	3.878	1.450
	60	**2.000**	**0.901**	3.809	1.409	4.378	1.512	3.480	1.330	2.156	1.017	2.250	1.027	5.452	1.662	2.241	1.033	3.908	1.397	2.366	1.071	4.148	1.474	3.978	1.407	4.002	1.465
	Avg	2.335	**0.929**	3.584	1.362	4.253	1.508	3.275	1.280	**2.323**	1.008	2.332	1.037	5.244	1.628	2.394	1.065	3.715	1.380	2.393	1.066	4.093	1.481	3.945	1.417	3.971	1.471
1st Count		**47**		3		3		10		8		21		2		0		9		0		1		5		0	

Table 15: Short-term forecasting results on PeMS traffic datasets (Chen et al., 2001). All PEMS datasets use $T = 96$ and $H = 12$. MAPE is reported as a percentage value. **Red**: best, Blue: second best.

Models		OSVIF (Ours)	TimeMixer (2024)	UniTS-ST (2024)	iTransformer (2024)	TimesNet (2023)	TiDE (2023)	Crossformer (2023)	PatchTST (2023)	MICN (2023b)	DLinear (2023)	FEDformer (2022)	Stationary (2022c)	Autoformer (2021)
PEMS03	MAE	**16.63**	17.08	21.33	18.14	19.71	18.39	17.93	20.39	16.73	19.70	19.00	17.64	18.08
	MAPE	**15.15**	17.31	20.74	17.81	20.43	17.39	19.22	20.49	17.87	18.35	18.57	17.56	18.75
	RMSE	26.18	26.50	33.54	28.61	31.11	30.59	28.91	31.61	**25.82**	32.35	30.05	28.37	27.82
PEMS04	MAE	**22.23**	23.24	37.58	25.13	26.78	23.99	23.62	27.79	23.24	24.62	26.51	22.34	25.00
	MAPE	**14.60**	17.05	28.04	18.33	20.92	14.76	18.08	21.19	17.26	16.12	16.76	14.85	16.70
	RMSE	**34.98**	35.89	56.33	39.12	40.53	37.30	36.03	41.86	35.90	39.51	41.81	35.47	38.02
PEMS07	MAE	24.30	24.79	32.63	26.78	29.85	25.31	26.82	30.90	**24.11**	28.65	27.92	26.02	26.92
	MAPE	**10.23**	10.95	14.58	11.71	13.76	11.02	12.66	15.07	10.42	12.15	12.29	11.75	11.83
	RMSE	38.00	38.41	48.47	41.32	45.26	39.12	40.36	45.37	**37.37**	45.02	42.29	42.34	40.60
PEMS08	MAE	**17.19**	18.87	31.64	19.74	22.69	19.73	18.99	22.22	18.97	20.26	20.56	19.29	20.47
	MAPE	**10.52**	12.07	20.55	12.42	14.71	12.02	12.10	14.13	11.97	12.09	12.41	12.21	12.27
	RMSE	**27.22**	29.28	46.48	30.91	34.29	30.67	29.00	33.58	29.64	32.38	32.97	38.62	31.52

Table 16: Imputation results (Part I). The input length is $T = 96$, and the random masking ratios M are 12.5%, 25%, 37.5%, and 50%. MSE and MAE are computed only on artificially masked positions. **Red**: best, Blue: second best.

Models		OSVIF (Ours)		TimeMixer (2024)		UniTS-ST (2024)		iTransformer (2024)		TimesNet (2023)		PatchTST (2023)		DLinear (2023)		ETSformer (2022)		LightTS (2022b)		FEDformer (2022)	
	Mask ratio	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	12.5%	**0.019**	**0.090**	0.063	0.171	0.367	0.400	0.054	0.162	**0.019**	0.091	0.040	0.127	0.058	0.162	0.067	0.188	0.075	0.180	0.035	0.135
	25%	**0.022**	**0.097**	0.068	0.177	0.631	0.556	0.080	0.201	0.023	0.100	0.042	0.132	0.080	0.193	0.096	0.229	0.093	0.206	0.052	0.166
	37.5%	**0.027**	**0.105**	0.072	0.182	0.682	0.601	0.121	0.251	0.028	0.110	0.047	0.139	0.103	0.219	0.133	0.271	0.113	0.231	0.069	0.191
	50%	**0.033**	**0.115**	0.082	0.193	0.755	0.650	0.166	0.293	0.036	0.126	0.054	0.148	0.132	0.248	0.186	0.323	0.134	0.255	0.089	0.218
	Avg	**0.025**	**0.102**	0.071	0.181	0.609	0.552	0.105	0.226	0.027	0.107	0.046	0.137	0.093	0.206	0.120	0.253	0.104	0.218	0.061	0.177
ETTm2	12.5%	**0.016**	**0.071**	0.040	0.131	0.157	0.286	0.058	0.160	0.019	0.080	0.025	0.092	0.062	0.166	0.108	0.239	0.034	0.127	0.056	0.159
	25%	**0.018**	**0.075**	0.040	0.131	0.304	0.416	0.099	0.209	0.021	0.086	0.028	0.098	0.085	0.196	0.164	0.294	0.042	0.143	0.080	0.195
	37.5%	**0.020**	**0.080**	0.042	0.135	0.509	0.539	0.197	0.294	0.023	0.091	0.030	0.103	0.106	0.222	0.237	0.356	0.051	0.159	0.110	0.231
	50%	**0.024**	**0.087**	0.043	0.135	0.773	0.668	0.331	0.378	0.025	0.097	0.034	0.109	0.131	0.247	0.323	0.421	0.059	0.174	0.156	0.276
	Avg	**0.020**	**0.078**	0.042	0.133	0.436	0.477	0.171	0.260	0.022	0.088	0.029	0.101	0.096	0.208	0.208	0.328	0.046	0.151	0.101	0.215
ETTh1	12.5%	**0.053**	**0.154**	0.199	0.313	0.630	0.530	0.116	0.240	**0.053**	0.158	0.096	0.202	0.151	0.267	0.126	0.263	0.240	0.345	0.070	0.190
	25%	**0.067**	**0.170**	0.214	0.320	0.683	0.568	0.160	0.285	0.071	0.182	0.112	0.220	0.180	0.292	0.169	0.304	0.265	0.364	0.106	0.236
	37.5%	**0.086**	**0.192**	0.207	0.318	0.723	0.609	0.241	0.357	0.088	0.203	0.128	0.235	0.215	0.318	0.220	0.347	0.296	0.382	0.124	0.258
	50%	0.114	**0.220**	0.226	0.330	0.785	0.656	0.345	0.435	**0.107**	0.223	0.153	0.255	0.257	0.347	0.293	0.402	0.334	0.404	0.165	0.299
	Avg	**0.080**	**0.184**	0.211	0.320	0.705	0.591	0.216	0.329	**0.080**	0.191	0.122	0.228	0.201	0.306	0.202	0.329	0.284	0.374	0.116	0.246
ETTh2	12.5%	**0.040**	**0.122**	0.102	0.219	0.216	0.334	0.106	0.222	**0.040**	0.131	0.056	0.150	0.100	0.216	0.187	0.319	0.101	0.231	0.095	0.212
	25%	**0.045**	**0.130**	0.099	0.216	0.359	0.440	0.159	0.272	0.047	0.143	0.061	0.158	0.127	0.247	0.279	0.390	0.115	0.246	0.137	0.258
	37.5%	**0.051**	**0.140**	0.100	0.217	0.555	0.550	0.269	0.351	0.052	0.152	0.067	0.166	0.158	0.276	0.400	0.465	0.126	0.257	0.187	0.304
	50%	**0.058**	**0.152**	0.102	0.218	0.816	0.673	0.452	0.454	0.060	0.163	0.073	0.174	0.183	0.299	0.602	0.572	0.136	0.268	0.232	0.341
	Avg	**0.048**	**0.136**	0.101	0.218	0.487	0.499	0.246	0.325	0.050	0.147	0.065	0.162	0.142	0.260	0.367	0.436	0.120	0.251	0.163	0.279
Electricity	12.5%	**0.053**	**0.152**	0.176	0.292	0.301	0.391	0.080	0.201	0.089	0.204	0.059	0.167	0.092	0.214	0.196	0.321	0.102	0.229	0.107	0.237
	25%	**0.063**	**0.166**	0.182	0.298	0.712	0.690	0.105	0.234	0.092	0.208	0.069	0.181	0.118	0.247	0.207	0.332	0.121	0.252	0.120	0.251
	37.5%	**0.074**	**0.180**	0.197	0.306	0.820	0.759	0.134	0.266	0.097	0.214	0.082	0.196	0.144	0.276	0.219	0.344	0.141	0.273	0.136	0.266
	50%	**0.088**	**0.196**	0.201	0.310	0.843	0.769	0.154	0.283	0.103	0.222	0.098	0.216	0.175	0.305	0.235	0.357	0.160	0.293	0.158	0.284
	Avg	**0.070**	**0.174**	0.189	0.302	0.669	0.652	0.118	0.246	0.095	0.212	0.077	0.190	0.132	0.261	0.214	0.339	0.131	0.262	0.130	0.260
Weather	12.5%	0.028	**0.042**	0.061	0.105	0.094	0.162	0.040	0.088	**0.026**	0.049	0.029	0.048	0.039	0.084	0.057	0.141	0.047	0.101	0.041	0.107
	25%	0.030	**0.043**	0.059	0.104	0.149	0.255	0.052	0.115	0.031	0.059	0.031	0.053	0.048	0.103	0.065	0.155	0.052	0.111	0.064	0.163
	37.5%	**0.031**	**0.043**	0.059	0.105	0.197	0.316	0.077	0.159	0.033	0.063	0.033	0.057	0.057	0.117	0.081	0.180	0.058	0.121	0.107	0.229
	50%	**0.034**	**0.046**	0.061	0.105	0.263	0.381	0.119	0.213	**0.034**	0.062	0.037	0.062	0.066	0.134	0.102	0.207	0.065	0.133	0.183	0.312
	Avg	**0.031**	**0.043**	0.060	0.105	0.176	0.278	0.072	0.144	**0.031**	0.058	0.033	0.055	0.053	0.110	0.076	0.171	0.056	0.117	0.099	0.203
1st Count		**57**		0		0		0		8		0		0		0		0		0	

Table 16: Imputation results (continued; Part II). The input length is $T = 96$, and the random masking ratios M are 12.5%, 25%, 37.5%, and 50%. MSE and MAE are computed only on artificially masked positions. **Red**: best, Blue: second best.

Models		OSVIF (Ours)		Stationary (2022c)		LSSL (2021)		Autoformer (2021)		Pyraformer (2022b)		Informer (2021)		Reformer (2020)		LogTrans (2019)		TCN (2018)		LSTM (1997)	
Mask ratio		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	12.5%	**0.019**	**0.090**	0.026	0.107	0.101	0.231	0.034	0.124	0.670	0.541	0.047	0.155	0.032	0.126	0.041	0.141	0.510	0.493	0.974	0.780
	25%	**0.022**	**0.097**	0.032	0.119	0.106	0.235	0.046	0.144	0.689	0.553	0.063	0.180	0.042	0.146	0.044	0.144	0.518	0.500	1.032	0.807
	37.5%	**0.027**	**0.105**	0.039	0.131	0.116	0.246	0.057	0.161	0.737	0.581	0.079	0.200	0.063	0.182	0.052	0.158	0.516	0.499	0.999	0.792
	50%	**0.033**	**0.115**	0.047	0.145	0.129	0.260	0.067	0.174	0.770	0.605	0.093	0.218	0.082	0.208	0.063	0.173	0.519	0.496	0.952	0.763
	Avg	**0.025**	**0.102**	0.036	0.126	0.113	0.243	0.051	0.151	0.717	0.570	0.071	0.188	0.055	0.165	0.050	0.154	0.516	0.497	0.989	0.786
ETTm2	12.5%	**0.016**	**0.071**	0.021	0.088	0.150	0.298	0.023	0.092	0.394	0.470	0.133	0.270	0.108	0.228	0.103	0.229	0.307	0.441	1.013	0.805
	25%	**0.018**	**0.075**	0.024	0.096	0.159	0.306	0.026	0.101	0.421	0.482	0.135	0.272	0.136	0.262	0.120	0.248	0.263	0.402	1.039	0.814
	37.5%	**0.020**	**0.080**	0.027	0.103	0.180	0.321	0.030	0.108	0.478	0.521	0.155	0.293	0.175	0.300	0.138	0.260	0.250	0.396	0.917	0.744
	50%	**0.024**	**0.087**	0.030	0.108	0.210	0.353	0.035	0.119	0.568	0.560	0.200	0.333	0.211	0.329	0.117	0.247	0.246	0.389	1.140	0.835
	Avg	**0.020**	**0.078**	0.026	0.099	0.175	0.320	0.029	0.105	0.465	0.508	0.156	0.292	0.158	0.280	0.119	0.246	0.267	0.407	1.027	0.799
ETTh1	12.5%	**0.053**	**0.154**	0.060	0.165	0.422	0.461	0.074	0.182	0.857	0.609	0.114	0.234	0.074	0.194	0.229	0.330	0.599	0.554	1.265	0.896
	25%	**0.067**	**0.170**	0.080	0.189	0.412	0.456	0.090	0.203	0.829	0.672	0.140	0.262	0.102	0.227	0.207	0.323	0.610	0.567	1.262	0.883
	37.5%	**0.086**	**0.192**	0.102	0.212	0.421	0.461	0.109	0.222	0.830	0.675	0.174	0.293	0.135	0.261	0.210	0.328	0.628	0.577	1.200	0.867
	50%	0.114	**0.220**	0.133	0.240	0.443	0.473	0.137	0.248	0.854	0.691	0.215	0.325	0.179	0.298	0.230	0.348	0.648	0.587	1.174	0.849
	Avg	**0.080**	**0.184**	0.094	0.202	0.424	0.463	0.103	0.214	0.843	0.662	0.161	0.279	0.122	0.245	0.219	0.332	0.621	0.571	1.225	0.874
ETTh2	12.5%	**0.040**	**0.122**	0.042	0.133	0.521	0.555	0.044	0.138	0.976	0.754	0.305	0.431	0.163	0.289	0.173	0.308	0.410	0.494	2.060	1.120
	25%	**0.045**	**0.130**	0.049	0.147	0.487	0.535	0.050	0.149	1.037	0.774	0.322	0.444	0.206	0.331	0.175	0.310	0.419	0.490	2.007	1.105
	37.5%	**0.051**	**0.140**	0.056	0.158	0.487	0.529	0.060	0.163	1.107	0.800	0.353	0.462	0.252	0.370	0.185	0.315	0.429	0.498	2.033	1.111
	50%	**0.058**	**0.152**	0.065	0.170	0.484	0.523	0.068	0.173	1.193	0.838	0.369	0.472	0.316	0.419	0.212	0.339	0.467	0.529	2.054	1.119
	Avg	**0.048**	**0.136**	0.053	0.152	0.495	0.536	0.056	0.156	1.078	0.791	0.337	0.452	0.234	0.352	0.186	0.318	0.431	0.503	2.038	1.114
Electricity	12.5%	**0.053**	**0.152**	0.093	0.210	0.217	0.341	0.089	0.210	0.297	0.383	0.218	0.326	0.190	0.308	0.164	0.296	0.621	0.620	0.277	0.366
	25%	**0.063**	**0.166**	0.097	0.214	0.219	0.341	0.096	0.220	0.294	0.380	0.219	0.326	0.197	0.312	0.169	0.299	0.559	0.585	0.281	0.369
	37.5%	**0.074**	**0.180**	0.102	0.220	0.223	0.343	0.104	0.229	0.296	0.381	0.222	0.328	0.203	0.315	0.178	0.305	0.567	0.588	0.275	0.364
	50%	**0.088**	**0.196**	0.108	0.228	0.229	0.347	0.113	0.239	0.299	0.383	0.228	0.331	0.210	0.319	0.187	0.312	0.581	0.597	0.273	0.361
	Avg	**0.070**	**0.174**	0.100	0.218	0.222	0.343	0.101	0.225	0.296	0.382	0.222	0.328	0.200	0.314	0.174	0.303	0.582	0.597	0.277	0.365
Weather	12.5%	0.028	**0.042**	0.027	0.051	0.036	0.095	**0.026**	0.047	0.140	0.220	0.037	0.093	0.031	0.076	0.037	0.072	0.176	0.287	0.296	0.379
	25%	0.030	**0.043**	**0.029**	0.056	0.042	0.104	0.030	0.054	0.147	0.229	0.042	0.100	0.035	0.082	0.038	0.074	0.187	0.293	0.327	0.409
	37.5%	**0.031**	**0.043**	0.033	0.062	0.047	0.112	0.032	0.060	0.156	0.240	0.049	0.111	0.040	0.091	0.039	0.078	0.172	0.281	0.406	0.463
	50%	**0.034**	**0.046**	0.037	0.068	0.054	0.123	0.037	0.067	0.164	0.249	0.053	0.114	0.046	0.099	0.042	0.082	0.195	0.303	0.431	0.483
	Avg	**0.031**	**0.043**	0.032	0.059	0.045	0.108	**0.031**	0.057	0.152	0.234	0.045	0.105	0.038	0.087	0.039	0.076	0.182	0.291	0.365	0.433
1st Count		**57**		1		0		2		0		0		0		0		0		0	

Table 17: Classification results on UEA datasets (Bagnall et al., 2018). Values are classification accuracy. (%)**Red**: best, Blue: second best.

Datasets / Models	Classical methods			RNN			TCN	Transformers										MLP			CNN	
	DTW (1994)	XGBoost (2016)	Rocket (2020)	LSTM (1997)	LSTNet (2018)	LSSL (2021)	TCN (2018)	Trans. (2017)	Re. (2020)	In. (2021)	Pyra. (2022b)	Auto. (2021)	Station. (2022c)	FED. (2022)	ETS. (2022)	Flow. (2022)	iTrans. (2024)	DLinear (2023)	LightTS (2022b)	TiDE (2023)	TimesNet (2023)	OSVIF (Ours)
EthanolConcentration	32.3	43.7	**45.2**	32.3	39.9	31.1	28.9	32.7	31.9	31.6	30.8	31.6	32.7	28.1	31.2	33.8	28.1	32.6	29.7	27.1	35.7	31.6
FaceDetection	52.9	63.3	64.7	57.7	65.7	66.7	52.8	67.3	**68.6**	67.0	65.7	68.4	68.0	66.0	66.3	67.6	66.3	68.0	67.5	65.3	**68.6**	67.1
Handwriting	28.6	15.8	**58.8**	15.2	25.8	24.6	53.3	32.0	27.4	32.8	29.4	36.7	31.6	28.0	32.5	33.8	24.2	27.0	26.1	23.2	32.1	42.9
Heartbeat	71.7	73.2	75.6	72.2	77.1	72.7	75.6	76.1	77.1	**80.5**	75.6	74.6	73.7	73.7	71.2	77.6	75.6	75.1	75.1	74.6	78.0	79.0
JapaneseVowels	94.9	86.5	96.2	79.7	98.1	98.4	98.9	98.7	97.8	98.9	98.4	96.2	**99.2**	98.4	95.9	98.9	96.6	96.2	96.2	95.6	98.4	97.3
PEMS-SF	71.1	**98.3**	75.1	39.9	86.7	86.1	68.8	82.1	82.7	81.5	83.2	82.7	87.3	80.9	86.0	83.8	87.9	75.1	88.4	86.9	89.6	89.2
SelfRegulationSCP1	77.7	84.6	90.8	68.9	84.0	90.8	84.6	92.2	90.4	90.1	88.1	84.0	89.4	88.7	89.6	**92.5**	90.2	87.3	89.8	89.2	91.8	91.1
SelfRegulationSCP2	53.9	48.9	53.3	46.6	52.8	52.2	55.6	53.9	56.7	53.3	53.3	50.6	57.2	54.4	55.0	56.1	54.4	50.5	51.1	53.4	57.2	**58.9**
SpokenArabicDigits	96.3	69.6	71.2	31.9	**100.0**	**100.0**	95.6	98.4	97.0	**100.0**	99.6	**100.0**	**100.0**	**100.0**	**100.0**	98.8	96.0	81.4	**100.0**	95.0	99.0	98.8
UWaveGestureLibrary	90.3	75.9	**94.4**	41.2	87.8	85.9	88.4	85.6	85.6	85.6	83.4	85.9	87.5	85.3	85.0	86.6	85.9	82.1	80.3	84.9	85.3	81.4
Average Accuracy	67.0	66.0	72.5	48.6	71.8	70.9	70.3	71.9	71.5	72.1	70.8	71.1	72.7	70.7	71.0	73.0	70.5	67.5	70.4	69.5	73.6	**73.7**

Table 18: Anomaly detection results on SMD, MSL, SMAP, SWaT, and PSM (Su et al., 2019; Hundman et al., 2018; Mathur & Tippenhauer, 2016; Abdulaal et al., 2021). Avg denotes the mean over the five datasets. **Red**: best, Blue: second best.

Method	MSL		PSM		SMAP		SMD		SWaT		Avg.	
Metric	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
OCSVM (2001)	0.5240	0.1530	0.6190	0.4180	0.3930	0.1020	0.6020	0.1040	0.6570	0.1700	0.5590	0.1894
IForest (2008)	0.5240	0.1140	0.5420	0.3340	0.4870	0.1220	0.6640	0.1220	0.3460	0.0930	0.5126	0.1570
LSTM-VAE (2020)	0.5212	0.0452	0.4915	0.4022	0.5083	0.0419	0.5005	0.0415	0.4959	0.0413	0.5035	0.1144
TranAD (2022)	0.5755	0.1038	0.6279	0.4300	0.4068	0.1020	0.7212	0.2575	0.3230	0.1260	0.5309	0.2039
Anomaly Transformer (2022)	0.5080	0.1070	0.5140	0.2980	0.5110	0.1300	0.5080	0.0460	0.4050	0.1210	0.4892	0.1404
TFAD (2022a)	0.5000	0.1060	0.5000	0.2790	0.5000	0.1280	0.5000	0.0420	0.5000	0.1230	0.5000	0.1356
DCdetector (2023)	0.5070	0.1070	0.5010	0.2810	0.5220	0.1320	0.5020	0.0430	0.5010	0.1210	0.5066	0.1368
D3R (2023a)	0.6525	0.1699	0.5003	0.2631	0.4135	0.1062	0.6420	0.1224	0.5665	0.1330	0.5550	0.1589
MEMTO (2023)	0.4999	0.1048	0.4975	0.2696	0.5959	0.1629	0.7324	0.1035	0.4541	0.1145	0.5560	0.1511
TimesNet (2023)	0.6130	0.1460	0.5920	0.3910	0.4530	0.1140	0.7270	0.1410	0.2880	0.1070	0.5346	0.1798
PatchTST (2023)	0.6370	0.1570	0.5860	0.3780	0.4480	0.1150	0.7360	0.1470	0.2420	0.0850	0.5298	0.1764
DualTF (2024)	0.5760	0.1560	0.6000	0.4110	0.4780	0.1220	0.6310	0.0690	0.5670	0.1430	0.5704	0.1802
iTransformer (2024)	0.6110	0.1510	0.5920	0.3830	0.4090	0.1140	0.7450	0.1460	0.2420	0.0840	0.5198	0.1756
UniTS-ST (2024)	0.6593	0.1651	0.5918	0.3832	0.4581	0.1223	0.7487	0.1749	0.2381	0.0830	0.5392	0.1857
TimeMixer (2024)	0.6551	0.1555	0.5962	0.3860	0.4501	0.1125	0.7498	0.1467	0.2525	0.1021	0.5407	0.1805
CATCH (2025b)	0.6640	0.1670	0.6520	0.4340	0.5040	0.1310	0.8110	0.1720	0.3450	0.1660	0.5952	0.2140
OSVIF (Ours)	0.5692	0.1384	0.6059	0.3879	0.5661	0.1470	0.6440	0.1096	0.8213	0.7283	**0.6413**	**0.3023**

G Visualization Analysis

To qualitatively evaluate model behavior, we visualize representative outputs on the test sets. Figure 8–Figure 10 show representative forecasting and imputation windows; Figures Figure 11 and Figure 12 present classification confusion matrices and anomaly detection ROC and PR curves computed on the full test sets; and Figure Figure 13 shows anomaly intervals selected by a fixed rule. Across these task types, OSVIF exhibits strong qualitative behavior.

G.1 Short-Term Forecasting Trajectories

Figure 8 shows representative node windows from PEMS04 and PEMS08 under $T = 96$ and $H = 12$. Both cases contain clear upward changes, making them useful for comparing how different models track short-term trends and local turning points. The prediction curves of OSVIF generally preserve the direction of change and the increase in magnitude of the ground truth. In particular, in the second half of the PEMS08 prediction window, OSVIF responds sufficiently to the rapid upward trend, whereas several baselines show more pronounced smoothing or amplitude shrinkage. This indicates that OSVIF better preserves local dynamics in the displayed windows.

G.2 Long-Term Forecasting Trajectories

Figure 9 compares Solar-Energy forecasting with $H = 96$ and Traffic forecasting with $H = 192$. The former contains rapid rise, peak, and decay patterns, while the latter exhibits multi-period fluctuations, reflecting the difficulty of preserving amplitude changes and periodic structure, respectively. On Solar-Energy, OSVIF recovers the main peak and the subsequent decline more effectively, whereas some baselines show insufficient peak magnitude or excessive smoothing. On Traffic, the phase and amplitude of OSVIF are overall stable, with the main deviations concentrated around local peaks and valleys.

G.3 Imputation Trajectories and Mask Recovery

Figure 10 shows imputation results for Weather with $M = 37.5\%$ and ETTm2 with $M = 12.5\%$. Red markers indicate masked positions, while the blue and orange curves denote the ground-truth and imputed sequences, respectively. On Weather, OSVIF better preserves high-value structures and boundary continuity within masked regions. On ETTm2, its imputed trajectory follows the main fluctuation range and level changes. In contrast, some baselines show amplitude compression, over-smoothing, or bias around local peaks.

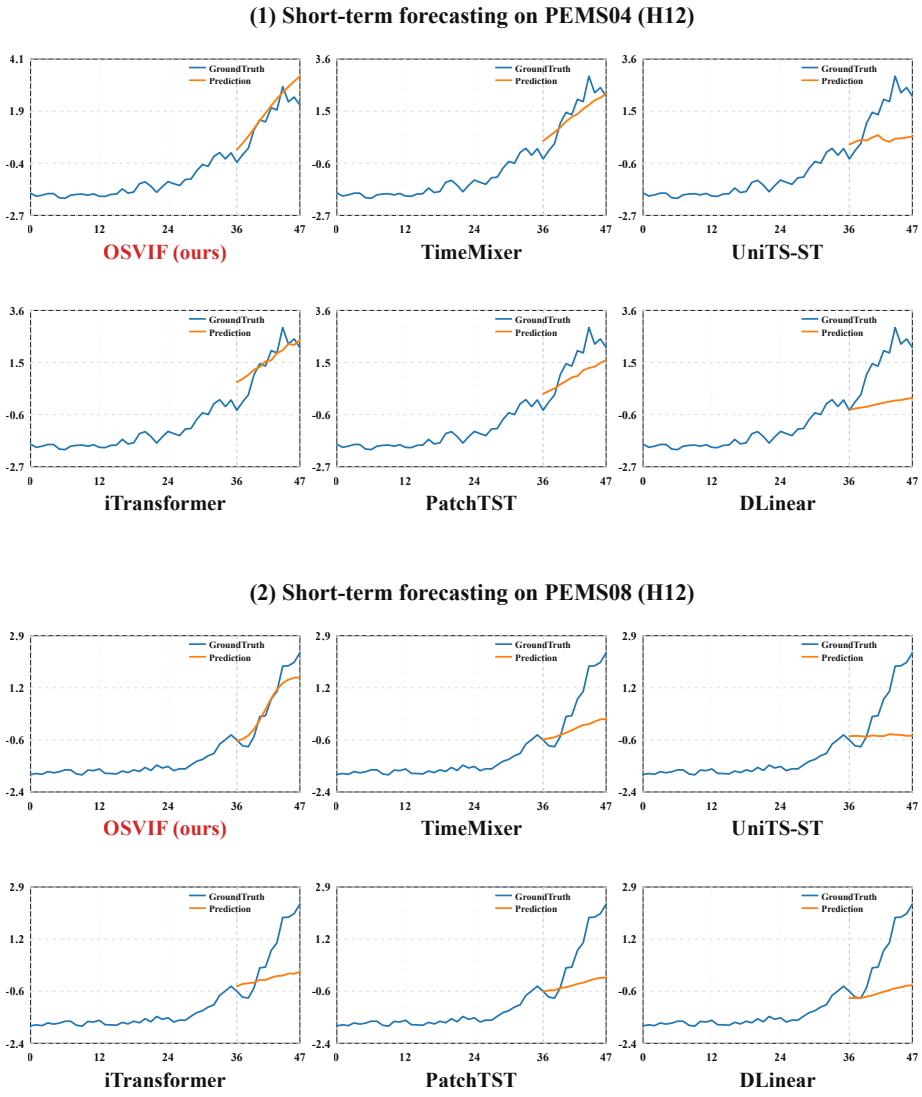

Figure 8: Visualization comparison for short-term forecasting. The upper half shows PEMS04 with input length 96 and prediction length 12; the lower half shows PEMS08 with input length 96 and prediction length 12.

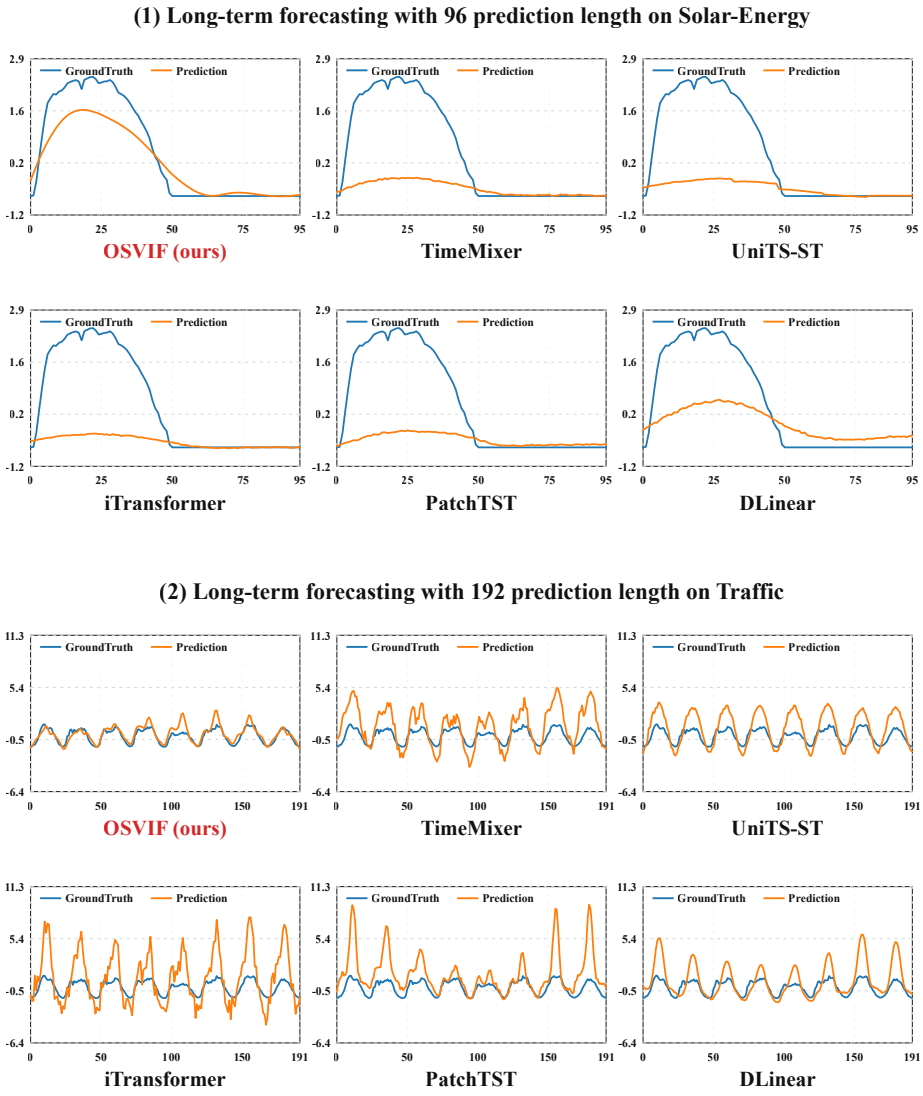

Figure 9: Visualization comparison for long-term forecasting. The upper half shows Solar-Energy with $H = 96$, and the lower half shows Traffic with $H = 192$; both use input length 96.

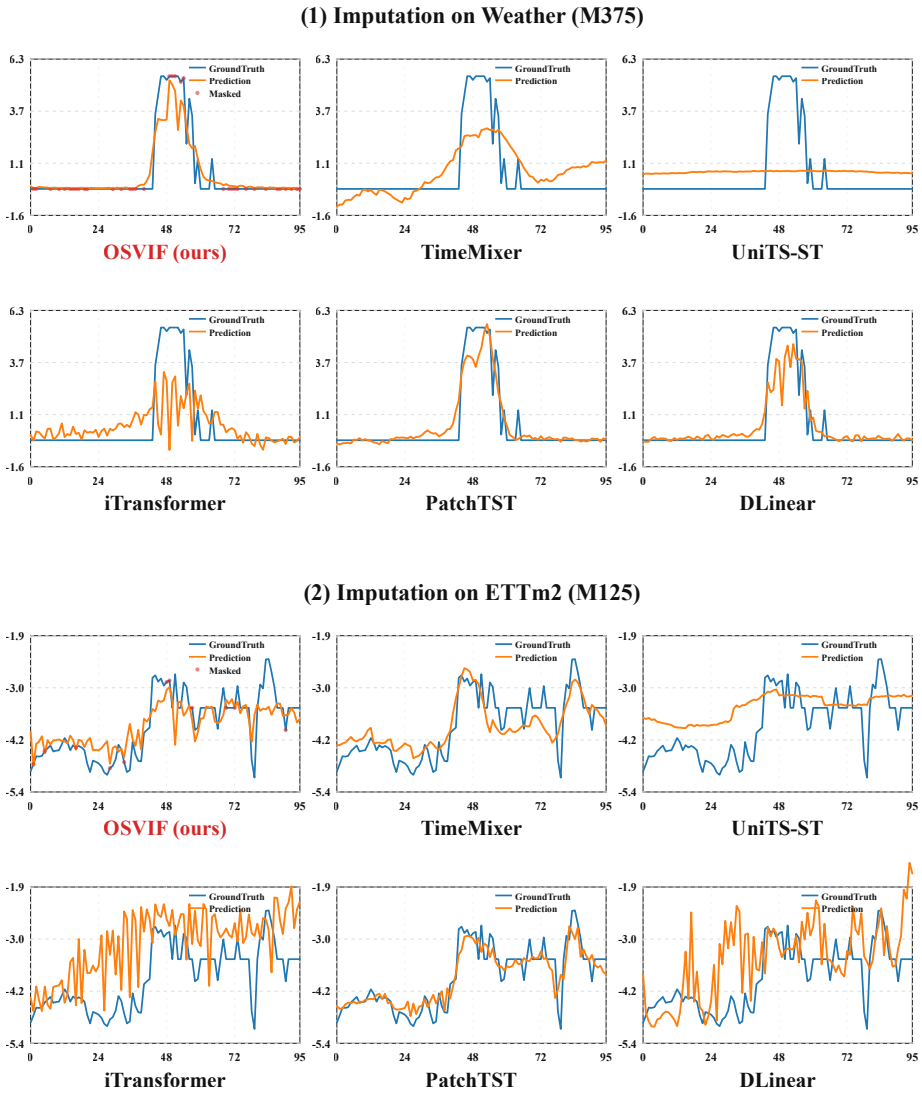

Figure 10: Visualization comparison for imputation. The upper half shows Weather with random masking ratio $M = 37.5\%$, and the lower half shows ETTm2 with random masking ratio $M = 12.5\%$. Red markers indicate masked positions.

G.4 Class Confusion Structure in Classification

Figure 11 reports confusion matrices on the full official test sets of JapaneseVowels and SelfRegulationSCP1. Rows denote ground-truth classes, columns denote predicted classes, and each cell reports both the sample count and the proportion normalized by the ground-truth class. The JapaneseVowels matrix shows a clear diagonal-dominant structure, with errors concentrated in a small number of class pairs; its Accuracy and Macro-F1 are 97.3% and 97.01%, respectively. For SelfRegulationSCP1, the two diagonal entries are 134 and 133, and both off-diagonal entries are 13, corresponding to Accuracy and Macro-F1 of approximately 91.13%, indicating relatively symmetric errors between the two classes.

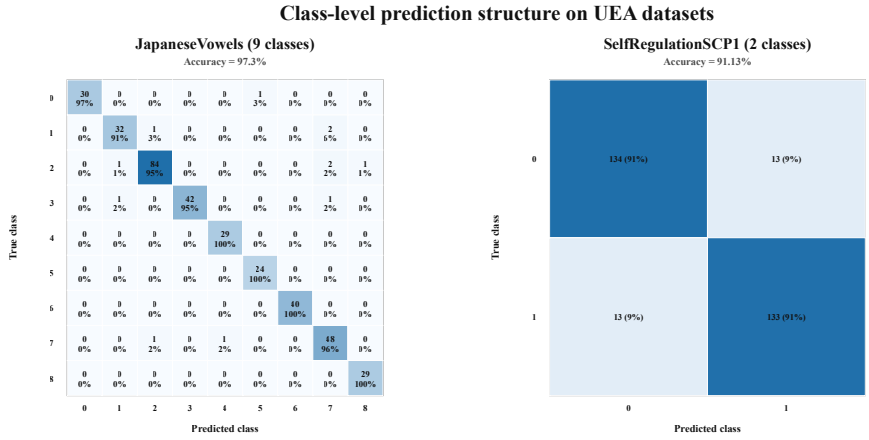

Figure 11: Class confusion structure for UEA classification tasks. Left: JapaneseVowels; right: SelfRegulationSCP1. Each cell reports both prediction count and the proportion normalized by the ground-truth class.

G.5 Threshold-Free Ranking Performance in Anomaly Detection

Figure 12 compares the ROC and PR curves of OSVIF, TimeMixer, UniTS-ST, iTransformer, and TimesNet on the full SWaT and PSM test sequences, using the saved seed-0 raw anomaly scores from each method. OSVIF uses the Value score, while baselines use their native anomaly scores. The zoomed views only change the displayed range and do not affect curve or metric computation. The gray dashed line in each PR plot denotes the anomaly ratio, which is 0.1214 for SWaT and 0.2778 for PSM.

On SWaT, OSVIF obtains AUROC/AUPRC of $0.8213/0.7283$, and its ROC and PR curves are overall better than those of the four baselines. On PSM, OSVIF obtains AUROC/AUPRC of $0.6059/0.3879$: its AUROC is slightly lower than iTransformer's 0.6086, but its AUPRC is higher than the other methods shown in the figure. Compared with SWaT, the curves on PSM are less separated across methods, indicating that anomaly-ranking performance is strongly dataset-dependent.

G.6 Temporal Response over Anomaly Intervals

Figure 13 shows representative anomaly intervals selected by a fixed rule from SWaT and PSM. Red shading denotes ground-truth anomaly intervals, while solid and dashed curves represent the Value score and Relation score, respectively. On SWaT, both scores change clearly after entering an anomaly interval, but their accumulation and recovery patterns differ. On PSM, the Value score mainly appears as sparse spikes, whereas the Relation score exhibits a more persistent interval response.

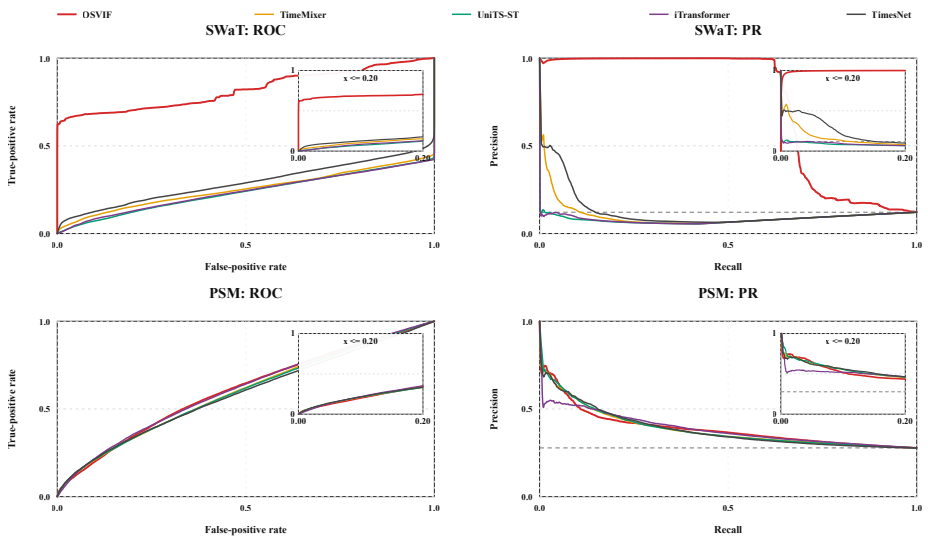

Figure 12: Threshold-free anomaly-ranking performance on SWaT and PSM. The left column shows ROC curves and the right column shows PR curves. OSVIF uses the Value score, while the comparison methods use their native anomaly scores. Each panel contains both the full coordinate range and a zoomed view of the low-horizontal-axis region; the gray dashed line in each PR plot indicates the anomaly-ratio reference.

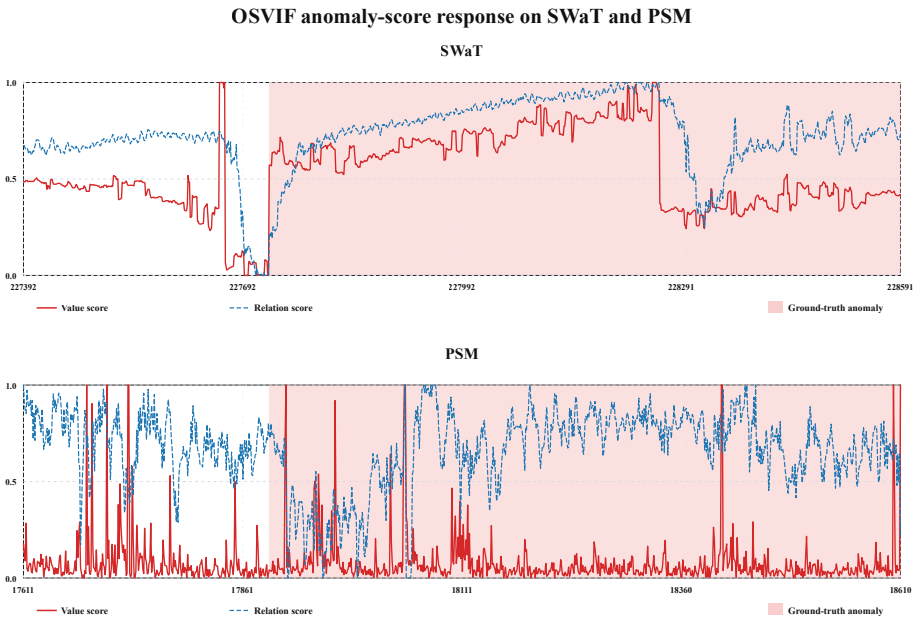

Figure 13: Temporal response of OSVIF over anomaly intervals on SWaT and PSM. Red shading denotes ground-truth anomaly intervals. The Value score is the score stream used in Table E5, while the Relation score is used for internal response diagnosis. The two scores are independently normalized within each panel only for visualization.