FAME: Adaptive Functional Attention with Expert Routing for Function-on-Function Regression

Yifei Gao

Industrial Engineering, Tsinghua University

Dec 3, 2025

Industrial Statistics and
Data Analytics Lab

CONTENTS

Introduction

Related work

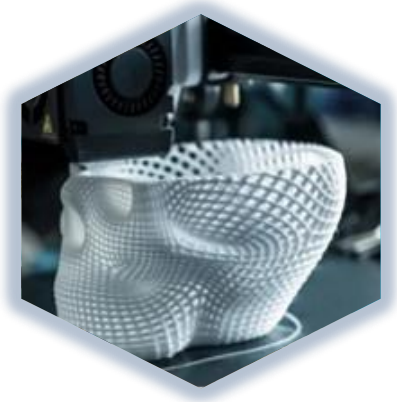
Problem Definition

Method

Experiments

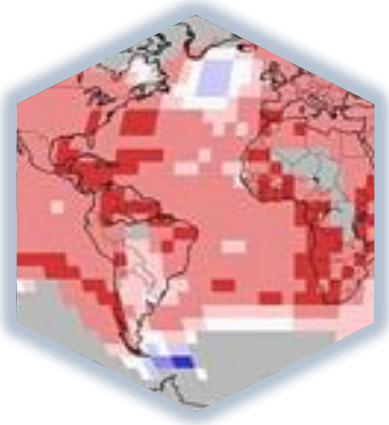
Conclusion

- Functional data are samples whose individual elements are random continuous functions, which now play a pivotal role across industries and the sciences.



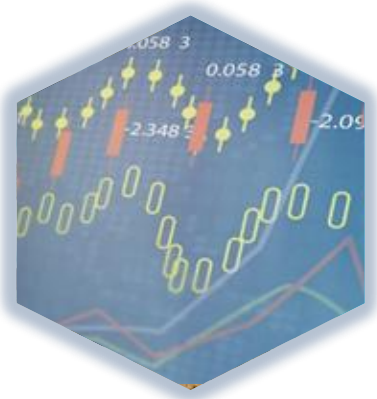
Healthcare

Environment



Manufacturing

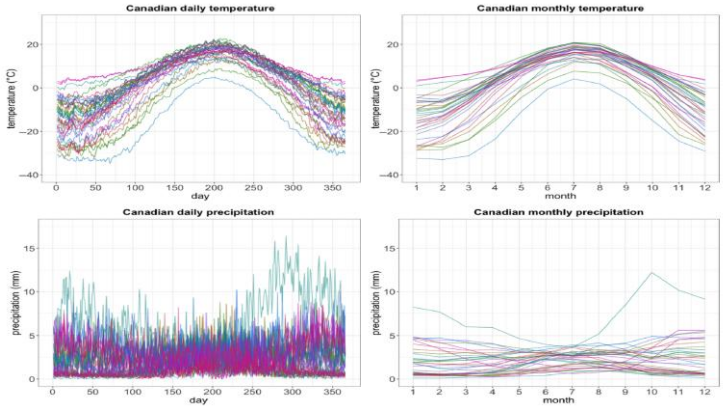
Finance

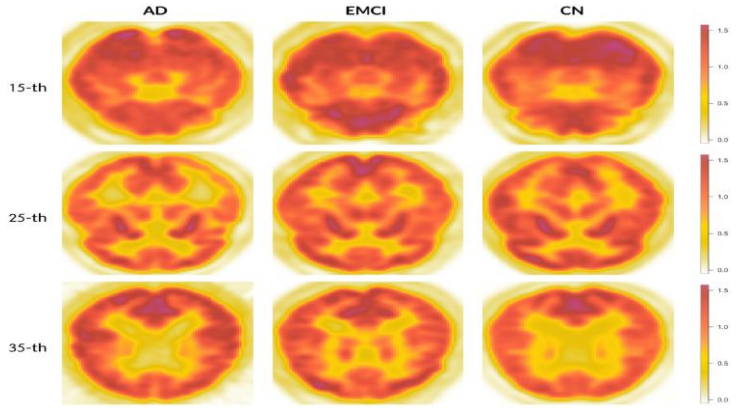


Transportation

Challenges

- Intra-functional continuity: each function lies in infinite-dimensional continuous functional space with features such as local dynamics and global trends;
- Inter-functional interactions: different dimensions of the input function can have nonlinear couplings with each other;
- Feature heterogeneity: different functions may exhibit vastly different properties such as scale, smoothness, or noise, and may even reside in different functional spaces.





Function-on-function regression

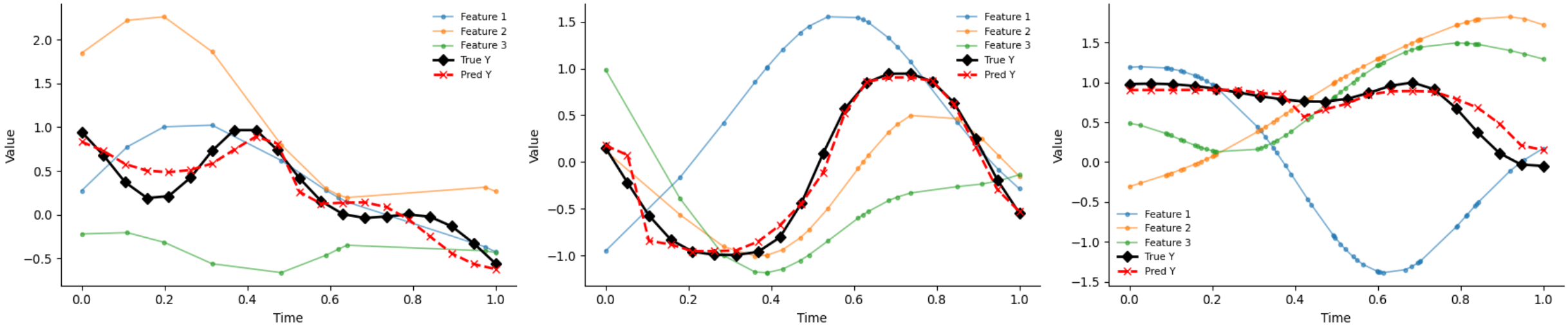
- Linear decomposition → Linear approaches employ finite bases such as FPCA, wavelets, or splines to project input and response curves into low-dimensional coordinates, enabling conventional regression techniques.
- Reproducing kernel Hilbert space → RKHS-based FoFR models nonlinear input→output operators in vector-/operator-valued RKHS via the representer theorem, supporting robust losses and estimators, but is kernel-sensitive and still scales cubically without low-rank approximations.
- Deep learning methods → Deep learning methods, including functional neural networks (FNNs) , FPCA-based neural networks, and adaptive basis expansions , offer data-driven flexibility.

We study function-on-function regression, where the goal is to learn an operator:

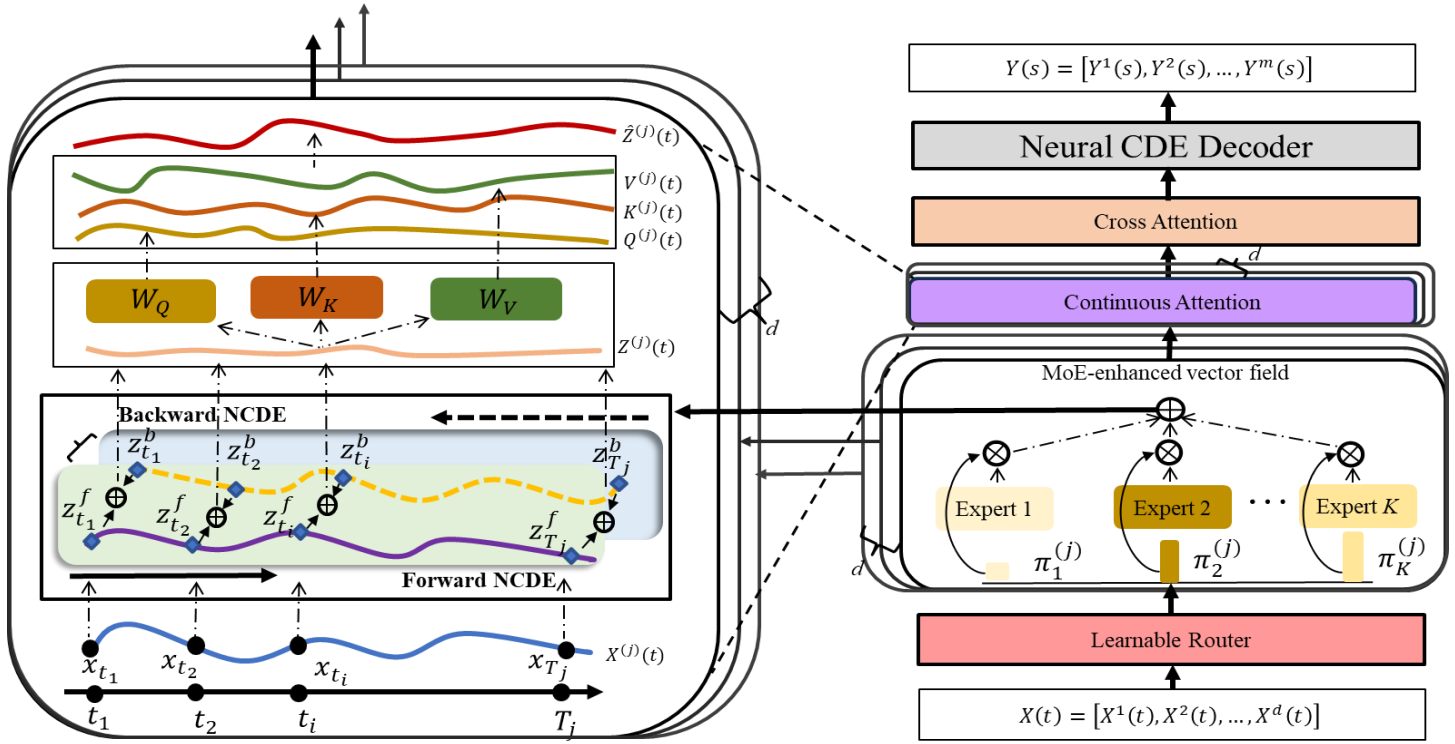
$$\mathcal{T} = \mathcal{X} \to \mathcal{Y}, \mathrm{Y} = \mathcal{T}(X),$$

that maps d input trajectories $X = \left(X^{(1)}, \ldots, X^{(d)}\right)$ to m continuous output trajectories $\mathrm{Y} = \left(Y^{(1)}, \ldots, Y^{(m)}\right)$.

In practice, we do not observe full trajectories X and Y; we only get a few irregular, non-synchronous samples from each function. Our goal is to learn a finite-parameter operator $\mathcal{T}_\theta \approx \mathcal{T}$ that predicts the entire output functions Y from these sparse samples.



The architecture of FAME



FAME is a continuous attention architecture for function-on-function regression. It (i) builds smooth query/key/value trajectories for each input function using a bidirectional NCDE, (ii) adapts to heterogeneous behaviors via a mixture-of-experts field, and (iii) fuses all functions through cross attention before decoding to continuous outputs.

Continuous bidirectional attention

We build a continuous representation for each input function $X^{(j)}$ that can see both past and future, and performs self-attention in continuous function.

Bidirectional NCDE :

$$Z_{\text{fwd}}^{(j)}(t) = Z^{(j)}(t_0) + \int_{t_0}^{t} f_{\theta_j^{\text{fwd}}}\left(Z_{\text{fwd}}^{(j)}(\tau)\right) dX^{(j)}(\tau),$$

$$Z_{\text{bwd}}^{(j)}(t) = Z^{(j)}(T_j) - \int_{t}^{T_j} f_{\theta_j^{\text{bwd}}}\left(Z_{\text{bwd}}^{(j)}(\tau)\right) d\tilde{X}^{(j)}(\tau).$$

Continuous self-attention over the function:

$$\hat{Z}^{j}(t) = \int_{t_0}^{T_j} \hat{\alpha}^{(j)}(t,\tau) V^{(j)}(\tau) d\tau \,,\; \hat{\alpha}^{(j)}(t,\tau) = \frac{\exp(\langle Q^{(j)}(t), K^{(j)}(\tau)\rangle / \sqrt{d_f})}{\int_{t_0}^{T_j} \exp(\langle Q^{(j)}(t), K^{(j)}(u)\rangle / \sqrt{d_f}) du}.$$

MoE-enhanced vector fields

Instead of a separate encoder per function, we use a shared pool of expert vector fields with a router that assigns function-specific mixture weights. Given experts $\{f_{\theta_k}\}_{k=1}^{K}$, the router outputs softmax weights $\pi_k^{(j)}$ and forms a customised field:

$$f_{\Theta}^{(j)}(z) = \sum_{k=1}^{K} \pi_k^{(j)} f_{\theta_k}(z), \Theta = \{\theta_1, \dots, \theta_K, \phi\}.$$

Cross-function fusion via cross-attention

After we build per-function continuous representations, we let functions talk to each other. For each function j, we attend over the other functions $\ell \in \{1, \dots, d\}$ to model their nonlinear couplings. The cross-attention weight from function j to function l at time t (for head p) is:

$$\hat{\beta}^{(j,l,p)}(t) = \frac{\exp(\langle Q^{(j,p)}(t), K^{(l,p)}(\tau) \rangle / \sqrt{d_c})}{\sum_{r=1}^{d} \exp(\langle Q^{(j,p)}(t), K^{(r,p)}(\tau) \rangle / \sqrt{d_c})}$$

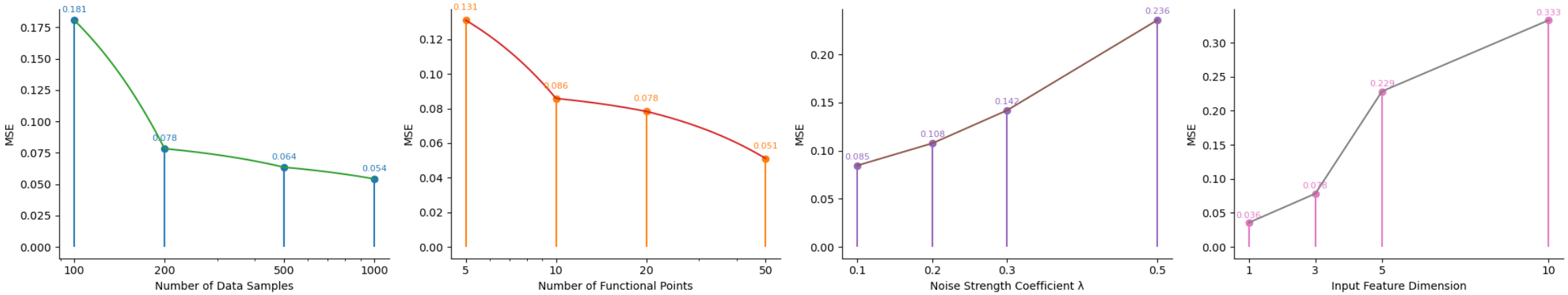
Experimental Results

Table 1: Average test MSE for different methods in regression. Detailed results (mean ± standard deviation) are provided in Appendix B. The best MSE for each case is highlighted in bold.

Model		Case 1			Case 2			Case 3		
		100	200	500	100	200	500	100	200	500
B-spline	Linear	0.4720	0.3947	0.3123	0.4135	0.3412	0.2822	0.4958	0.4002	0.3260
	Ridge	0.4264	0.3893	0.3117	0.3960	0.3393	0.2817	0.3869	0.3740	0.3222
	Lasso	0.4098	0.3830	0.3052	0.3856	0.3351	0.2766	0.4132	0.3584	0.3188
	Elastic Net	0.4510	0.3650	0.2874	0.3725	0.3152	0.2583	0.3986	0.3618	0.3190
Fourier	Linear	0.5002	0.4092	0.3224	0.4923	0.3636	0.2896	0.4762	0.3921	0.3547
	Ridge	0.4255	0.3780	0.3149	0.3616	0.3343	0.2841	0.3755	0.3568	0.3328
	Lasso	0.3550	0.3493	0.3135	0.3361	0.3280	0.3001	0.3808	0.3560	0.3247
	Elastic Net	0.3540	0.3325	0.2914	0.3167	0.3070	0.2720	0.3737	0.3496	0.3366
FPCA		0.3717	0.3554	0.3200	0.2890	0.2733	0.2624	0.3812	0.3563	0.3295
Kernel Method		0.2441	0.1728	0.1058	0.1741	0.0923	0.0700	0.2654	0.2445	0.1485
Gaussian Process		0.3405	0.2941	0.2036	0.3905	0.3917	0.2588	0.3031	0.2926	0.3945
FNN		0.3123	0.2083	0.1013	0.1941	0.1142	0.0811	0.3678	0.3571	0.1366
FAME w/o Bi-dir		0.1832	0.0812	0.0654	0.1530	0.0528	0.0355	0.1919	0.0813	0.0368
FAME w/o MoE		0.1870	0.0828	0.0663	0.1578	0.0538	0.0362	0.1972	0.0856	0.0374
FAME w/o Cross-attn		0.1902	0.0815	0.0668	0.1602	0.0544	0.0375	0.1997	0.0879	0.0381
FAME		**0.1806**	**0.0783**	**0.0635**	**0.1532**	**0.0511**	**0.0342**	**0.1954**	**0.0796**	**0.0352**

FAME attains the lowest test MSE across benchmarks.

Experimental Results



(a) Test MSE by sample size (b) Test MSE by sampling density (c) Test MSE by noise strength (d) Test MSE by input dimensionality

Across parameter sweeps, test error decreases with larger sample size and accuracy improves with more sampling points—even at very sparse settings—while performance remains stable under mixed/irregular resolutions, indicating sampling invariance; as noise and dimensionality increase, errors rise for all methods but FAME stays within a practical range, revealing a systematic dependence on data characteristics and aligning with the theoretical guarantees.

Experimental Results

Table 2: Average test set MSE in regression under simulation. Basis Expansion (best) shows the best result among the 8 basis expansion methods presented in Table 1.

Model	case 4	case 5			case 6	case 7		case 8
		0.1	0.2	0.3		5	10	
Basis Expansion(best)	0.3610	0.3665	0.3944	0.4419	0.3669	0.4501	0.4705	0.5204
FPCA	0.3802	0.3879	0.3956	0.4570	0.3844	0.5284	0.5573	0.5737
Kernel Method	0.1928	0.1440	0.1919	0.2435	0.2022	0.4659	0.5236	0.5895
Gaussian Process	0.2302	0.3079	0.3356	0.3830	0.3434	0.4120	0.4498	0.4762
FNN	0.2102	0.1879	0.2250	0.3438	0.1744	0.4801	0.5164	0.5384
FAME	**0.0798**	**0.0846**	**0.1076**	**0.1420**	**0.0824**	**0.2285**	**0.3330**	**0.3530**

FAME consistently maintains superior predictive performance across stress tests varying feature heterogeneity, noise, output structure, and input dimensionality.

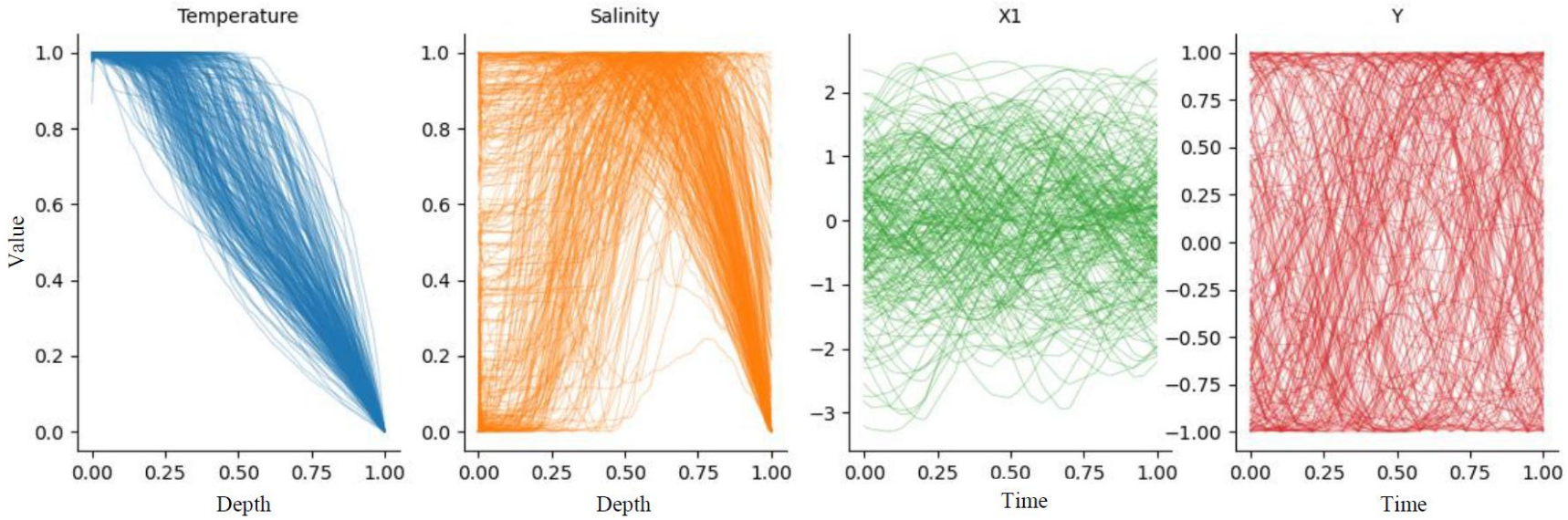
Experimental Results

Table 3: Average test set MSE in regression on different real-world datasets.

Model	Hawaii ocean		Human3.6M			ETDataset
	Salinity	Temp	Walking	Eating	Sitting down	Oil Temp
Basis Expansion(best)	0.0780	0.0014	0.0359	0.04841	0.0122	0.0365
FPCA	0.0865	0.0025	0.0373	0.0099	0.0121	0.0355
Kernel Method	0.0754	0.0025	0.0373	0.0099	0.0121	0.0355
Gaussian Process	0.0931	0.0022	0.0360	0.0075	0.0107	0.0380
FNN	0.0766	0.0020	0.0373	0.0099	0.0121	0.0355
FAME w/o Bi-dir	0.0751	0.0012	0.0327	0.0035	0.0071	0.0264
FAME w/o MoE	0.0759	0.0013	0.0332	0.0038	0.0075	0.0271
FAME w/o Cross-attn	0.0773	0.0014	0.0344	0.0044	0.0083	0.0286
FAME	**0.0748**	**0.0012**	**0.0325**	**0.0034**	**0.0070**	**0.0262**

FAME also achieves superior performance on real-world datasets.

Experimental Results



(a) Real data functional features (b) Synthetic data functional features

On the real-world dataset (smooth, low variability), fixed-basis and kernel methods already perform well; on the synthetic datasets (broader, irregular trajectories), FAME achieves substantially larger gains—i.e., it excels as function-space complexity increases.

- FAME is the first FoFR model to operate directly on irregularly sampled functional space—no predefined bases or grids—with effectiveness backed by theory and extensive experiments.
- FAME introduces a functional attention mechanism: continuous attention via bidirectional NCDEs (intra-functional continuity) and multi-head cross attention (inter-functional interactions).
- FAME augments a mixture-of-experts (MoE) architecture for adaptive heterogeneity modeling and uses an NCDE decoder to produce continuous outputs at arbitrary query locations, naturally handling misaligned targets.

Industrial Statistics and Data Analytics Lab

Thanks!