INTERPRETABLE TRANSFORMER REGRESSION FOR FUNCTIONAL AND LONGITUDINAL COVARIATES

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider scalar-on-function prediction from functional covariates that may be measured sparsely and irregularly over time with noise, which is common in longitudinal studies. We propose a dual-attention Transformer that operates on a discretized time grid with missing-value masks and trains end-to-end without any imputation. The model couples time-point attention, which encodes local and long-range temporal structure, with inter-sample attention, which shares information across similar subjects. We derive prediction error bounds and prove consistency under a random-effects framework that accommodates sparse/irregular sampling, measurement error, and label noise. In simulations across varying sparsity levels, our method outperforms 19 strong baselines (ensemble, statistical/functional, deep learning methods, tabular Transformers, and pre-trained models such as TabPFN) in regimes with $\leq 50\%$ observations and remains competitive in denser settings, highlighting the importance of end-to-end missingness-aware modeling. The learned attention weights are interpretable, revealing predictive time windows and cluster structure. In real-world data, our approach achieves the best prediction and classification performance, surpassing leading imputation methods paired with competitive learners. This underscores that explicitly modeling sparsity is preferable. In summary, the dual-attention mechanism is interpretable, consistently identifying predictive time windows and cohort clusters that align with domain knowledge. The proposed Transformer also outperforms state-of-the-art methods while preserving robustness and interpretability.

1 INTRODUCTION

Longitudinal data consist of repeated measurements on the same subjects over time, often coupled with time-varying covariates and subject-specific heterogeneity (Diggle et al., 2002; Fitzmaurice et al., 2004; 2009). Such data are ubiquitous in biomedicine, environmental monitoring, and digital health, where dynamic trajectories inform prognosis, treatment response, and risk stratification. It is common to assume that the underneath dynamic trajectory of each subject is a smooth function of time, while the observed longitudinal measurements may be noisy and measured at irregular and subject-specific time points. This adopts a perspective in the field of functional data analysis (FDA) (Wang et al., 2016; Ramsay & Silverman, 2005), where each trajectory is a realization of a latent smooth process observed with noise. When the sampling plan is intensive, a nonparametric approach is typically deployed to model such functional data; whereas a parametric approach, such as a mixed-effect model (Laird & Ware, 1982b) was the norm to model sparse or irregularly sampled functional data, until a nonparametric approach was proposed in (Yao et al., 2005a). Since longitudinal data are discretely sampled functional data (possibly with noise/measurement error), we aim to develop a unified nonparametric approach that handles a broad range of sampling schemes, whether they are intensive, sparse, or in between, including irregular and subject-specific time schedules.

We address the scalar-on-function regression problem, where a functional covariate, possibly observed on a sparse and irregular time schedule with noise, is used to predict a scalar outcome. The method must learn an unspecified functional of the whole trajectories without restrictive parametric forms. Since accurate scalar prediction hinges on correctly capturing the trajectory–outcome relationship in the presence of missingness and noise, an effective model should therefore (i) accommodate irregular sampling and missing data, (ii) encode temporal order and dependencies, (iii)

borrow strength across similar subjects, and (iv) learn the trajectory–outcome relationship without restrictive parametric assumptions.

Given that sparse and irregular longitudinal data are the most challenging type of functional data, we introduce the **I**nterpretable **D**ual-**A**ttention **T**ransformer (IDAT), a unified architecture inspired by Transformers (Vaswani et al., 2017) that is tailored to handle high missingness and irregular sampling. Meanwhile, this approach is broadly applicable to any type of functional/longitudinal data. In contrast to "sparse Transformers" that impose artificially designed sparse attention patterns for computational efficiency (Jaszczur et al., 2021; Correia et al., 2019), our notion of "sparsity" refers to missing-data sparsity arising from the sampling scheme. We discretize the time interval into a working grid and use explicit missing-value masks to encode each subject's observation pattern. The model combines (i) time-point attention, which serves as a data-adaptive functional encoder capturing smooth trajectory structure, with (ii) inter-sample attention, which learns nearest-neighbor–like weights across subjects via a learned similarity metric. A regressor layer then summarizes the learned representation to deliver end-to-end scalar-on-function prediction.

1.1 Main Contributions

End-to-end scalar-on-function regressors. We propose a dual-attention Transformer that predicts a scalar outcome from a sparse, irregular longitudinal trajectory in an end-to-end manner. The model uses explicit missing-value masks under supervised setup in the training stage, avoiding ad-hoc imputation while respecting each subject's observation pattern. The architecture jointly captures within-trajectory structure and cross-subject borrowing, yielding a trainable pipeline without restrictive parametric assumptions.

Interpretable dual-attention Transformer. The learned attention weights act as nonparametric regression coefficients along two axes. Time-point attention acts as a data-adaptive smoother, aggregating local and long-range temporal information to encode functional structure. Inter-sample attention implements learned and convex pooling over subjects, akin to nearest neighbors with a learned metric, which stabilizes predictions for missing and noisy data. Under masking, attention conditions on the observed entries and propagates signal about Y, effectively behaving as an informative weighting mechanism rather than simple imputation. Empirically, attention maps reveal domain-relevant windows and cluster structure, aiding interpretability.

Theoretical and numerical justification. We derive prediction error bounds and show consistency of the training and testing phases MSE. Extensive simulation and real data studies across a wide range of sparsity demonstrate robustness to high missingness and superior accuracy relative to 19 baselines (ensemble methods, statistical/functional models, deep learning methods, tabular Transformers, and pre-trained models TabPFN).

1.2 Related Work

Modeling longitudinal data, defined as repeated measurements over time that are often irregularly sampled and of varying length, poses challenges for representation learning and prediction. In biomedical settings (e.g., EHR), recent surveys document a rapid expansion of machine learning and deep learning approaches (Cascarano et al., 2023; Carrasco-Ribelles et al., 2023). Early neural models typically flatten a temporal history into fixed feature vectors for feedforward networks (e.g., cardiovascular risk prediction (Zhao et al., 2019)), thereby discarding ordering information. To retain functional structure, Yao et al. (2021) propose a basis-learning layer in which hidden units act as adaptive basis functions, enabling end-to-end, task-specific trajectory expansions for fully observed functional data.

Convolutional neural networks (CNNs) capture local temporal structure via 1D convolutions and are competitive for time-series classification (Wang et al., 2017), but modeling long-range dependencies often requires very deep networks or large receptive fields. Recurrent architectures (RNNs/LSTMs) maintain hidden states that aggregate past information and naturally handle variable sequence lengths and missingness patterns, with applications in clinical prognosis (e.g., Alzheimer's disease (Cui et al., 2019; Aghili et al., 2018)) and broader EHR modeling (Lipton et al., 2016). Nonetheless, their inherently sequential computation can be a bottleneck for long sequences.

Transformers (Vaswani et al., 2017) replace the sequential recurrence approach in CNN, RNN, and LSTM by "self-attention" to relate all time points within a sequence, capturing both local and global dependencies and supporting parallel computation. Empirically, reviews report strong performance on longitudinal biomedical tasks (Siebra et al., 2024). Early EHR applications such as BEHRT encoded patient histories as sequences of medical concepts to learn contextualized representations for downstream prediction (Li et al., 2020), while general frameworks demonstrated effectiveness across multivariate time-series classification and regression (Zerveas et al., 2021). The architecture has also been adapted to domain-specific objectives, including survival modeling (Öğretir et al., 2024; Zhang et al., 2025). Beyond prediction, specialized self-attention modules have been proposed for functional data imputation: SAND (Hong et al., 2024) introduces attention weights on derivatives to promote smooth reconstructions under irregular sampling. More broadly, efficient attention variants (e.g., sparse or kernelized forms (Jaszczur et al., 2021; Lou et al., 2024; Correia et al., 2019; Chen et al., 2023)) have been explored to mitigate quadratic time/memory costs on long sequences, though their approximation properties in sparse/irregular regimes require careful validation.

2 Method and Model

The proposed architecture in Figure 1 is designed to handle sparse, irregular longitudinal inputs and is applicable to all longitudinal and functional data. The top panel shows two subjects measured at only a few subject-specific times (blue circles and light-blue crosses) with additional measurement errors. Because the full latent trajectories are infinite-dimensional, we discretize the time domain into grid points, then inject positional encodings and apply explicit masks to tokens for unobserved grid points. The resulting input is a tabular form, whose columns index time grid locations, so each column at a time grid acts as a feature, while positional encodings preserve temporal ordering and proximity. The key difference from standard tabular data is the substantial missingness induced by sparse and irregular sampling and strong temporal dependence between adjacent features.

The proposed method, IDAT, is a dual-attention (encoder-only) Transformer for scalar-on-function regression. *Time–point attention* operates along each subject's time grid to learn both local and long-range temporal dependencies, encoding trajectory structure while respecting order and smoothness. Unlike imputation methods that reconstruct the entire trajectory without using the outcome, our model is trained with a Y-token: during training, attention flows between covariate tokens and the response, providing supervision that turns time-point attention into a nonparametric weighting scheme over time (low weights mark uninformative windows). At test time, the Y-token is masked, so predictions rely solely on observed covariates but still benefit from the supervision-shaped representation learned in training. *Inter–sample attention* acts across the mini-batch at each time grid, assigning data-adaptive weights to similar subjects to share information between subjects. In practice, these weights reveal the cluster structure when it exists. With underlying smooth latent trajectories, both attentions yield smooth reconstructions. Theoretical details are provided in the Appendix.

This modular design, which separates time-point and inter-sample attention, allows us to quantify the within-trajectory encoding and across-subjects contributions respectively. The model prioritizes prediction over imputation, learning attention weights end-to-end for regression rather than reconstructing trajectories. Empirically, the model outperforms state-of-the-art methods under $\leq 50\%$ observed data and irregular sampling, remains robust to measurement error, and performs competitively in denser regimes. Its strong performance spans a wide range of sampling schemes, enabling real-world applicability.

2.1 Model Setup

Without loss of generality, we assume all subjects have trajectories in the time interval $I = [0, 1]$ with latent smooth trajectories $X_i(\cdot)$, under contamination of measurement errors and irregular sampling scheme, the observation times for subject i is $\tilde{t}_i = (t_{i1}, \ldots, t_{i,n_i}) \subset I$, and we observe

$$X_i^*(\tilde{t}_i) = X_i(\tilde{t}_i) + \eta_i(\tilde{t}_i), \qquad \eta_i(\tilde{t}_i) \overset{iid}{\sim} N(0, \sigma_X^2). \tag{1}$$

The scalar response is generated from the functional regression model with an unspecified $\mathcal{F}$,

$$Y_i = \mathcal{F}\big(X_i(\cdot)\big) + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma_Y^2). \tag{2}$$

We refer to ϵ_i as the label noise on Y_i, in contrast to the measurement noise η_i on X_i.

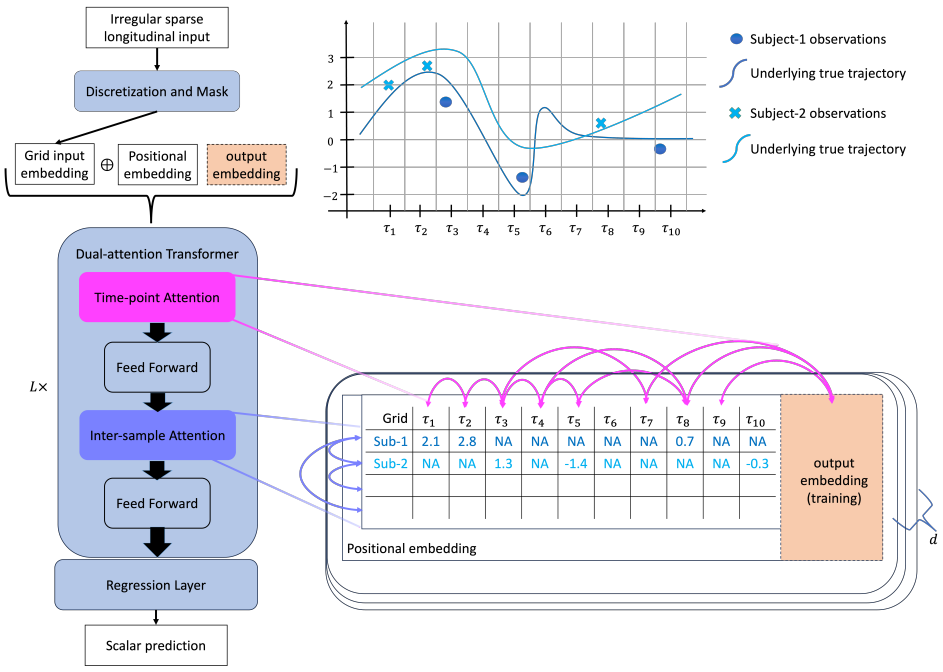


Figure 1: IDAT Model Architecture.

Discretization, masking and positional encoding. Despite irregular measurement times and varying n_i across subjects, we align all observations to a shared fixed grid $\tilde{\tau} = (\tau_1, \ldots, \tau_T) \subset I$,

$$M_i(\tau_j) = \mathbf{1}\{\tau_j \in \tilde{t}_i\}, \qquad X_i^*(\tau_j) = X_i^*(t_{i,k})\mathbf{1}\{\tau_j = t_{i,k}\}, \tag{3}$$

and form the length-$T+1$ vector $D_i = (X_i^*(\tilde{\tau}) \odot M_i(\tilde{\tau}),\, Y_i) \in \mathbb{R}^{T+1}$ with a mesh size defined by $\Delta = \max_j |\tau_{j+1} - \tau_j|$. Linear embeddings E_X and E_Y from $\mathbb{R}$ to $\mathbb{R}^d$ are applied token–wise along with sinusoidal positional encodings $P(\tilde{\tau})$,

$$\tilde{D}_i = \Big[E_X\big(X_i^*(\tilde{\tau}) \odot M_i(\tilde{\tau})\big) + P(\tilde{\tau}),\; E_Y(Y_i)\Big] \in \mathbb{R}^{d\times(T+1)}. \tag{4}$$

Dual–attention Transformer. Given a batch $\tilde{\mathbf{D}}$ of size B, the dual-attention Transformer block is

$$\mathcal{TB} = \mathrm{FF}_1 \circ A_I \circ \mathrm{FF}_2 \circ A_T, \tag{5}$$

omitting normalization layers for brevity, where A_T (time-point attention) applies attention along the temporal axis within each sample, A_I (inter-sample attention) applies attention across samples in the batch (enabling cross-sample borrowing), and $\mathrm{FF}_1, \mathrm{FF}_2$ are position-wise two-layer ReLU MLPs. Stacking L dual-attention Transformer blocks yields the embedding $\mathcal{T} = \mathcal{TB}^{\circ L}$.

Regression layer. Given the output of dual–attention Transformer $\mathcal{T}(\tilde{D}_i) = \big[\, Z_X \;\big|\; Z_Y \,\big]_i \in \mathbb{R}^{d\times(T+1)}$, we write $Z_X = \mathcal{T}(\tilde{D}_i)_{:,:,1:T} \in \mathbb{R}^{d\times T}$ for the longitudinal covariate embeddings and $Z_Y = \mathcal{T}(\tilde{D}_i)_{:,:,T+1} \in \mathbb{R}^d$ for the Y-token embedding. Each sequence Z_X is summarized to a single d-vector via a trainable pooling map $\phi : \mathbb{R}^{d\times T} \to \mathbb{R}^d$. The pooled representation is then mapped to a scalar by a MLP g (e.g., two-layer ReLU), yielding the prediction

$$\hat{Y}_i \;=\; g\left(\phi\big([Z_X]_i\big)\right). \tag{6}$$

During training, we minimize the loss function $\ell\big(\hat{Y}_i,\, g([Z_Y]_i)\big)$ (e.g., MSE for regression), treating the response embedding Z_Y as an informative target. At test time, the Y token is masked (set to zero) prior to encoding, so predictions depend solely on the covariate sequence.

3 KEY THEORETICAL RESULTS

Let $\tilde{\mathbf{D}}$ denote the discretized, noisy and masked embedding, and $\tilde{\mathbf{S}}$ the oracle embedding. The input–embedding discrepancy (discretization error) is controlled by $\delta_0 = \|\tilde{\mathbf{D}} - \tilde{\mathbf{S}}\|_\infty$ (Lemma 3). In training, the dual–attention block components, including time–point self–attention A_T, feed–forward networks $\mathrm{FF}_1, \mathrm{FF}_2$, and inter–sample attention A_I, are Lipschitz with constants $L_T, L_{\mathrm{FF}}, L_I$ (Lemmas 4, 8) and admit uniform approximation on compact sets (Lemmas 5, 6, 7), giving deterministic approximation errors $\varepsilon_{A_T}, \varepsilon_{\mathrm{FF}}, \varepsilon_{A_I}$. Hence, a dual-attention block $\mathcal{TB}$ is $L_{\mathcal{TB}}$-Lipschitz (Lemma 8). Inter–sample attention further reduces a stochastic embedding error $\varepsilon_{\mathrm{var}}$ by up to a $B^{-1/2}$ factor (Lemma 9) with batch size B.

Except for Theorems 1 and 2 below, all the lemmas and other theorems are listed in the Appendix.

Theorem 1 (Consistency of $\mathcal{T}$)**.** *Under the assumptions and notations of Lemmas 3-9, the oracle mapping is $H(\tilde{\mathbf{S}}) = G \circ f_{\mathrm{I}} \circ G \circ f_{\mathrm{T}}(\tilde{\mathbf{S}})$. Let*

$$\varepsilon_{\mathcal{TB}} := L_{\mathrm{FF}}\big(\varepsilon_{\mathrm{var}} + \varepsilon_{A_I} + L_I\big(\varepsilon_{\mathrm{FF}} + L_{\mathrm{FF}}\varepsilon_{A_T}\big)\big) + \varepsilon_{\mathrm{FF}}.$$

Then, with probability at least $1-\delta$,

$$\big\|\mathcal{TB}(\tilde{\mathbf{D}}) - H(\tilde{\mathbf{S}})\big\|_\infty \le \varepsilon_{\mathcal{TB}}, \qquad \big\|\mathcal{T}(\tilde{\mathbf{D}}) - H(\tilde{\mathbf{S}})\big\|_\infty \le L_{\mathcal{TB}}^L\,(\delta_0 + \varepsilon_{\mathcal{TB}}) := \varepsilon_{\mathcal{T}}. \quad (7)$$

In particular, if (i) the mesh shrinks $\Delta \to 0$ so that $\delta_0 \to 0$; (ii) the dual-attention Transformer has sufficient capacity so that $\varepsilon_{\mathrm{FF}}, \varepsilon_{A_T}, \varepsilon_{A_I}, \varepsilon_{\mathrm{var}} \to 0$; and (iii) the block Lipschitz constant is uniformly bounded with training size n and batch size B, i.e., $\sup_{n,B} L_{\mathcal{TB}}(n,B) < \infty$ for fixed depth L, then

$$\big\|\mathcal{T}(\tilde{\mathbf{D}}) - H(\tilde{\mathbf{S}})\big\|_\infty \xrightarrow{\mathbb{P}} 0,$$

i.e., the dual-attention Transformer $\mathcal{T}$ is consistent for the oracle mapping H.

Generalization bounds for $\mathcal{TB}$ and $\mathcal{T}$ are given in Lemmas 10 and 11. Let $S(U) = U_{:,:,1:T}$ denote the covariate-token slice (excluding the Y-token), define the predictor with L_ϕ-Lipschitz pooling function ϕ and L_g-Lipschitz MLP,

$$\widehat{Y}^{(L)}(\tilde{D}) = g\left(\phi(S[\mathcal{T}(\tilde{D})])\right).$$

Theorem 2 (Training MSE generalization and consistency)**.** *Let $\mathcal{T} = \mathcal{TB}^{\circ L}$ be the L-block encoder, and set $p := B\,d\,(T+1)$. Assume (i) Boundedness: $\|\tilde{D}\|_\infty \le R_{\mathrm{in}}$, $|\widehat{Y}^{(L)}(\tilde{D})| \le R_{\mathrm{out}}$, and $|Y| \le M_f$ almost surely with $L_\ell := 2(R_{\mathrm{out}} + M_f)$. (ii) Optimization: a stable SGD regime in which the empirical risk approaches its minimum (within the hypothesis class) with high probability (Hardt et al., 2016). (iii) $\sup_n L_{\mathcal{T}}(n) < \infty$ for fixed depth L and (iv) $p/n \to 0$. Let $\mathrm{MSE}_n^{\mathrm{train}}$ be the training MSE over n samples. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\mathrm{MSE}_n^{\mathrm{train}} - \mathbb{E}\big(\widehat{Y}^{(L)}(\tilde{D}) - Y\big)^2\right| \le 2\,L_\ell\,L_g\,L_\phi\,L_{\mathcal{T}}\,R_{\mathrm{in}}\sqrt{\tfrac{2p}{n}} + 3\sqrt{\tfrac{\ln(2/\delta)}{2n}}.$$

Moreover, if the regressor approximation error vanishes as capacity grows (Hornik, 1991; Stinchcombe, 1999; Cybenko, 1989; Hornik et al., 1989; Yarotsky, 2017) and $\mathcal{T}$ is consistent (Theorem 1), the population training MSE is consistent.

At test time, with N independent samples, the Y-token is masked so that

$$\mathbb{E}\big|\widehat{Y}^*(\tilde{D}^*) - \widehat{Y}^{(L)}(\tilde{D})\big| \le L_\ell L_g L_\phi L_{\mathcal{T}} L_Y\,\mathbb{E}|Y|.$$

Lemma 12 quantifies this testing phase Y-token perturbation. Theorem 13 further bounds the pointwise and uniform test errors via a decomposition (embedding, head approximation, discretization, label noise), and controls the population test MSE by the training MSE plus a Rademacher term and an expectation bridge; the empirical test MSE adds an extra concentration term. If the training MSE is consistent and the Y-token embedding is scaled so that $L_Y \to 0$, the bridge term vanishes then both population and empirical test MSE are consistent asymptotically (Corollary 14).

Mesh–size trade-off. For an α–Hölder trajectory (Lemma 3), the grid discretization bias $\|\tilde{\mathbf{D}} - \tilde{\mathbf{S}}\|_\infty$ scales as $O(\Delta^\alpha)$, with $\Delta = \max_j |\tau_{j+1} - \tau_j|$. Since $T \asymp \Delta^{-1}$, time-point attention incurs $O(T^2 d)$ time and $O(T^2)$ memory, so halving Δ quadruples computational cost, while improving the bias by only $2^{-\alpha}$. Δ must be chosen to balance statistical accuracy (smaller Δ) against compute and memory (larger Δ). The inter–sample component scales linear in T, so a similar trade–off applies.

Variance reduction by inter–sample attention. When subject-level heterogeneity is not prominent, queries/keys align across subjects and the inter-sample attention weights are nearly uniform ($\approx 1/B$), yielding an $O_{\mathbb{P}}(1/\sqrt{B})$ reduction in embedding noise (Lemma 9). This variance contraction is orthogonal to deterministic approximation biases, so increasing B stabilizes the embedding without changing those bias terms. MSE improves when the induced averaging does not introduce substantial pooling bias, namely, when the attended neighbors are genuinely similar for the target.

Masking Y-token. During training, the final token carries the response embedding to help learn the $X \to Y$ relation. By Lemma 12, masking the Y-token induces an approximation perturbation scaled by $L_Y = \|E_Y\|_{\mathrm{op}}$. To reduce train–test mismatch and improve robustness, one can randomly mask the Y-token during training: draw $d_i \sim \mathrm{Bernoulli}(q)$ and feed $\tilde{Z}_Y = d_i E_Y Y_i$ with a corresponding mask-reweighted loss. The resulting train phase embedding error remains bounded up to constants, with the fully masked case (matching test time) serves as a worst-case upper bound.

4 Experiments

We compare our method against a diverse set of 19 baselines that collectively cover statistical, functional, attention-based tabular, deep learning methods and ensemble approaches, each adapted to irregular and sparse longitudinal inputs. The statistical and/or functional baselines comprise ordinary linear regression (LR), functional linear regression (FLR) (Yao et al., 2005a; Cai & Hall, 2006), and functional principal components analysis followed by a regression neural layer (FPCA+NN) (Yao et al., 2005b; Wang et al., 2016).

For deep learning and tabular Transformer methods, we provide mean-imputed inputs on a fixed grid, where each feature corresponds to a time point. Compared methods include SAINT (Somepalli et al., 2021), FTTransformer (Gorishniy et al., 2021), TabNet (Arik & Pfister, 2021), AutoInt (Song et al., 2019), and a vanilla Transformer trained solely on covariates without a Y-token followed by a regression neural layer (VT+NN). We also include the most recent state-of-the-art tabular model, TabPFN (Hollmann et al., 2022; 2025), a generative Transformer-based foundation model pretrained on millions of synthetic datasets. To assess the value of end-to-end training relative to decoupled imputation, we also evaluate SAND (Hong et al., 2024) augmented with a prediction multilayer perceptron (SAND+NN)[1], thereby testing how well a learned imputer performs when the regression layer is trained separately. Other deep learning approaches including multilayer perceptron (MLP) and ResNet (He et al., 2016) are considered. As well as AdaFNN (Yao et al., 2021), a basis-specified neural method tailored to completely observed functional data.

Finally, we benchmark strong ensemble and tree-based systems, including AutoGluon (Erickson et al., 2020), which automatically trains, tunes, and stacks diverse models on tabular tasks. Gradient-boosted approaches, including XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018), and NODE (Popov et al., 2019) are included. We also evaluate xRFM (Beaglehole et al., 2025), a very recent method that combines random-feature (kernel-style) learning with tree-based partitioning. These 19 existing models are compared with our proposed dual-attention method (IDAT) and a variant without inter-sample attention (IDAT w/o A_I).

4.1 Simulation

Without loss of generality, we consider functions on the unit interval $\mathcal{I} = [0, 1]$ generated as follows. In all simulations, the time grid has length $T = 100$. For subject i in group $g \in \{1, \ldots, G\}$, the (noise-free) latent trajectory is

$$X_i(t) \;=\; \mu_g(t) \;+\; \sum_{k=1}^{20} \left[a_k^s \sin(2\pi k t) + a_k^c \cos(2\pi k t)\right]/k, \qquad t \in \mathcal{I},$$

where μ_g is a group-specific mean function and $\{a_k^s, a_k^c\}$ are Fourier coefficients with smoothness controlled by the decay rate of k^{-1}. To induce smooth but heavy-tailed trajectories (beyond the sub-Gaussian assumption), these coefficients are independently drawn from a zero-mean exponential distribution. The response is generated by a functional operator $Y_i = \mathcal{F}(X_i) + \varepsilon_i$ with $\varepsilon_i \overset{iid}{\sim} N(0, 1)$.

[1] SAND is an imputation method and is not applied when the covariates are 100% observed.

Case I: Functional linear regression. Set $\mu_g \equiv 0$. Let $\mathcal{F}(X) = \int_0^1 \beta_1(t)\, X(t)\, dt$ with

$$\beta_1(t) \;=\; (3-6t)\,\mathbf{1}\{t \le 0.5\} \;+\; (2t-1)\,\mathbf{1}\{t > 0.5\}.$$

Case II: Nonlinear model. Set $\mu_g \equiv 0$. Define

$$\mathcal{F}(X) = \int_0^1 \beta_2(t)\, X(t)\, dt \;+\; \Big(\int_0^1 \beta_3(t)\, X(t)\, dt\Big)^2,$$

with $\beta_2(t) = (4-16t)\,\mathbf{1}\{t \le 0.25\}$ and $\beta_3(t) = (4-16|t-0.5|)\,\mathbf{1}\{0.25 \le t \le 0.75\}$. Here, time points $t > 0.75$ are non-informative; the interval just beyond $t = 0.25$ contributes weakly.

Case III: Cluster analysis. Let $G = 2$ and with $\mathcal{F}(X) = \int_0^1 0.5\, t\, X(t)\, dt$,

$$\mu_1(t) = 1 + 4(t-0.5)\mathbf{1}\{t \ge 0.5\}, \qquad \mu_2(t) = -6(t-0.5)\mathbf{1}\{t \ge 0.5\}.$$

This scenario is explicitly designed to demonstrate the clustering ability, where group-specific mean shifts create separable subpopulations that inter-sample attention can cluster.

In addition to Cases I-III, we also add measurement error on the functional covariate[2]. We set the signal-to-noise ratio to SNR $= \Big(\int_0^1 X_i(t)^2\, dt\Big) / \mathrm{sd}(\eta_i) = 4$. Across the six simulation cases, we evaluate five observation regimes at various levels of sparsity: *ssparse* (10%), *vsparse* (20%), *sparse* (50%), *dense* (80%), and *full* (100%). The mean squared errors (MSE) for all 30 settings are reported in Tables 4–9, evaluating predictive accuracy and computational efficiency across the full spectrum of sampling scenarios. Consistently, IDAT performs best in regimes when at most 50% of time points are observed (out of 100 grid points), as summarized in Table 2. As the sparsity decreases, TabPFN becomes the primary competitor. TabPFN's pre-trained prior favors simple, additive, and near-linear relationships, allowing it to fit dominant trends in dense settings. Intuitively, inter–sample attention lowers the variance by borrowing strength from "similar" subjects; under measurement error, however, similarity can be inaccurately estimated and borrowing from mismatched neighbors raises the bias. When the increase in bias outweighs variance reduction, the time–point–only variant (no A_I) can prevail, though the dual-attention model still outperforms other baselines. Empirically, IDAT delivers the largest gains in sparse settings with clear cluster structures (Table 3). In practice, we recommend the dual–attention model by default for detecting and exploiting clustering; if the learned attention maps are diffuse or uniform across subjects, a pragmatic fallback is the A_T–only variant.

4.2 Real Data

National Child Development Study: We analyze data from the 1958 National Child Development Study (NCDS)[3]. The task is to predict BMI at age 62 from prior BMI trajectories observed at ages 11, 16, 23, 33, 42, 44, 46, 50 and 55. The cohort is relatively homogeneous, including individuals born in Great Britain during a single week in 1958, reducing potential confounding by ethnicity. All models are adjusted for baseline covariates measured at age 7: sex, baseline BMI, and an early-life social adversity index computed as the average of 13 binary indicators (e.g., housing problems, financial hardship, parental divorce, unemployment, illness, disability, or death) (Flèche et al., 2021). We fit sex-stratified models and compare the mean squared error (MSE) and mean absolute error (MAE) of two imputation pipelines for missing BMI trajectories: (i) mean imputation (a simple but commonly used approach in tabular workflows) and (ii) multiple imputation by chained equations (MICE) (van Buuren & Groothuis-Oudshoorn, 2011). The $n = 4952$ longitudinal BMI series exhibit substantial missingness (mean 25%, range 8%–96%), spanning dense to super-sparse regimes. On average, subjects have 6.2 observations (SD 0.9) across the sweeps (Figure 7). Across both imputation settings, IDAT consistently outperforms all competing models, while requiring no imputation or preprocessing and operating directly on sparse and irregular inputs.

Synthetic HIV Dataset (Health Gym Project): We use the HIV dataset from the Health Gym project (Kuo et al., 2022), a public collection of synthetic yet realistic clinical datasets, and focus exclusively on the HIV cohort. Since the measurements are monthly and equally spaced, alongside

[2]Cases with measurement errors are denoted by Case I*, II* and III*

[3]University College London, UCL Social Research Institute, Centre for Longitudinal Studies (2024)

mean imputation and MICE, we also evaluate the last-observation-carried-forward (LOCF) method (Lachin, 2016; Woolley et al., 2009) and report accuracy and F1 score[4] in Table 10. The binary response label indicates whether a patient achieves viral suppression (VL$<$ 200 copies/mL) at any time during the prediction window (months 20–30). For each patient, we use VL measurements from the first 20 months as longitudinal covariates and include sex as a baseline covariate. The $n = 8683$ VL series in the feature window is very sparse (mean missingness 65%, range 47–88%). Subjects receive a mean of 5.8 observations (SD 1.8) during the 20-month covariate interval (Figure 9).

Table 1: Performance on real-world tasks. For NCDS BMI (regression), we report MSE/MAE under mean and MICE imputation. For Synthetic HIV (classification), we report F1 score under mean, MICE, and LOCF imputation. The best method is in **bold** and the top three are in *italics*.

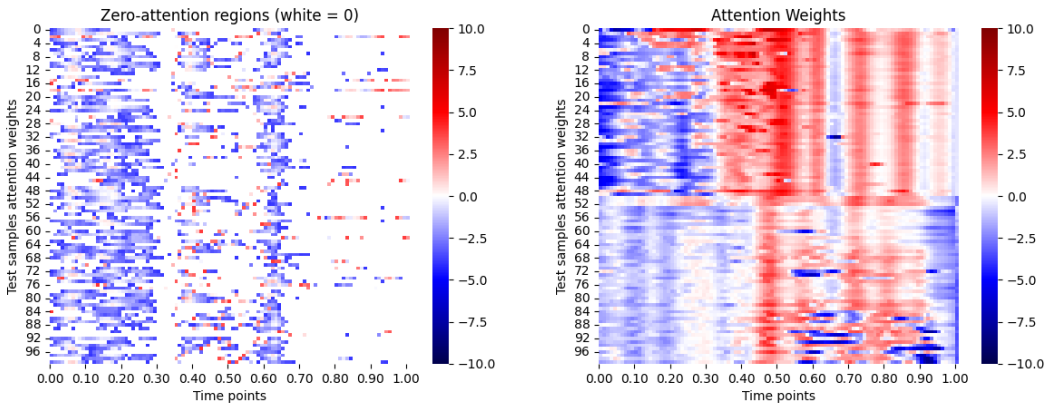
	NCDS BMI				Synthetic HIV		
	mean imputed		MICE		mean	MICE	LOCF
Method	MSE	MAE	MSE	MAE	F1	F1	F1
LM/GLM	21.2318	3.5864	9.1630	2.2935	*0.9746*	*0.9751*	0.9716
FLR/FGLM	11.1655	2.5682	11.1655	2.5682	*0.9736*	0.9736	0.9736
FPCA+NN	10.2562	2.3870	10.2562	2.3870	0.9698	0.9698	0.9698
TabNet	9.6078	2.3381	*8.1461*	2.1843	0.9734	0.9740	0.9752
SAINT	9.7213	2.3524	9.4844	2.3231	**0.9752**	*0.9745*	0.9752
FTTransformer	9.8154	2.3606	9.2480	2.3276	0.9721	0.9739	0.9757
AutoInt	8.6933	2.2375	8.7121	2.2607	0.9703	0.9740	*0.9758*
TabPFN	*8.5794*	**2.1621**	8.5164	**2.1503**	0.9722	0.9734	*0.9758*
VT+NN	19.1824	3.3388	19.1824	3.3388	0.9611	0.9611	0.9611
SAND+NN	19.1647	3.3413	19.1647	3.3413	0.9647	0.9647	0.9647
MLP	9.4439	2.3142	9.2378	2.3191	**0.9752**	0.9727	*0.9764*
ResNet	9.3321	2.3526	8.9923	2.2780	0.9715	0.9751	0.9752
AdaFNN	9.2237	2.2716	8.6085	2.2147	0.9649	0.9649	0.9649
NODE	9.6511	2.2955	8.6911	2.2528	0.9727	*0.9751*	**0.9769**
CatBoost	9.5655	2.3125	9.1693	2.2969	0.9679	0.9715	0.9751
XGBoost	10.9604	2.4617	10.3945	2.4074	0.9652	0.9727	0.9745
LightGBM	9.5160	2.3066	9.9092	2.3795	0.9678	0.9715	0.9733
AutoGluon	9.0408	2.2623	8.9595	2.2543	*0.9746*	*0.9745*	0.9751
xRFM	11.5200	2.5628	20.5166	2.7625	0.9698	0.9698	0.9698
IDAT	**8.0061**	*2.1728*	**8.0061**	*2.1728*	**0.9752**	**0.9752**	0.9752
IDAT w/o A_I	*8.1177*	*2.1966*	*8.1177*	*2.1966*	**0.9752**	**0.9752**	0.9752

5 Conclusion

Across datasets and simulations with diverse missingness, our end-to-end dual-attention Transformer IDAT, without external imputation and with supervision via the Y-token, consistently performs the best when less or equal to 50% of time points are observed, a regime typical of many longitudinal applications, and remains competitive as sparsity decreases. In practice, sparsity varies widely across cohorts, time windows, and variables; thus, robustness across sampling densities is essential. By adaptively leveraging inter-sample attention to borrow strength when data are scarce and emphasizing time-point structure as coverage improves, IDAT offers a unified solution across the full sparsity spectrum, while yielding interpretable dual-attention patterns that clarify when and where each mechanism contributes.

Domain detection with time–point attention. Time-point attention learns a data-driven weighting scheme over time, highlighting the segments of a trajectory that are most predictive. As shown in Figures 2a and 8b, when portions of the time axis are not informative for the response, the learned

[4]The F1 score is a classification evaluation metric that represents the harmonic mean of precision and recall.



(a) Domain detection in case II.

(b) Embedding clustering in case III.

Figure 2: Interpretable dual-attention. All the attention weights are are scaled by a factor of 1000.

weights shrink toward (near) zero in those intervals, effectively performing time domain selection. This behavior mirrors the known informative window in the simulation and is corroborated by external domain knowledge in the real cohort as shown in Figure 8.

Clustering with inter–sample attention. Inter–sample attention acts as a learned, end-to-end nearest-neighbor mechanism: it computes attention across data points (rows) within a batch using a learned similarity and aggregates information from the most relevant samples. This cross-sample sharing of information is particularly helpful for missing or noisy features. In clustered data, it produces cluster-specific attention profiles that better align with the regression signal. Figure 2b illustrates a two-group setting with different group mean functions: the clusters display clearly distinct attention patterns (e.g., Group 1 assigns negative weights in the first third of the trajectory and near-zero weights in the last third, whereas Group 2 shows near-zero weights early and coherent within-cluster structure thereafter). This demonstrates the clustering capability of the dual-attention mechanism. In contrast, removing inter-sample attention (Figure 6) makes profiles of the two clusters much more similar, differing only slightly in the early trajectory. This highlights that time-point and inter-sample attention provide complementary and additive gains in predictive performance.

Extension of the dual-attention method. On real data (Table 1), we replace the regression head with a classification head while keeping the dual-attention encoder unchanged, demonstrating that IDAT adapts seamlessly to classification. However, in imbalanced settings, inter-sample attention can amplify majority signals and attenuate minority patterns. Time-independent covariates can be included by simple concatenation (without positional encodings), so the model jointly learns their relationships with the longitudinal covariates and the outcome. For multi-dimensional functional input, an intra-functional attention layer could be added to capture cross-channel relations. Tabular data can be cast as functional via "stringing" (Chen et al., 2011): order features by similarity and treat each reordered row as samples from a smooth curve, carrying induced positions and masks. Sinusoidal encodings do not extrapolate beyond the largest trained horizon or unseen time grids; relative or seasonal encoders may better capture temporal structure (Zhou et al., 2021; Woo et al., 2022). Although dual-attention can increase computational cost, Table 2 shows that IDAT is relatively fast compared to existing Transformer-based methods. To balance variance reduction and pooling bias under measurement noise, we further introduce a learnable, data-adaptive gate $\lambda \in [0, 1]$, trained end-to-end:

$$\mathcal{T}_B^{(\lambda)} \;=\; \big[\lambda\, A_T + (1-\lambda)\, A_I\big] \circ \mathrm{FF}.$$

REFERENCES

Maryamossadat Aghili, Vittorio Murino, and Diego Sona. Predictive modeling of longitudinal data for alzheimer's disease diagnosis using rnns. In *International Workshop on PRedictive Intelligence In MEdicine*, pp. 13–21. Springer, 2018.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6679–6687, 2021.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Daniel Beaglehole, David Holzmüller, Adityanarayanan Radhakrishnan, and Mikhail Belkin. xrfm: Accurate, scalable, and interpretable feature learning models for tabular data, 2025. URL https://arxiv.org/abs/2508.10053.

T Tony Cai and Peter Hall. Prediction in functional linear regression. 2006.

Lucia A Carrasco-Ribelles, Fernando Martín-Sánchez, Niels Peek, and Carlos Sáez. Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review. *Journal of the American Medical Informatics Association*, 30 (12):2072–2082, 2023.

Anna Cascarano, Ilaria D'Aloisio, Giusy Guastamacchia, and Francesca Ieva. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artificial Intelligence Review*, 56(Suppl 2):1711–1771, 2023.

Centers for Disease Control and Prevention. Hiv diagnoses, deaths, and prevalence, April 2025. URL https://www.cdc.gov/hiv-data/nhss/hiv-diagnoses-deaths-prevalence.html. Accessed: 2025-09-12.

Kun Chen, Kehui Chen, Hans-Georg Müller, and Jane-Ling Wang. Stringing high-dimensional data for functional analysis. *Journal of the American Statistical Association*, 106(493):275–284, 2011.

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.

Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5896–5905, 2023.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019.

Ruoxuan Cui, Manhua Liu, and Alzheimer's Disease Neuroimaging Initiative. Rnn-based longitudinal analysis for diagnosis of alzheimer's disease. *Computerized Medical Imaging and Graphics*, 73:1–10, 2019.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Anna K Dahl, Chandra A Reynolds, Tove Fall, Patrik KE Magnusson, and Nancy L Pedersen. Multifactorial analysis of changes in body mass index across the adult life course: a study with 65 years of follow-up. *International journal of obesity*, 38(8):1133–1141, 2014.

Peter J. Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L. Zeger. *Analysis of Longitudinal Data*. Oxford Statistical Science Series. Oxford University Press, Oxford, 2 edition, 2002.

Robert W Eisinger, Carl W Dieffenbach, and Anthony S Fauci. Hiv viral load and transmissibility of hiv infection: undetectable equals untransmittable. *Jama*, 321(5):451–452, 2019.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs (eds.). *Longitudinal Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2009.

Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2004.

Sarah Flèche, Warn N Lekfuangfu, and Andrew E Clark. The long-lasting effects of family and childhood on adult wellbeing: Evidence from british cohort data. *Journal of Economic Behavior & Organization*, 181:290–311, 2021.

Steven Garasky, Susan D Stewart, Craig Gundersen, Brenda J Lohman, and Joey C Eisenmann. Family stressors and child obesity. *Social science research*, 38(4):755–766, 2009.

Matthew W Gillman, Sheryl Rifas-Shiman, Catherine S Berkey, Alison E Field, and Graham A Colditz. Maternal gestational diabetes, birth weight, and adolescent obesity. *Pediatrics*, 111(3): e221–e226, 2003.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Jennifer A Halliday, Cassandra L Palma, David Mellor, Julie Green, and AMN Renzaho. The relationship between family functioning and child and adolescent overweight and obesity: a systematic review. *International journal of obesity*, 38(4):480–493, 2014.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Ju-Sheng Hong, Junwen Yao, Jonas Mueller, and Jane-Ling Wang. SAND: Smooth imputation of sparse and noisy functional data with transformer networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MXRO5kukST.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4 (2):251–257, 1991.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Lukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, and Jonni Kanerva. Sparse is enough in scaling transformers. *Advances in Neural Information Processing Systems*, 34:9895–9907, 2021.

BARTLETT JG. Panel on antiretroviral guidelines for adults and adolescents. guidelines for the use of antiretroviral agents in hiv-1-infected adults and adolescents. *The Department of Health and Human Services Panel on Antiretroviral Guidelines for Adult and Adolescents*, pp. 42–43, 2008.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? *arXiv preprint arXiv:2307.14023*, 2023.

Michail Katsoulis, Martina Narayanan, Brian Dodgeon, George Ploubidis, and Richard Silverwood. A data driven approach to address missing data in the 1970 british birth cohort. *medRxiv*, pp. 2024–02, 2024.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Nicholas I-Hsien Kuo, Mark N Polizzotto, Simon Finfer, Federico Garcia, Anders Sönnerborg, Maurizio Zazzi, Michael Böhm, Rolf Kaiser, Louisa Jorm, and Sebastiano Barbieri. The health gym: synthetic health-related datasets for the development of reinforcement learning algorithms. *Scientific data*, 9(1):693, 2022.

John M Lachin. Fallacies of last observation carried forward analyses. *Clinical trials*, 13(2):161–168, 2016.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974, 1982a. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529876.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pp. 963–974, 1982b.

Michel Ledoux and Michel Talagrand. Probability in banach spaces. classics in mathematics, 2011.

Yikuan Li, Shuvro Rao, José Roberto Ayala Solares, Ahmed Hassaine, Peter Ramadge, Hui Li, and Jim Sun. Behrt: Transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Zachary C Lipton, David C Kale, and Randall Wetzel. Modeling missing data in clinical time series with rnns. In *Machine Learning for Healthcare Conference*, pp. 231–245. PMLR, 2016.

Chao Lou, Zixia Jia, Zilong Zheng, and Kewei Tu. Sparser is faster and less is more: Efficient sparse attention for long-range transformers. *arXiv preprint arXiv:2406.16747*, 2024.

Tarek Mostafa, Martina Narayanan, Benedetta Pongiglione, Brian Dodgeon, Alissa Goodman, Richard J Silverwood, and George B Ploubidis. Missing at random assumption made more plausible: evidence from the 1958 british birth cohort. *Journal of Clinical Epidemiology*, 136:44–54, 2021.

Hans-Georg Mu et al. Functional modeling of longitudinal data. In *Longitudinal data analysis*, pp. 237–266. Chapman and Hall/CRC, 2008.

Mine Öğretir, Miika Koskinen, Juha Sinisalo, Risto Renkonen, and Harri Lähdesmäki. Seqrisk: Transformer-augmented latent variable model for improved survival prediction with longitudinal data. *arXiv preprint arXiv:2409.12709*, 2024.

World Health Organization et al. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing hiv infection: recommendations for a public health approach june 2013. In *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach June 2013*, pp. 272–272. 2013.

Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

Tessa J Parsons, Chris Power, Stuart Logan, and CD Summerbelt. Childhood predictors of adult obesity: a systematic review. *International journal of obesity*, 23, 1999.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312*, 2019.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

James O Ramsay and Bernard W Silverman. *Functional data analysis*. Springer, 2005.

Imogen Rogers. The influence of birthweight and intrauterine environment on adiposity and fat distribution in later life. *International journal of obesity*, 27(7):755–777, 2003.

Caroline A Sabin, Helen Devereux, Andrew N Phillips, Andrew Hill, George Janossy, Christine A Lee, and Clive Loveday. Course of viral load throughout hiv-1 infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 23(2):172–177, 2000.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Clauirton A Siebra, Mascha Kurpicz-Briki, and Katarzyna Wac. Transformers in health: a systematic review on architectures for longitudinal data analysis. *Artificial Intelligence Review*, 57(2): 32, 2024.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1161–1170, 2019.

Christina Stenhammar, Gunilla Maria Olsson, S Bahmanyar, A-L Hulting, Björn Wettergren, Birgitta Edlund, and Scott M Montgomery. Family stress and bmi in young children. *Acta paediatrica*, 99(8):1205–1212, 2010.

Maxwell B Stinchcombe. Neural network approximation of continuous functionals and continuous functions on compactifications. *Neural Networks*, 12(3):467–477, 1999.

Naoki Takeshita and Masaaki Imaizumi. Approximation of permutation invariant polynomials by transformers: Efficient construction in column-size. *arXiv preprint arXiv:2502.11467*, 2025.

Anton Frederik Thielmann, Manish Kumar, Christoph Weisser, Arik Reuter, Benjamin Säfken, and Soheila Samiee. Mambular: A sequential model for tabular deep learning. *arXiv preprint arXiv:2408.06291*, 2024.

University College London, UCL Social Research Institute, Centre for Longitudinal Studies. National child development study. UK Data Service, 2024. URL `https://doi.org/10.5255/UKDA-Series-2000032`. Data series, 14th Release. SN: 2000032.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and its application*, 3(1):257–295, 2016.

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pp. 1578–1585. IEEE, 2017.

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.

Stephen B Woolley, Alex A Cardoni, and John W Goethe. Last-observation-carried-forward imputation method in clinical efficacy trials: review of 352 antidepressant studies. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 29(12):1408–1416, 2009.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. 2005a.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590, 2005b.

Junwen Yao, Jonas Mueller, and Jane-Ling Wang. Deep learning for functional data analysis with adaptive basis layers. In *International conference on machine learning*, pp. 11898–11908. PMLR, 2021.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94: 103–114, 2017.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Maurizio Zazzi, Francesca Incardona, Michal Rosen-Zvi, Mattia Prosperi, Thomas Lengauer, Andre Altmann, Anders Sonnerborg, Tamar Lavee, Eugen Schülter, and Rolf Kaiser. Predicting response to antiretroviral treatment by machine learning: the euresist project. *Intervirology*, 55(2):123–127, 2012.

George Zerveas, Srideep Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eforn. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124, 2021.

Zhiyue Zhang, Yao Zhao, and Yanxun Xu. Transformerlsr: Attentive joint model of longitudinal data, survival, and recurrent events with concurrent latent structure. *Artificial intelligence in medicine*, 160:103056, 2025.

Juan Zhao, QiPing Feng, Patrick Wu, Roxana A Lupu, Russell A Wilke, Quinn S Wells, Joshua C Denny, and Wei-Qi Wei. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Scientific reports*, 9(1):717, 2019.

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.