TOTEM: TOKENIZED TIME SERIES EMBEDDINGS FOR GENERAL TIME SERIES ANALYSIS

Sabera Talukder, Yisong Yue & Georgia Gkioxari
California Institute of Technology
{sabera,yyue,georgia}@caltech.edu

ABSTRACT

Learning with time series health data poses many challenges such as variability in sensor semantics (e.g. neural voltage recordings vs US birth rate), difficulty in accessing data, and the relatively smaller data volume compared to other time series domains. Given these limitations, and the fact that the field of general time series analysis has recently begun to explore unified modeling, we approach unification from a complementary vantage point to ultimately benefit zero-shot performance to health time series. Historically general time series analysis unification entails when a common architectural backbone is retrained on a specific task for a specific dataset; we study the unification of time series data representations across domains in many tasks. To this end, we explore the impact of discrete, learnt, time series data representations that enable generalist, cross-domain training. Our method, TOTEM, or Tokenized Time Series Embeddings, proposes a simple tokenizer architecture that embeds time series data from varying domains using a discrete vectorized representation learned in a self-supervised manner. TOTEM works across multiple tasks and domains with minimal to no tuning. We study TOTEM's efficacy with an extensive evaluation on 17 real world time series datasets across 3 tasks. Notably, the majority of our zero-shot datasets are time series health datasets from the neuroscience and birth domains. We evaluate both the specialist (i.e., train a model on each domain) and generalist (i.e., train a single model on many domains), and show that TOTEM matches or outperforms previous best methods on several popular benchmarks. Please find the full paper here: https://arxiv.org/pdf/2402.16412.pdf, and the code here: https://github.com/SaberaTalukder/TOTEM.

1 INTRODUCTION

Time series analysis, both for health and more generally, encompasses a wide range of datasets, tasks, and applications in the real world. When considering training paradigms, time series analysis has historically been conducted via *specialist-training*, meaning that models are trained on a single time series domain (Zhou et al., 2023; Wu et al., 2022a; Nie et al., 2022; Zhang & Yan, 2022). *Generalist-training*, where models are simultaneously trained on multiple time series domains, contrasts the specialist paradigm. Both specialist and generalist models can be tested under various regimes. Within *in-domain-testing*, a model is tested on the same domain(s) it was trained on. In *zero-shot-testing*, a model is tested on different domains(s) than it was trained on. Some methods have begun to explore the idea of zero-shot forecasting where (1) a forecaster trains on one dataset then predicts on a separate dataset (Zhou et al., 2023), or (2) a forecaster trains on a subset of channels (which we call *sensors*) from one dataset then zero-shot forecasts on the remaining sensors in the same dataset (Liu et al., 2023). Both of these models would be considered specialists, as they were trained on only one (or a subset of one) dataset. In order to fully enable generalist training and zero shot testing we explore the value of unified time series data representations.

Further, time series analysis has typically been restricted by task, where methods study only *forecasting* (Wu et al., 2021; Woo et al., 2022), *anomaly detection* (Xu et al., 2021; He & Zhao, 2019), or *imputation* (Luo et al., 2018; 2019), among others. Recently, the field has become increasingly unified with respect to model architecture, with methods (Zhou et al., 2023; Wu et al., 2022a) ex-

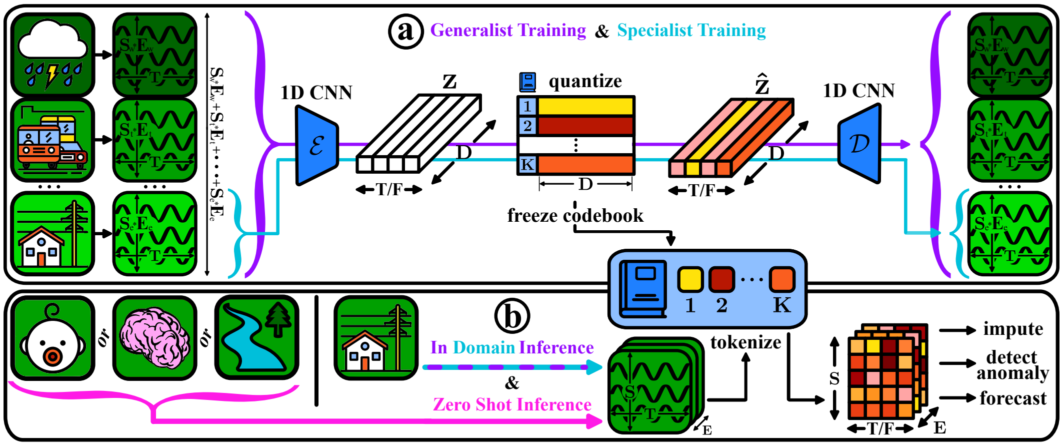

Figure 1: **TOTEM & Evaluation Regimes.** **(a)** The TOTEM VQVAE architecture consists of an 1D strided CNN encoder $\mathcal{E}$, quantizer, latent codebook, and 1D strided transpose CNN decoder $\mathcal{D}$. TOTEM's VQVAE enables generalist training, i.e. on all datasets jointly, and specialist training, i.e. on one dataset at a time. **(b)** TOTEM's discrete, self-supervised codebook can be leveraged for both in domain and zero shot testing. We utilize US birth and neuroscience domains for zero-shot testing.

ploring language and vision backbones on various time series tasks. These backbones, like previous methods, utilize specialist training (e.g., training separate anomaly detectors on each dataset).

The field has also become increasingly unified with respect to data representation, with growing emphasis on learning performant data representations. For instance, Franceschi et al. (2019) utilize an exponentially dilated causal convolutional encoder to discover in-domain embeddings, Tonekaboni et al. (2021) leverage temporal neighborhood coding, Yang & Hong (2022) utilize temporal-spectral fusion, and Yue et al. (2022) employs hierarchical contrasting across time and batch dimensions.

At a technical level, our approach bears closest affinity to methods that use vector quantized variational autoencoders (VQVAEs) (Van Den Oord et al., 2017; Duan et al., 2023; Rasul et al., 2022b;a). As we discuss further in Section 2, Our goal is to develop a streamlined framework for learning a tokenized data representation (using VQVAEs) in a way that permits easy applicability and holistic empirical evaluation on a broad range of time series modeling tasks and data domains (including zero-shot generalization to new test domains) with minimal to no tuning.[1]

Motivated by the difficulty of training on health time series and the trend of time series analysis unification, we explore the value of a VQVAE-based tokenizer for time series imputation, anomaly detection (Appendix E), and forecasting (Appendix F). Unlike previous methods, we utilize self-supervised, discrete tokens, and extensively explore their utility in varied training and testing regimes. Neuro2 [N2], Neuro5 [N5], and US Births [B] are health datasets we utilize to test zero-shot performance, see Appendix D for more discussion. Our contributions are as follows:

1. We present TOTEM, a simple tokenizer architecture for time series analysis that works across domains and tasks with minimal to no tuning.
2. Despite its simplicity, TOTEM matches or outperforms the state-of-the-art on several popular benchmark datasets and tasks.
3. With an extensive evaluation in the generalist setting (training a single model on multiple domains), we show that TOTEM outperforms the leading state-of-the-art model in both in-domain and zero-shot testing regimes.

2 Method

Our proposed discrete time series tokenization enables the design of general models across a variety of time series domains, tasks, and evaluation schemas, Figure 1. We design a single tokenizer architecture that is generally applicable without extensive data engineering while being suitable for varying data dimensionalities across different tasks. There are many possibilities for how to introduce a discrete time series tokenizer, we extensively study one such methodology that satisfies the aforementioned design criteria.

[1]As an aside, our approach to studying what is a performant general time series data representation shares a philosophical alignment with the development of large generalist models in natural language processing, which are also based on having a common tokenized representation (Gage, 1994; Radford et al., 2018).

Data Engineering. Prior work leverages data engineering such as the use of auxiliary features (e.g. day of the month, or minute in the hour, etc.) (Chen et al., 2023; Salinas et al., 2020), or frequency transformations (Wu et al., 2022a; Zhou et al., 2022). We forego any data engineering and operate directly on time steps. This enables generalist-training as differing data domains have widely varying sampling rates leading to distinct auxiliary features and frequency profiles.

Varying Dimensionality. A time series dataset consists of E examples (i.e. number of distinct recordings), S sensor channels, and T time steps, and can be formally expressed as $\{\mathbf{x}_j\}_{j=1}^{E} \subset \mathbb{R}^{S\times T}$. Even within a single task and single data domain where S does not change, E and T take on a wide range of values. As an example, canonical forecasting predictions lengths range from 96 to 720 time steps. When moving to generalist-training, datasets additionally have wide ranging sensor dimensionalities S. Our tokenizer handles varying dimensionality across E, S, and T by creating non-overlapping tokens along the time-dimension that are smaller than the dimension T.

Differing Tasks. There are numerous tasks to tackle in health time series analysis. Three significant ones are imputation, anomaly detection (Appendix E), and forecasting (Appendix F). In *imputation*, models intake a masked time series $\mathbf{x_m} \in \mathbb{R}^{S\times T_{\text{in}}}$, and then reconstruct and impute $\mathbf{x} \in \mathbb{R}^{S\times T_{\text{in}}}$. In *anomaly detection*, models intake a corrupted time series $\mathbf{x_{corr}} \in \mathbb{R}^{S\times T_{\text{in}}}$ and reconstruct the data $\mathbf{x} \in \mathbb{R}^{S\times T_{\text{in}}}$. The amount of corruption is considered known, at A%. In *forecasting*, models intake a time series $\mathbf{x} \in \mathbb{R}^{S\times T_{\text{in}}}$ and predict future readings $\mathbf{y} \in \mathbb{R}^{S\times T_{\text{out}}}$, where S is the number of sensors and $T_{\text{in}}, T_{\text{out}}$ signify the durations of the preceding and succeeding time series, respectively. Our tokenizer is performant across all tasks despite their distinct representational requirements.

TOTEM Implementation. To realize a single tokenizer architecture that enables generalist modeling across differing domains and tasks we take inspiration from the VQVAE (Van Den Oord et al., 2017). The original VQVAE leverages a dilated convolutional architecture with a stride of 2 and window-size of 4, similar to the WaveNet (Oord et al., 2016) dilated, causal, convolutional decoder. A dilated convolution skips inputs allowing a filter to operate on a larger input area / coarser scale. Utilizing dilated convolutions is an architectural decision rooted in the high sampling rates of raw audio waveforms (Oord et al., 2016; Van Den Oord et al., 2017). High sampling rates are not a trait shared by many time series domains.

When adapting the VQVAE for general time series analysis, the TOTEM VQVAE:

1. Operates directly on time steps; no data engineering.
2. Creates discrete, non-overlapping tokens along the time dimension of length F, where $F < T$, thereby promoting training and testing on variable length examples, E, sensors, S, and time steps T.
3. Maintains the same architecture and objective regardless of the downstream task.
4. Aims to capture maximal information within a large receptive field by: (1) using a strided non-causal convolutional architecture with no dilation, (2) training on long time series inputs, (3) pre-striding the data by a stride of 1 so the tokenizer learns from maximal inputs.

The TOTEM VQVAE consists of an encoder, quantizer, latent codebook, and decoder. It takes in a univariate time series $\{\mathbf{x}_i \in \mathbb{R}^T\}_{i=1}^{E\cdot S}$ obtained by flattening the multivariate sensor channel. This makes TOTEM's VQVAE sensor-agnostic, enabling TOTEM's generalist-training and zero-shot-testing. The encoder $\mathcal{E}$ consists of strided 1D convolutions compressing the time series by a cumulative stride of F. $\mathcal{E}$ maps a univariate time series $\mathbf{x} \in \mathbb{R}^T$ to a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{T/F \times D}$, where D is the the hidden dimension. The latent codebook $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^{K}$ consists of K D-dim codewords $\mathbf{c}_i \in \mathbb{R}^D$. During quantization, the codebook is used to replace $\mathbf{z}$ with $\hat{\mathbf{z}} \in \mathbb{R}^{T/F \times D}$ such that $\hat{\mathbf{z}}_j = \mathbf{c}_k$, where $k = \arg\min_i ||\mathbf{z}_j - c_i||_2$. The decoder $\mathcal{D}$ follows the reverse architecture of the encoder $\mathcal{E}$, consisting of 1D transpose convolutions with a cumulative stride of $1/F$ mapping the quantized $\hat{\mathbf{z}}$ to a reconstructed time series $\hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}}) \in \mathbb{R}^T$. We learn $\mathcal{E}$, $\mathcal{D}$, and $\mathcal{C}$ by optimizing the objective $\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cmt}}$ consisting of a reconstruction loss $\mathcal{L}_{\text{rec}} = \frac{1}{E\cdot S}\sum_i ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2$ and a commitment loss $\mathcal{L}_{\text{cmt}}$, which allows the codebook to update despite the the non-differentiable $\arg\min$ operation during quantization. The final objective is $\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha \cdot \mathcal{L}_{\text{cmt}}$, where α is a scalar that weights the two losses. This objective does not change even when the underlying task, time series length, data masking, normalization schema, or data domain changes.

Table 1: **Specialist Imputation** (↓). Across all datasets, metrics, and masking percentages, TOTEM has the highest AvgWins (**52.1%**), followed by GPT2 (**35.4%**). TOTEM values are means from 3 seeds; baseline values are from Zhou et al. (2023); Wu et al. (2022a).

Model		TOTEM		GPT2		TiNet		Patch		ETS		FED		Stat		Auto		Inf		Re		LiTS		Dlin	
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
W	12.5%	0.028	0.046	0.026	0.049	0.025	0.045	0.029	0.049	0.057	0.141	0.041	0.107	0.027	0.051	0.026	0.047	0.037	0.093	0.031	0.076	0.047	0.101	0.039	0.084
	25%	0.029	0.047	0.028	0.052	0.029	0.052	0.031	0.053	0.065	0.155	0.064	0.163	0.029	0.056	0.030	0.054	0.042	0.100	0.035	0.082	0.052	0.111	0.048	0.103
	37.5%	0.031	0.048	0.033	0.060	0.031	0.057	0.035	0.058	0.081	0.180	0.107	0.229	0.033	0.062	0.032	0.060	0.049	0.111	0.040	0.091	0.058	0.121	0.057	0.117
	50%	0.033	0.052	0.037	0.065	0.034	0.062	0.038	0.063	0.102	0.207	0.183	0.312	0.037	0.068	0.037	0.067	0.053	0.114	0.046	0.099	0.065	0.133	0.066	0.134
E	12.5%	0.054	0.154	0.080	0.194	0.085	0.202	0.055	0.160	0.196	0.321	0.107	0.237	0.093	0.210	0.089	0.210	0.218	0.326	0.190	0.308	0.102	0.229	0.092	0.214
	25%	0.059	0.160	0.087	0.203	0.089	0.206	0.065	0.175	0.207	0.332	0.120	0.251	0.097	0.214	0.096	0.220	0.219	0.326	0.197	0.312	0.121	0.252	0.118	0.247
	37.5%	0.067	0.169	0.094	0.211	0.094	0.213	0.076	0.189	0.219	0.344	0.136	0.266	0.102	0.220	0.104	0.229	0.222	0.328	0.203	0.315	0.141	0.273	0.144	0.276
	50%	0.079	0.183	0.101	0.220	0.100	0.221	0.091	0.208	0.235	0.357	0.158	0.284	0.108	0.228	0.113	0.239	0.228	0.331	0.210	0.319	0.160	0.293	0.175	0.305
m1	12.5%	0.049	0.125	0.017	0.085	0.019	0.092	0.041	0.130	0.067	0.188	0.035	0.135	0.026	0.107	0.034	0.124	0.047	0.155	0.032	0.126	0.075	0.180	0.058	0.162
	25%	0.052	0.128	0.022	0.096	0.023	0.101	0.044	0.135	0.096	0.229	0.052	0.166	0.032	0.119	0.046	0.144	0.063	0.180	0.042	0.146	0.093	0.206	0.080	0.193
	37.5%	0.055	0.132	0.029	0.111	0.029	0.111	0.049	0.143	0.133	0.271	0.069	0.191	0.039	0.131	0.057	0.161	0.079	0.200	0.063	0.182	0.113	0.231	0.103	0.219
	50%	0.061	0.139	0.040	0.128	0.036	0.124	0.055	0.151	0.186	0.323	0.089	0.218	0.047	0.145	0.067	0.174	0.093	0.218	0.082	0.208	0.134	0.255	0.132	0.248
m2	12.5%	0.016	0.078	0.017	0.076	0.018	0.080	0.026	0.094	0.108	0.239	0.056	0.159	0.021	0.088	0.023	0.092	0.133	0.270	0.108	0.228	0.034	0.127	0.062	0.166
	25%	0.017	0.081	0.020	0.080	0.020	0.085	0.028	0.099	0.164	0.294	0.080	0.195	0.024	0.096	0.026	0.101	0.135	0.272	0.136	0.262	0.042	0.143	0.085	0.196
	37.5%	0.018	0.084	0.022	0.087	0.023	0.091	0.030	0.104	0.237	0.356	0.110	0.231	0.027	0.103	0.030	0.108	0.155	0.293	0.175	0.300	0.051	0.159	0.106	0.222
	50%	0.020	0.088	0.025	0.095	0.026	0.098	0.034	0.110	0.323	0.421	0.156	0.276	0.030	0.108	0.035	0.119	0.200	0.333	0.211	0.329	0.059	0.174	0.131	0.247
h1	12.5%	0.119	0.212	0.043	0.140	0.057	0.159	0.093	0.201	0.126	0.263	0.070	0.190	0.060	0.165	0.074	0.182	0.114	0.234	0.074	0.194	0.240	0.345	0.151	0.267
	25%	0.127	0.220	0.054	0.156	0.069	0.178	0.107	0.217	0.169	0.304	0.106	0.236	0.080	0.189	0.090	0.203	0.140	0.262	0.102	0.227	0.265	0.364	0.180	0.292
	37.5%	0.138	0.230	0.072	0.180	0.084	0.196	0.120	0.230	0.220	0.347	0.124	0.258	0.102	0.212	0.109	0.222	0.174	0.293	0.135	0.261	0.296	0.382	0.215	0.318
	50%	0.157	0.247	0.107	0.216	0.102	0.215	0.141	0.248	0.293	0.402	0.165	0.299	0.133	0.240	0.137	0.248	0.215	0.325	0.179	0.298	0.334	0.404	0.257	0.347
h2	12.5%	0.040	0.129	0.039	0.125	0.040	0.130	0.057	0.152	0.187	0.319	0.095	0.212	0.042	0.133	0.044	0.138	0.305	0.431	0.163	0.289	0.101	0.231	0.100	0.216
	25%	0.041	0.131	0.044	0.135	0.046	0.141	0.061	0.158	0.279	0.390	0.137	0.258	0.049	0.147	0.050	0.149	0.322	0.444	0.206	0.331	0.115	0.246	0.127	0.247
	37.5%	0.043	0.136	0.051	0.147	0.052	0.151	0.067	0.166	0.400	0.465	0.187	0.304	0.056	0.158	0.060	0.163	0.353	0.462	0.252	0.370	0.126	0.257	0.158	0.276
	50%	0.047	0.142	0.059	0.158	0.060	0.162	0.073	0.174	0.602	0.572	0.232	0.341	0.065	0.170	0.068	0.173	0.369	0.472	0.316	0.419	0.136	0.268	0.183	0.299
AvgWins		52.1%		35.4%		18.8%		0%		0%		0%		0%		0%		0%		0%		0%		0%	

For further discussion see: reproducibility (A), ethical considerations (B), related work (C), experimental setup (D), anomaly detection (E), forecasting (F), ablations (G), exploratory studies in generalist modeling (H), and std. devs. (I). Following the field standard, we bold the **best**, **second** best, and **third** best and calculate the average number of best results, or AvgWins , for each method. We compare to two approach families: methods designed for multiple tasks (**multitask**) – TOTEM's category – and methods designed for a specific task (**singletask**), and are adapted to other tasks.

3 Imputation

In imputation, models intake a masked time series $\mathbf{x_m} \in \mathbb{R}^{S \times T_{\text{in}}}$, and then reconstruct and impute $\mathbf{x} \in \mathbb{R}^{S \times T_{\text{in}}}$ We experiment with four canonical masking percentages at 12.5%, 25%, 37.5%, 50%, and report MSE and MAE ; lower is better (↓). **Specialist.** In Table 1 we compare TOTEM to baselines. All models are trained and evaluated on the same dataset (in-domain). TOTEM has the highest AvgWins with 52.1%, followed by GPT2 at 35.4%, and TiNet at 18.8%. TOTEM performance for m1 and h1 is lower; notably these datasets are the minute and hour resampling of the same raw data respectively. We investigate and discuss TOTEM's success across different domains in Table 9. **Generalist.** In Table 2 we compare TOTEM to GPT2 (best performing models above), when both models are trained on the aggregate of W, E, m1, m2, h1, h2. We test them on the in-domain and zero-shot test sets. TOTEM outperforms GPT2 in-domain, 58.3% vs. 43.8% , and by a much larger margin in zero-shot, 80% vs. 20%. TOTEM's performance across all experiments demonstrate that tokens are a performant representation for imputation.

Table 2: **Generalist Imputation** (↓). TOTEM & GPT2 simultaneously train on all in domain datasets, 3 seeds each. **A. In-Domain Performance.** TOTEM has the highest AvgWins at **58.3%**. **B. Zero-Shot Performance.** We test on unseen datasets zero-shot. TOTEM again has the highest AvgWins at **80.0%**.

A. In-Domain Performance

Model		TOTEM		GPT2	
Metric		MSE	MAE	MSE	MAE
W	12.5%	0.029	0.060	0.029	0.045
	25%	0.030	0.060	0.033	0.048
	37.5%	0.032	0.062	0.037	0.054
	50%	0.036	0.067	0.043	0.061
E	12.5%	0.065	0.171	0.080	0.186
	25%	0.071	0.179	0.091	0.197
	37.5%	0.080	0.189	0.108	0.213
	50%	0.095	0.205	0.132	0.236
m1	12.5%	0.041	0.132	0.052	0.141
	25%	0.044	0.135	0.065	0.154
	37.5%	0.048	0.139	0.085	0.171
	50%	0.058	0.152	0.117	0.196
m2	12.5%	0.040	0.125	0.029	0.095
	25%	0.041	0.126	0.033	0.101
	37.5%	0.043	0.129	0.038	0.110
	50%	0.048	0.136	0.045	0.121
h1	12.5%	0.100	0.201	0.113	0.217
	25%	0.108	0.209	0.131	0.231
	37.5%	0.122	0.220	0.153	0.247
	50%	0.144	0.237	0.182	0.266
h2	12.5%	0.075	0.175	0.067	0.155
	25%	0.076	0.177	0.071	0.160
	37.5%	0.093	0.195	0.077	0.167
	50%	0.089	0.192	0.086	0.179
AvgWins		58.3%		43.8%	

B. Zero-Shot Performance

Model		TOTEM		GPT2	
Metric		MSE	MAE	MSE	MAE
N2	12.5%	0.029	0.120	0.047	0.145
	25%	0.033	0.127	0.064	0.164
	37.5%	0.041	0.139	0.090	0.191
	50%	0.056	0.160	0.131	0.228
N5	12.5%	0.017	0.085	0.021	0.095
	25%	0.019	0.090	0.028	0.107
	37.5%	0.022	0.098	0.039	0.123
	50%	0.029	0.110	0.055	0.145
R	12.5%	0.071	0.109	0.093	0.119
	25%	0.087	0.117	0.125	0.134
	37.5%	0.112	0.129	0.167	0.154
	50%	0.148	0.147	0.220	0.182
B	12.5%	0.632	0.642	0.392	0.496
	25%	0.693	0.665	0.444	0.523
	37.5%	0.761	0.692	0.498	0.553
	50%	0.827	0.718	0.591	0.599
S	12.5%	0.057	0.160	0.070	0.173
	25%	0.061	0.168	0.084	0.189
	37.5%	0.069	0.178	0.103	0.209
	50%	0.082	0.193	0.128	0.234
AvgWins		80.0%		20.0%	

4 Conclusions, Limitations & Future Work

We present TOTEM: a simple, performant tokenizer that creates unified time series data representations across domains in many tasks thereby enabling generalist modeling. TOTEM demonstrates strong in-domain and zero-shot capabilities that match or outperform existing state-of-the-art approaches. Through dataset selection we emphasize the ability to train on varying domains and test on health domains. We leave discussion of anomaly detection E, forecasting F, ablations G, and further studies of generalist modeling H to the Appendix. Moving forward, an interesting limitation is that TOTEM does not support variable token lengths. Future work includes exploring dynamic token lengths as they could enhance unified representations and further improve task performance.

REFERENCES

Oliver D Anderson. Time-series. 2nd edn., 1976.

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.

Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6989–6997, 2023.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. *arXiv preprint arXiv:2303.06053*, 2023.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*, 2023.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems*, 32, 2019.

Philip Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Yangdong He and Jiabao Zhao. Temporal convolutional networks for anomaly detection in time series. In *Journal of Physics: Conference Series*, volume 1213, pp. 042050. IOP Publishing, 2019.

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

Charles C Holt. Forecasting trends and seasonals by exponentially weighted moving averages. *ONR Memorandum*, 52(52):5–10, 1957.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2021.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

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.

Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on linear mapping. *arXiv preprint arXiv:2305.10721*, 2023.

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*, 2021.

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, 2022b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. E2gan: End-to-end generative adversarial network for multivariate time series imputation. In *Proceedings of the 28th international joint conference on artificial intelligence*, pp. 3094–3100. AAAI Press Palo Alto, CA, USA, 2019.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.

Steven M Peterson, Satpreet H Singh, Benjamin Dichter, Michael Scheid, Rajesh PN Rao, and Bingni W Brunton. Ajile12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific data*, 9(1):184, 2022.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Kashif Rasul, Umang Gupta, Hena Ghonia, and Yuriy Nevmyvaka. Vq-tr: Vector quantized attention for time series forecasting. 2022a.

Kashif Rasul, Young-Jin Park, Max Nihlén Ramström, and Kyung-Min Kim. Vq-ar: Vector quantized autoregressive probabilistic time series forecasting. *arXiv preprint arXiv:2205.15894*, 2022b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Lifeng Shen, Zhuocong Li, and James Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems*, 33:13016–13026, 2020.

Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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.

Peter R Winters. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342, 1960.

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.

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, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022a.

Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. *arXiv preprint arXiv:2202.06258*, 2022b.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*, 2021.

Ling Yang and Shenda Hong. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In *International Conference on Machine Learning*, pp. 25038–25054. PMLR, 2022.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022.

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.

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*, 2022.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

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.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained lm. *arXiv preprint arXiv:2302.11939*, 2023.

APPENDIX

A REPRODUCIBILITY STATEMENT

To ensure reproducibility all results are run on three seeds; see section I for standard deviations. All code will be released. All datasets are already popular, public time series benchmark datasets. In imputation, anomaly detection, and forecasting the VQVAE is trained with a learning rate of 0.001, embedding dimension of 64, commitment cost of 0.25, and compression factor of 4. In forecasting the downstream model is a transformer encoder with 4 layers and 4 attention heads and a feed-forward hidden dimension of 256. We train using Adam with a base learning rate of 0.0001 and a one cycle learning rate scheduler in accordance with Nie et al. (2022) on A100s.

B ETHICAL CONSIDERATIONS

There are no immediate ethical concerns that arise from our work. However, as with all data driven methods, certain societal consequences are important to be discussed, in this case surrounding time series modeling. A few are reported below:

Privacy Concerns. Time series data, especially when sourced from personal devices or applications, can contain sensitive information about individuals, e.g. for health domains. In this work, no time series were sourced from personal devices, and all data is publicly available.

Reliability. Time series models can be unreliable. For instance, if a model forecasts incorrect health predictions, it could cause undue patient concern. In this work, we focused on unified data representations across many tasks as opposed to a single task.

C RELATED WORK

Time series modeling methods utilize many techniques, ranging from statistical methods (Winters, 1960; Holt, 1957; Anderson, 1976; Hyndman & Athanasopoulos, 2018; Taylor & Letham, 2018) to multilayer perceptrons (MLPs) (Zeng et al., 2023; Li et al., 2023; Das et al., 2023; Challu et al., 2023; Chen et al., 2023; Zhang et al., 2022; Oreshkin et al., 2019) to convolutional neural networks (CNNs) (Wu et al., 2022a; Liu et al., 2022a; He & Zhao, 2019; Franceschi et al., 2019; Bai et al., 2018) to recurrent neural networks (RNNs) (Salinas et al., 2020; Shen et al., 2020; Hochreiter & Schmidhuber, 1997) to transformers (Zhou et al., 2023; Liu et al., 2023; Nie et al., 2022; Zhang & Yan, 2022; Woo et al., 2022; Zhou et al., 2022; Liu et al., 2022b; Wu et al., 2022b; Xu et al., 2021; Wu et al., 2021; Liu et al., 2021; Zhou et al., 2021; Kitaev et al., 2020; Li et al., 2019). Many models are hybrid solutions that blend aforementioned approaches.

Most of these methods intake time and then perform various combinations of normalization (Kim et al., 2021), frequency transformations (Wu et al., 2022a; Zhou et al., 2022), and patchification either along the time dimension (Liu et al., 2023; Zhang & Yan, 2022; Nie et al., 2022), or sensor dimension (Li et al., 2019; Zhou et al., 2021; Wu et al., 2021; Liu et al., 2021).[2] Patch lengths range from a single time-step / sensor, also known as point-wise, to the length of the entire time series / all sensors. Time and sensor patch dependencies are then learned, via an attention mechanism, convolution, recurrence, or linear layer, across the temporal dimension, sensor dimension, or both the temporal and sensor dimensions (Zhang & Yan, 2022). For multisensor modeling, one can model all sensors jointly or independently (i.e., forecast each sensor independently) (Nie et al., 2022). These methods learn the underlying data representations end-to-end with the downstream task (e.g., forecasting).

Specialist-training, where models are only trained on a single time series domain, is the most common regime amongst prior work (Zhou et al., 2023; Wu et al., 2022a; Nie et al., 2022; Zhang & Yan, 2022). These specialist models are primarily evaluated via in-domain-testing, where the test set is from the same domain as the train set. Recently, some methods (Zhou et al., 2023; Liu et al., 2023) have begun to explore specialist zero-shot forecasting capabilities.

[2] In time series analysis, sensors, channels, and variates are synonymous terms; in this paper we adopt the sensor terminology.

The time series analysis field is undergoing unification along both the modeling axis (Zhou et al., 2023; Wu et al., 2022a) and data representation axis (Franceschi et al., 2019; Tonekaboni et al., 2021; Yang & Hong, 2022; Yue et al., 2022). Unified data representations, both statistical and learnt, have been more extensively studied in language and vision modeling (Gage, 1994; Van Den Oord et al., 2017; Esser et al., 2021; Rombach et al., 2022). The vision modeling field distinguishes between discrete, learnt, tokens (Van Den Oord et al., 2017; Esser et al., 2021; Rombach et al., 2022) and patches (Dosovitskiy et al., 2020). Patches have been studied in time series modeling (Zhou et al., 2023; Nie et al., 2022; Zhang & Yan, 2022). In this work, we propose to use discrete, learnt tokenized representations, which we show lead to strong performance in both specialist and generalist settings, as well as in-domain and zero-shot testing regimes.

D Experimental Setup

Through experiments in imputation (§3), anomaly detection (§E), and forecasting (§F), our goal is to explore the efficacy of TOTEM on standard benchmark datasets and tasks, and domain general settings. To briefly refresh: specialist refers to training on a single domain (Tables 1, 3, 5). Generalist refers to training on multiple domains (Tables 2, 4, 6). Finally, in-domain refers to testing on the training domain, and zero-shot to testing on a separate domain from training.

For all experiments & models, we run three seeds and report the mean; standard deviations are reported in section I. Following the field standard, we bold the **best** metric in all tables. Evaluation metrics differ across tasks. We report mean squared error `MSE` (↓), mean absolute error `MAE` (↓), precision `P` (↑), recall `R` (↑), and `F1` score (↑); (↓) means lower is better, (↑) means higher performance is better. Given the varied metrics we calculate the average number of best results, or `AvgWins` , for each method and highlight the **best**, **second** best, and **third** best methods.

Notably imputation and anomaly detection can be directly solved with just TOTEM's VQVAE, as they are fundamentally data representation tasks, whereas in forecasting further modeling is required, Figure 2. In forecasting, the trained, frozen, codebook representation converts a sensor's observed measurements $\mathbf{x}_s \in \mathbb{R}^{T_{\text{in}}}$ to a sequence of T_{in}/F discrete tokens.

Baselines. We compare to two families of approaches: methods designed for multiple tasks (**multitask**) – TOTEM belongs in this category – and methods designed for a specific task (**singletask**), and are adapted to other tasks.

We compare against two recent **multitask** models, the transformer based GPT2 Zhou et al. (2023) and the convolutional TimesNet[TiNet] Wu et al. (2022a). For **singletask** models we compare against PatchTST Patch] Nie et al. (2022), ETSFormer[ETS] Woo et al. (2022), Fedformer[FED] Zhou et al. (2022), Non-stationary trans.[Stat] Liu et al. (2022b), Autoformer[Auto] Wu et al. (2021), Informer[Inf] Zhou et al. (2021), Reformer[Re] Kitaev et al. (2020), LightTS[LiTS] Zhang et al. (2022), DLinear[DLin] Zeng et al. (2023), Anomaly trans.[ATran]Xu et al. (2021), Pyraformer[Pyra] Liu et al. (2021), LogTrans.[LogTr] Li et al. (2019), Trans.[Trans] Vaswani et al. (2017), Crossformer[Cross] Zhang & Yan (2022), TiDE Das et al. (2023), RLinear[RLin] Li et al. (2023), SciNet[SCi] Liu et al. (2022a), & iTrans.[iTrans] Liu et al. (2023).

Datasets. We leverage 12 benchmark datasets: weather[W], electricity[E], traffic[T], ETTm1[m1], ETTm2[m2], ETTh1[h1], ETTh2[h2], SMD, MSL, SMAP, SWAT, PSM that are commonly used for imputation, anomaly detection and forecasting Zhou et al. (2023); Wu et al. (2022a); Xu et al. (2021); Zhang & Yan (2022); Nie et al. (2022). For the zero shot settings, we leverage 5 benchmark datasets: neuro2[N2], neuro5[N5] (from Peterson et al. (2022)), and saugeen river flow[R], U.S. births[B], and sunspot[S] (from Godahewa et al. (2021)). 17 datasets in total.

E ANOMALY DETECTION

In anomaly detection, models intake a corrupted time series $\mathbf{x_{corr}} \in \mathbb{R}^{S \times T_{\text{in}}}$ and reconstruct the data $\mathbf{x} \in \mathbb{R}^{S \times T_{\text{in}}}$, where the amount of corruption is considered known, at A%. We report % Precision `P` (↑), Recall `R` (↑), and `F1` Score (↑); higher is better (↑).

The standard practice in machine learning, which we adopt, is to have a held out test set that is not used for tuning the model or learning algorithm. One aspect that makes comparing with several prior works challenging is that they use the test set as a validation set for early stopping of the learning algorithm, which can often inflate their performance. Despite this inconsistency, we compare our performance against these reported performances, whenever available.

Specialist. In Table 3 we evaluate TOTEM against numerous specialist baselines. TOTEM has the highest `AvgWins` at 26.7% followed by a five-way tie between GPT2, TiNet, ATrans, ETS, and LogTr at 13.3%. **Generalist.** In Table 4 we compare generalist-trained TOTEM and GPT2. On the in-domain test sets TOTEM outperforms GPT2: 80% vs. 20%. In the zero-shot test sets TOTEM outperforms GPT2: 73.3% vs. 26.7%.

TOTEM's `AvgWins` across the specialist and generalist settings demonstrate that tokens are a performant representation for anomaly detection.

Table 3: **Specialist Anomaly Detection** (↑)**.** TOTEM has the highest `AvgWins` at **26.7%** followed by a five-way tie between GPT2, TiNet, ATrans, ETS, and LogTr at **13.3%**. Some prior methods use the test set as a validation set for early stopping of the learning algorithm, which can inflate performance. We do not adopt this practice and train TOTEM for a set number of iterations.

Metric	Model	TOTEM	GPT2	TiNet	ATran	Patch	ETS	FED	Stat	Auto	Pyra	Inf	Re	LogTr	Trans	LiTS	DLin
F1	SMD	79.62	**86.89**	84.61	**85.49**	84.62	83.13	85.08	84.62	**85.11**	83.04	81.65	75.32	76.21	79.56	82.53	77.10
	MSL	82.58	82.45	81.84	83.31	78.70	**85.03**	78.57	77.50	79.05	**84.86**	84.06	84.40	79.57	78.68	78.95	**84.88**
	SMAP	**94.02**	**72.88**	69.39	**71.18**	68.82	69.50	70.76	71.09	71.12	71.09	69.92	70.40	69.97	69.70	69.21	69.26
	SWAT	**94.27**	**94.23**	93.02	83.10	85.72	84.91	93.19	79.88	92.74	91.78	81.43	82.80	80.52	80.37	**93.33**	87.52
	PSM	95.87	97.13	**97.34**	79.40	96.08	91.76	**97.23**	**97.29**	93.29	82.08	77.10	73.61	76.74	76.07	97.15	93.55
R	SMD	76.06	**84.98**	81.54	82.23	82.14	79.23	**82.39**	81.21	**82.35**	80.61	77.23	69.24	70.13	76.13	78.42	71.52
	MSL	82.85	82.91	75.36	**87.37**	70.96	84.93	80.07	**89.14**	80.92	85.93	86.48	83.31	**87.37**	**87.37**	75.78	85.42
	SMAP	**94.04**	**60.95**	56.40	58.11	55.46	55.75	58.10	**59.02**	58.62	57.71	57.13	57.44	57.59	57.12	55.27	55.41
	SWAT	95.91	96.34	95.40	**97.32**	80.94	80.36	96.42	**96.75**	95.81	96.00	**96.75**	96.53	**97.32**	96.53	94.72	95.30
	PSM	94.21	95.68	96.20	94.72	93.47	85.28	**97.16**	**96.76**	88.15	96.02	96.33	95.38	**98.00**	96.56	95.97	89.26
P	SMD	83.54	**88.89**	87.91	**88.91**	87.26	87.44	87.95	**88.33**	88.06	85.61	86.60	82.58	83.46	83.58	87.10	83.62
	MSL	82.32	82.00	**89.54**	79.61	**88.34**	85.13	77.14	68.55	77.27	83.81	81.77	**85.51**	73.05	71.57	82.40	84.34
	SMAP	**94.00**	90.60	90.14	91.85	90.64	92.25	90.47	89.37	90.40	**92.54**	90.11	90.91	89.15	89.37	**92.58**	92.32
	SWAT	**92.68**	**92.20**	90.75	72.51	91.10	90.02	90.17	68.03	89.85	87.92	70.29	72.50	68.67	68.84	**91.98**	80.91
	PSM	97.58	98.62	98.51	68.35	**98.84**	**99.31**	97.31	97.82	**99.08**	71.67	64.27	59.93	63.06	62.75	98.37	98.28
`AvgWins`		**26.7%**	**13.3%**	**13.3%**	**13.3%**	0%	**13.3%**	0%	6.7%	0%	0%	0%	0%	**13.3%**	0%	0%	0%

Table 4: **Generalist Anomaly Detection** (↑)**.** We train TOTEM & GPT2 on all datasets and then perform in-domain and zero-shot evaluations. **A. In-Domain Performance.** TOTEM outperforms GPT2: **80.0%** vs. 20.0%. **B. Zero-Shot Performance.** TOTEM again outperforms GPT2: **73.3%** vs. 26.7%.

A. In-Domain Performance

Metric	Model	TOTEM	GPT2
F1	SMD	78.64	**79.73**
	MSL	**83.29**	80.17
	SMAP	**92.51**	67.05
	SWAT	**94.37**	89.62
	PSM	**95.78**	90.47
R	SMD	72.07	**73.42**
	MSL	**82.96**	78.48
	SMAP	**91.48**	53.42
	SWAT	**96.13**	87.53
	PSM	**93.90**	87.76
P	SMD	86.66	**87.44**
	MSL	**83.64**	81.95
	SMAP	**93.56**	90.01
	SWAT	**92.68**	91.83
	PSM	**97.74**	93.39
`AvgWins`		**80.0%**	20.0%

B. Zero-Shot Performance

Metric	Model	TOTEM	GPT2
F1	N2	**51.29**	39.02
	N5	**51.28**	42.19
	R	**49.39**	36.14
	B	**49.15**	20.81
	S	**52.17**	38.12
R	N2	**76.88**	33.69
	N5	**76.84**	36.77
	R	**70.49**	29.66
	B	**73.71**	17.67
	S	**77.36**	31.83
P	N2	38.49	**46.43**
	N5	38.48	**49.58**
	R	38.02	**46.30**
	B	**36.86**	25.33
	S	39.35	**47.72**
`AvgWins`		**73.3%**	26.7%

F FORECASTING

The forecaster transformer encoder processes the tokenized time series independently for each sensor, adding time-based positional encodings to each token along the time dimension. Using a series of multi-head attention layers, the model predicts the forecasted measurements $\bar{\mathbf{y}}_s \in \mathbb{R}^{T_{\text{out}}}$ for $s = 1, ..., S$, applying the attention mechanism along the time dimension T. In parallel, the forecaster takes in $\mathbf{x}_s$ and predicts the future's mean, μ_s, and standard deviation, σ_s, for each sensor $s = 1, ..., S$ to unnormalize the data. The final forecasted prediction is $\mathbf{y}_s = \sigma_s \cdot \bar{\mathbf{y}}_s + \mu_s$. The forecaster is trained in a supervised fashion by minimizing three smooth L1 losses between predictions $\{\bar{\mathbf{y}}_s, \mu_s, \sigma_s\}$ and their ground truth respectively.

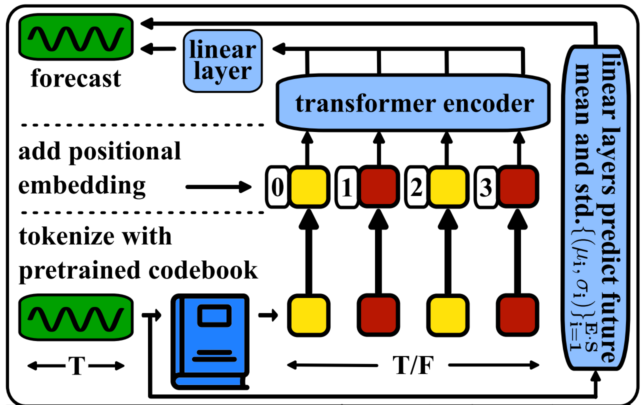

Figure 2: **Forecaster Modeling.** The forecasting task requires modeling beyond the VQVAE. We leverage TOTEM's pretrained, learnt, discrete codes as a the input data representation and train a transformer encoder. We add positional embeddings along the time dimension, and use linear layers before the final output as well as to unnormalize the resulting forecast.

In forecasting, models intake a time series $\mathbf{x} \in \mathbb{R}^{S \times T_{\text{in}}}$ and predict future readings $\mathbf{y} \in \mathbb{R}^{S \times T_{\text{out}}}$, where S is the number of sensors and $T_{\text{in}}, T_{\text{out}}$ signify the durations of the preceding and succeeding time series, respectively. The pairs $(\mathbf{x}, \mathbf{y})$ are generated by striding the original time series data.

All models have a lookback of $T_{\text{in}} = 96$, with prediction lengths $T_{\text{out}} = \{96, 192, 336, 720\}$. Numbers for other methods are from Liu et al. (2023). We run GPT2 with $T_{\text{in}} = 96$ as they originally report varying, dataset-specific, lookback lengths. We report `MSE` (↓) and `MAE` (↓); lower is better.

Specialist. From Table 5 we find that TOTEM achieves the highest `AvgWins` at 28.6% followed by iTrans at 26.8%. TOTEM has first finishes in five datasets while iTrans' first finishes are only electricity and traffic. **Generalist.** In Table 6 we compare generalist TOTEM and GPT2. TOTEM outperforms GPT2 for both in-domain (67.9% vs. 33.9%) and zero-shot (90.0% vs. 12.5%).

TOTEM's `AvgWins` forecasting performance across the training and testing regimes demonstrates that tokens are a performant representation for forecasting.

Table 5: **Specialist Forecasting** (↓)**.** TOTEM has the best `AvgWins` (**28.6%**), followed by iTrans (**26.8%**). Notably, TOTEM has first place finishes in 5 datasets, while iTrans' first places are concentrated in only electricity and traffic. All models have lookback $T_{in} = 96$.

Model		TOTEM		GPT2		TiNet		iTrans		Patch		Cross		FED		Stat		TiDE		RLin		DLin		SCi	
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
W	96	**0.165**	**0.208**	0.184	0.224	**0.172**	0.220	0.174	**0.214**	0.177	**0.218**	**0.158**	0.230	0.217	0.296	0.173	0.223	0.202	0.261	0.192	0.232	0.196	0.255	0.221	0.306
	192	**0.207**	**0.250**	0.231	0.263	**0.219**	0.261	0.221	**0.254**	0.225	**0.259**	**0.206**	0.277	0.276	0.336	0.245	0.285	0.242	0.298	0.240	0.271	0.237	0.296	0.261	0.340
	336	**0.257**	**0.291**	0.285	0.302	0.280	0.306	**0.278**	**0.296**	**0.278**	**0.297**	**0.272**	0.335	0.339	0.380	0.321	0.338	0.287	0.335	0.292	0.307	0.283	0.335	0.309	0.378
	720	**0.326**	**0.340**	0.362	0.351	0.365	0.359	0.358	**0.349**	0.354	**0.348**	0.398	0.418	0.403	0.428	0.414	0.410	**0.351**	0.386	0.364	0.353	**0.345**	0.381	0.377	0.427
E	96	0.178	**0.263**	0.186	**0.272**	**0.168**	**0.272**	**0.148**	**0.240**	0.195	0.285	0.219	0.314	0.193	0.308	**0.169**	0.273	0.237	0.329	0.201	0.281	0.197	0.282	0.247	0.345
	192	0.187	**0.272**	0.190	**0.278**	**0.184**	0.289	**0.162**	**0.253**	0.199	0.289	0.231	0.322	0.201	0.315	**0.182**	0.286	0.236	0.330	0.201	0.283	0.196	0.285	0.257	0.355
	336	**0.199**	**0.285**	0.204	**0.291**	**0.198**	0.300	**0.178**	**0.269**	0.215	0.305	0.246	0.337	0.214	0.329	0.200	0.304	0.249	0.344	0.215	0.298	0.209	0.301	0.269	0.369
	720	0.236	**0.318**	0.245	0.324	**0.220**	**0.320**	**0.225**	**0.317**	0.256	0.337	0.280	0.363	0.246	0.355	**0.222**	0.321	0.284	0.373	0.257	0.331	0.245	0.333	0.299	0.390
T	96	0.523	**0.303**	**0.471**	0.311	0.593	0.321	**0.395**	**0.268**	0.544	0.359	**0.522**	**0.290**	0.587	0.366	0.612	0.338	0.805	0.493	0.649	0.389	0.650	0.396	0.788	0.499
	192	**0.530**	**0.303**	**0.479**	0.312	0.617	0.336	**0.417**	**0.276**	0.540	0.354	**0.530**	**0.293**	0.604	0.373	0.613	0.340	0.756	0.474	0.601	0.366	0.598	0.370	0.789	0.505
	336	**0.549**	**0.311**	**0.490**	0.317	0.629	0.336	**0.433**	**0.283**	0.551	0.358	0.558	**0.305**	0.621	0.383	0.618	0.328	0.762	0.477	0.609	0.369	0.605	0.373	0.797	0.508
	720	0.598	**0.331**	**0.524**	0.336	0.640	0.350	**0.467**	**0.302**	**0.586**	0.375	0.589	**0.328**	0.626	0.382	0.653	0.355	0.719	0.449	0.647	0.387	0.645	0.394	0.841	0.523
m1	96	**0.320**	**0.347**	**0.328**	**0.363**	0.338	0.375	0.334	0.368	**0.329**	**0.367**	0.404	0.426	0.379	0.419	0.386	0.398	0.364	0.387	0.355	0.376	0.345	0.372	0.418	0.438
	192	0.379	**0.382**	**0.368**	**0.382**	**0.374**	0.387	0.377	0.391	**0.367**	**0.385**	0.450	0.451	0.426	0.441	0.459	0.444	0.398	0.404	0.391	0.392	0.380	0.389	0.439	0.450
	336	**0.406**	**0.402**	**0.400**	**0.404**	0.410	0.411	0.426	0.420	**0.399**	**0.410**	0.532	0.515	0.445	0.459	0.495	0.464	0.428	0.425	0.424	0.415	0.413	0.413	0.490	0.485
	720	**0.471**	**0.438**	**0.462**	**0.440**	0.478	0.450	0.491	0.459	**0.454**	**0.439**	0.666	0.589	0.543	0.490	0.585	0.516	0.487	0.461	0.487	0.450	0.474	0.453	0.595	0.550
m2	96	**0.176**	**0.253**	**0.178**	**0.263**	0.187	0.267	0.180	0.264	**0.175**	**0.259**	0.287	0.366	0.203	0.287	0.192	0.274	0.207	0.305	0.182	0.265	0.193	0.292	0.286	0.377
	192	0.247	**0.302**	**0.245**	0.307	0.249	0.309	0.250	0.309	**0.241**	**0.302**	0.414	0.492	0.269	0.328	0.280	0.339	0.290	0.364	**0.246**	**0.304**	0.284	0.362	0.399	0.445
	336	0.317	0.348	**0.307**	**0.346**	0.321	0.351	0.311	0.348	**0.305**	**0.343**	0.597	0.542	0.325	0.366	0.334	0.361	0.377	0.422	**0.307**	**0.342**	0.369	0.427	0.637	0.591
	720	0.426	0.410	0.410	0.409	**0.408**	**0.403**	0.412	0.407	**0.402**	**0.400**	1.730	1.042	0.421	0.415	0.417	0.413	0.558	0.524	**0.407**	**0.398**	0.554	0.522	0.960	0.735
h1	96	**0.380**	**0.394**	**0.379**	**0.397**	0.384	0.402	0.386	0.405	0.414	0.419	0.423	0.448	**0.376**	0.419	0.513	0.491	0.479	0.464	0.386	**0.395**	0.386	0.400	0.654	0.599
	192	**0.434**	**0.427**	0.438	**0.427**	**0.436**	0.429	0.441	0.436	0.460	0.445	0.471	0.474	**0.420**	0.448	0.534	0.504	0.525	0.492	0.437	**0.424**	0.437	0.432	0.719	0.631
	336	0.490	0.459	**0.474**	**0.448**	0.491	0.469	0.487	**0.458**	0.501	0.466	0.570	0.546	**0.459**	0.465	0.588	0.535	0.565	0.515	**0.479**	**0.446**	0.481	0.459	0.778	0.659
	720	0.539	0.513	**0.496**	**0.475**	0.521	0.500	0.503	0.491	**0.500**	**0.488**	0.653	0.621	0.506	0.507	0.643	0.616	0.594	0.558	**0.481**	**0.470**	0.519	0.516	0.836	0.699
h2	96	**0.293**	**0.338**	**0.295**	**0.348**	0.340	0.374	0.297	0.349	0.302	**0.348**	0.745	0.584	0.358	0.397	0.476	0.458	0.400	0.440	**0.288**	**0.338**	0.333	0.387	0.707	0.621
	192	**0.375**	**0.390**	0.384	0.402	0.402	0.414	**0.380**	**0.400**	0.388	**0.400**	0.877	0.656	0.429	0.439	0.512	0.493	0.528	0.509	**0.374**	**0.390**	0.477	0.476	0.860	0.689
	336	**0.422**	**0.431**	**0.418**	**0.432**	0.452	0.452	0.428	**0.432**	0.426	0.433	1.043	0.731	0.496	0.487	0.552	0.551	0.643	0.571	**0.415**	**0.426**	0.594	0.541	1.000	0.744
	720	0.610	0.567	**0.423**	**0.446**	0.462	0.468	**0.427**	**0.445**	0.431	**0.446**	1.104	0.763	0.463	0.474	0.562	0.560	0.874	0.679	**0.420**	**0.440**	0.831	0.657	1.249	0.838
`AvgWins`		**28.6%**		1.8%		1.8%		**26.8%**		14.3%		3.6%		5.4%		0%		0%		**25%**		0%		0%	

Table 6: **Generalist Forecasting** (↓). We evaluate generalist TOTEM and GPT2. **A. In-Domain.** TOTEM outperforms GPT2: 67.9% to 33.9%. **B. Zero-Shot.** TOTEM outperforms GPT2: 90.0% to 12.5%.

A. In-Domain Performance

Model		TOTEM		GPT2	
Metric		MSE	MAE	MSE	MAE
W	96	**0.172**	**0.216**	0.201	0.237
	192	**0.217**	**0.256**	0.247	0.275
	336	**0.266**	**0.295**	0.298	0.311
	720	**0.334**	**0.342**	0.372	0.360
E	96	**0.179**	**0.264**	0.194	0.278
	192	**0.181**	**0.267**	0.199	0.284
	336	**0.196**	**0.283**	0.214	0.300
	720	**0.230**	**0.314**	0.255	0.331
T	96	0.507	**0.284**	**0.484**	0.320
	192	0.511	**0.282**	**0.488**	0.320
	336	0.535	**0.292**	**0.502**	0.326
	720	0.580	**0.309**	**0.534**	0.343
m1	96	**0.374**	**0.384**	0.487	0.468
	192	**0.400**	**0.399**	0.516	0.480
	336	**0.432**	**0.424**	0.548	0.499
	720	**0.487**	**0.460**	0.581	0.511
m2	96	**0.198**	**0.275**	0.243	0.315
	192	**0.266**	**0.319**	0.297	0.346
	336	0.365	0.377	**0.349**	**0.376**
	720	0.588	0.511	**0.439**	**0.423**
h1	96	**0.382**	**0.404**	0.421	0.408
	192	**0.463**	**0.435**	0.480	0.436
	336	**0.507**	0.463	0.518	**0.453**
	720	**0.517**	0.500	**0.517**	**0.467**
h2	96	0.307	0.345	**0.298**	**0.343**
	192	0.406	0.403	**0.381**	**0.392**
	336	0.505	0.460	**0.406**	**0.419**
	720	0.661	0.557	**0.423**	**0.438**
AvgWins		**67.9%**		33.9%	

B. Zero-Shot Performance

Model		TOTEM		GPT2	
Metric		MSE	MAE	MSE	MAE
N2	96	**1.138**	**0.777**	1.332	0.830
	192	**1.149**	**0.785**	1.416	0.863
	336	**1.092**	**0.770**	1.358	0.851
	720	**1.045**	**0.754**	1.308	0.840
N5	96	**0.483**	**0.484**	0.528	0.499
	192	**0.495**	**0.491**	0.578	0.524
	336	**0.468**	**0.483**	0.548	0.515
	720	**0.451**	**0.477**	0.537	0.511
R	96	**1.120**	**0.582**	1.465	0.725
	192	**1.242**	**0.635**	1.638	0.785
	336	**1.237**	**0.626**	1.601	0.769
	720	**1.182**	**0.604**	1.552	0.760
B	96	**0.805**	**0.739**	0.838	0.762
	192	**0.836**	**0.752**	0.837	**0.752**
	336	0.809	0.748	**0.792**	**0.738**
	720	**0.896**	**0.794**	0.927	0.806
S	96	0.446	0.482	**0.443**	**0.478**
	192	**0.462**	**0.491**	0.481	0.499
	336	**0.521**	**0.525**	0.541	0.533
	720	**0.717**	**0.625**	0.773	0.643
AvgWins		**90.0%**		12.5%	

G Ablations

Tokens vs. Time. To evaluate if tokens enable TOTEM's performance, we implement TimeTOTEM. TimeTOTEM has the identical architecture to TOTEM, except we replace the VQVAE with an MLP trained end-to-end with the downstream forecaster. We compare Totem vs. TimeTOTEM in the specialist in-domain, and generalist in-domain and zero-shot regimes (Table 7). In all cases TOTEM outperforms TimeTOTEM - specialist: 67.9% vs. 39.3%, generalist in-domain: 78.6% vs. 23.2%, generalist zero-shot: 67.5% vs. 35.0%. TOTEM's performance demonstrates that tokens, when compared to time, lead to better performance.

Codebook Size. In Table 7 we explore the affect of the codebook size, K, on the VQVAE's `MSE` and `MAE` reconstruction performance. As expected, we find that as K increases from 32 to 256 to 512 the reconstruction performance improves.

Table 7: **Ablations** (↓). Across the Tokens vs. Time (TvT) experiments tokens out perform time. (A) specialist: 67.9% to 39.93%, (B) in-domain generalist: 78.6% to 23.2% , and (C) zero-shot generalist: 67.5% to 35%. (D) As the codebook size K increases the VQVAE reconstruction performance improves.

A. TvT Specialist

Model		TOTEM		TimeTOTEM	
Metric		MSE	MAE	MSE	MAE
W	96	0.165	**0.208**	**0.164**	0.209
	192	**0.207**	**0.250**	0.209	0.251
	336	**0.257**	**0.291**	0.261	0.293
	720	**0.326**	**0.340**	0.332	**0.340**
E	96	**0.178**	0.263	0.179	**0.262**
	192	0.187	0.272	**0.185**	**0.269**
	336	**0.199**	**0.285**	0.204	0.289
	720	**0.236**	**0.318**	0.244	0.325
T	96	**0.523**	**0.303**	0.528	0.310
	192	0.530	**0.303**	**0.500**	0.349
	336	0.549	**0.311**	**0.531**	0.365
	720	0.598	**0.331**	**0.578**	0.398
m1	96	**0.320**	**0.347**	0.326	0.355
	192	0.379	**0.382**	**0.377**	0.386
	336	**0.406**	**0.402**	0.409	0.409
	720	0.471	**0.438**	**0.469**	0.441
m2	96	**0.176**	**0.253**	**0.176**	0.254
	192	**0.247**	**0.302**	**0.247**	0.303
	336	**0.317**	**0.348**	0.318	0.350
	720	0.426	**0.410**	**0.419**	0.411
h1	96	0.380	**0.394**	**0.377**	0.395
	192	0.434	**0.427**	**0.428**	0.428
	336	0.490	**0.459**	**0.480**	0.462
	720	0.539	**0.513**	**0.530**	0.522
h2	96	**0.293**	**0.338**	0.294	**0.338**
	192	0.375	0.390	**0.373**	**0.389**
	336	**0.422**	**0.431**	0.423	0.433
	720	0.610	0.567	**0.591**	**0.556**
AvgWins		**67.9%**		39.3%	

B. TvT In-Domain Generalist

Model		TOTEM		TimeTOTEM	
Metric		MSE	MAE	MSE	MAE
W	96	**0.172**	**0.216**	0.173	0.218
	192	**0.217**	**0.256**	0.218	0.261
	336	**0.266**	**0.295**	0.267	0.299
	720	**0.334**	**0.342**	0.337	0.347
E	96	**0.179**	**0.264**	0.183	0.267
	192	**0.181**	**0.267**	0.189	0.275
	336	**0.196**	**0.283**	0.204	0.291
	720	**0.230**	**0.314**	0.242	0.325
T	96	**0.507**	**0.284**	0.517	0.293
	192	**0.511**	**0.282**	0.526	0.296
	336	**0.535**	**0.292**	0.552	0.304
	720	**0.580**	**0.309**	0.602	0.326
m1	96	**0.374**	**0.384**	0.428	0.420
	192	**0.400**	**0.399**	0.438	0.427
	336	**0.432**	**0.424**	0.469	0.447
	720	**0.487**	**0.460**	0.546	0.493
m2	96	**0.198**	**0.275**	0.207	0.286
	192	**0.266**	**0.319**	0.269	0.325
	336	0.365	**0.377**	**0.358**	**0.377**
	720	0.588	0.511	**0.521**	**0.482**
h1	96	**0.382**	**0.404**	0.401	0.410
	192	0.463	**0.435**	**0.453**	0.441
	336	0.507	**0.463**	**0.496**	0.468
	720	**0.517**	**0.500**	0.518	0.510
h2	96	0.307	**0.345**	**0.305**	0.346
	192	0.406	0.403	**0.396**	**0.402**
	336	0.505	0.460	**0.492**	**0.458**
	720	0.661	0.557	**0.599**	**0.531**
AvgWins		**78.6%**		23.2%	

C. TvT Zero-Shot Generalist

Model		TOTEM		TimeTOTEM	
Metric		MSE	MAE	MSE	MAE
N2	96	1.138	0.777	**1.127**	**0.773**
	192	**1.149**	**0.785**	1.169	0.793
	336	**1.092**	**0.770**	1.115	0.780
	720	**1.045**	**0.754**	1.070	0.766
N5	96	0.483	0.484	**0.481**	**0.483**
	192	**0.495**	**0.491**	0.508	0.500
	336	**0.468**	**0.483**	0.481	0.491
	720	**0.451**	**0.477**	0.467	0.488
R	96	1.120	0.582	**1.102**	**0.578**
	192	1.242	0.635	**1.207**	**0.628**
	336	1.237	0.626	**1.190**	**0.613**
	720	1.182	0.604	**1.149**	**0.596**
B	96	**0.805**	**0.739**	0.825	0.751
	192	**0.836**	**0.752**	0.847	0.761
	336	**0.809**	**0.748**	0.831	0.764
	720	**0.896**	**0.794**	0.928	0.813
S	96	**0.446**	0.482	**0.446**	**0.481**
	192	**0.462**	**0.491**	0.478	0.499
	336	**0.521**	**0.525**	0.535	0.532
	720	**0.717**	**0.625**	0.736	0.631
AvgWins		**67.5%**		35.0%	

D. Codebook Size Ablations

	Codebook Size K		
	32	256	512
	MSE		
All	0.0451	0.0192	**0.0184**
W	0.0393	0.0161	**0.0128**
E	0.0463	0.0209	**0.0152**
T	0.0312	0.0120	**0.0101**
	MAE		
All	0.1460	0.0937	**0.0913**
W	0.1122	0.0673	**0.0607**
E	0.1520	0.1027	**0.0878**
T	0.1204	0.0749	**0.0685**
AvgWins	0%	0%	100%

H EXPLORATORY STUDIES IN GENERALIST MODELING

Generalist Codebooks. To further explore the capabilities of a generalist codebook data representation we train models that utilize a general codebook but dataset-specific transformer forecasters, e.g. a TOTEM VQVAE trained on multiple domains with a forecaster trained only on electricity, Table 8. We compare these mixed models to generalist and specialist models trained on the same domains. All models use the same the codebook hyperparameters (number of codewords $K = 256$, compression factor $F = 4$, code dimensionality $D = 64$) as well as the forecaster transformer architecture to ensure a fair comparison.

Since we are evaluating the specialists, mixed-models, and generalist on in-domain test data one might expect that the TOTEM specialists will significantly outperform all models. Surprisingly this intuition is not correct. When comparing models trained using specialist codebooks to models trained using a single generalist codebook we find that generalist codebook models outperform specialist codebook models: 66.1% vs. 57.1%. Upon further inspection we find that the fully-generalist model (far right column Table 8) significantly outperforms the mixed-models (middle column Table 8) in traffic (T) and electricity (E). This dominant performance is puzzling until considering the training sizes.

The largest training set across domains belongs to traffic (T) at $10.2M$ training examples. In dataset T, the fully generalist models achieves 100% `AvgWins` . The second largest training set belongs to electricity (E) at $5.8M$ training examples, with 75% `AvgWins` for the fully-generalist model. Unfortunately there is a sharp drop off in training set sizes, with the rest of the data domains collectively comprising $1.6M$ training examples. These results evoke questions. For instance: does training on the smaller datasets act like form of regularization? Or: how does in-domain generalist performance scale with dataset size? We leave these exciting directions for future work. The generalist codebook's performance across datasets highlights the potential of unified, discrete, token representations for in-domain evaluations.

Zero Shot Vignette: Training Size & Data Diversity. Here we further explore generalist and specialist zero-shot testing capabilites, Table 9. We take the two largest TOTEM specialist, traffic at $10.2M$ and electricity at $5.8M$ training examples, and test their zero-shot capabilities compared to the TOTEM generalist. We expect that the generalist will perform best as it was trained on the most data at $17.6M$ training examples as well as the most domains. We predict the generalist will be followed by TOTEM-traffic then TOTEM-electricity as they are both trained on only one domain but traffic has $4.4M$ more training examples than electricity. As expected the generalist outperforms both TOTEM-traffic and TOTEM-electricity with 85.0% `AvgWins` . However, curiously TOTEM-electricity outperforms TOTEM-traffic: 12.5% vs. 2.5% despite having $4.4M$ fewer training examples. Why is the smaller training set outperforming the larger training set? One possible explanation is that the electricity domain is more similar than the traffic domain to neuro, river, births, and sunspot. Another possible explanation comes from the raw time series dimensionality. Despite having fewer training examples, electricity has a higher number of raw time steps[3] compared to traffic: 26304 vs. 17544. However, traffic has a larger number of sensors: 862 vs. 321. This limited analysis suggests that a higher number of raw time steps is more valuable than more sensor readings. Untangling these possibilities and beginning to answer the questions: what is a unit of data in time series? And how this unit scale as the time steps, sensors, and examples scale? are valuable future directions. The zero shot vignette has demonstrated the power of the token-enabled generalist over the traffic and electricity specialists, and has opened up exciting training size and data diversity questions.

[3]Raw time steps for all data. The train:val:test ratio is 7:1:2.

Table 8: Generalist codes beat specialist codes: 66.1% vs 57.1%.

Codebook		Specialist		Generalist		Generalist	
Forecaster		Specialist		Specialist		Generalist	
Metric		MSE	MAE	MSE	MAE	MSE	MAE
W	96	0.165	0.208	0.164	0.208	0.172	0.216
	192	0.207	0.250	0.208	0.251	0.217	0.256
	336	0.257	0.291	0.258	0.290	0.266	0.295
	720	0.326	0.340	0.329	0.338	0.334	0.342
E	96	0.178	0.263	0.178	0.263	0.179	0.264
	192	0.187	0.272	0.187	0.273	0.181	0.267
	336	0.199	0.285	0.199	0.285	0.196	0.283
	720	0.236	0.318	0.238	0.320	0.230	0.314
T	96	0.523	0.303	0.521	0.301	0.507	0.284
	192	0.530	0.303	0.530	0.303	0.511	0.282
	336	0.549	0.311	0.555	0.313	0.535	0.292
	720	0.598	0.331	0.605	0.337	0.580	0.309
m1	96	0.320	0.347	0.328	0.352	0.374	0.384
	192	0.379	0.382	0.377	0.383	0.400	0.399
	336	0.406	0.402	0.408	0.404	0.432	0.424
	720	0.471	0.438	0.470	0.440	0.487	0.460
m2	96	0.176	0.253	0.175	0.253	0.198	0.275
	192	0.247	0.302	0.247	0.302	0.266	0.319
	336	0.317	0.348	0.318	0.348	0.365	0.377
	720	0.426	0.410	0.427	0.410	0.588	0.511
h1	96	0.380	0.394	0.382	0.395	0.382	0.404
	192	0.434	0.427	0.437	0.427	0.463	0.435
	336	0.490	0.459	0.490	0.460	0.507	0.463
	720	0.539	0.513	0.536	0.512	0.517	0.500
h2	96	0.293	0.338	0.294	0.339	0.307	0.345
	192	0.375	0.390	0.375	0.391	0.406	0.403
	336	0.422	0.431	0.421	0.431	0.505	0.460
	720	0.610	0.567	0.610	0.567	0.661	0.557
AvgWins		57.1%		66.1%			

Table 9: Zero Shot Vignette: Training Size & Diversity

Model		TOTEM Generalist		TOTEM Specialist		TOTEM Specialist	
Train Domain		ALL		Traffic		Electricity	
Sensor Num (S)		-		862		321	
Raw Length (T)		-		17544		26304	
Train Size		17.6M		10.2M		5.8M	
Metric		MSE	MAE	MSE	MAE	MSE	MAE
N2	96	1.138	0.777	1.194	0.798	1.193	0.802
	192	1.149	0.785	1.218	0.808	1.300	0.845
	336	1.092	0.770	1.190	0.804	1.260	0.837
	720	1.045	0.754	1.117	0.784	1.234	0.832
N5	96	0.483	0.484	0.515	0.505	0.489	0.490
	192	0.495	0.491	0.535	0.514	0.555	0.527
	336	0.468	0.483	0.524	0.513	0.538	0.525
	720	0.451	0.477	0.500	0.507	0.533	0.527
R	96	1.120	0.582	1.171	0.635	1.141	0.579
	192	1.242	0.635	1.273	0.673	1.297	0.652
	336	1.237	0.626	1.232	0.653	1.247	0.628
	720	1.182	0.604	1.198	0.642	1.236	0.633
B	96	0.805	0.739	0.812	0.749	0.820	0.756
	192	0.836	0.752	0.858	0.767	0.843	0.759
	336	0.809	0.748	0.826	0.759	0.791	0.741
	720	0.896	0.794	0.919	0.803	0.886	0.790
S	96	0.446	0.482	0.476	0.508	0.460	0.487
	192	0.462	0.491	0.511	0.528	0.505	0.511
	336	0.521	0.525	0.576	0.568	0.569	0.545
	720	0.717	0.625	0.795	0.685	0.764	0.641
AvgWins		**85.0%**		**2.5%**		**12.5%**	

I MEANS AND STANDARD DEVIATIONS

I.1 IMPUTATION RESULTS - MEANS AND STANDARD DEVIATIONS

Table 10: **TOTEM - Specialist Imputation** (↓)

Metric		MSE	MAE
W	12.5%	0.028 ± 0.0000	0.046 ± 0.0006
	25%	0.028 ± 0.0000	0.046 ± 0.0006
	37.5%	0.029 ± 0.0000	0.047 ± 0.0010
	50%	0.031 ± 0.0006	0.048 ± 0.00015
E	12.5%	0.054 ± 0.0006	0.154 ± 0.0015
	25%	0.059 ± 0.0006	0.160 ± 0.0010
	37.5%	0.067 ± 0.006	0.169 ± 0.0012
	50%	0.079 ± 0.0012	0.183 ± 0.0012
m1	12.5%	0.049 ± 0.0000	0.125 ± 0.0006
	25%	0.052 ± 0.0006	0.128 ± 0.0006
	37.5%	0.055 ± 0.0000	0.132 ± 0.0006
	50%	0.061 ± 0.0006	0.139 ± 0.0006
m2	12.5%	0.016 ± 0.0006	0.078 ± 0.0010
	25%	0.017 ± 0.0006	0.081 ± 0.0006
	37.5%	0.018 ± 0.0000	0.084 ± 0.0006
	50%	0.020 ± 0.0000	0.088 ± 0.0000
h1	12.5%	0.119 ± 0.0010	0.212 ± 0.0006
	25%	0.127 ± 0.0015	0.220 ± 0.0006
	37.5%	0.138 ± 0.0012	0.230 ± 0.0006
	50%	0.157 ± 0.0006	0.247 ± 0.0010
h2	12.5%	0.040 ± 0.0006	0.129 ± 0.0017
	25%	0.041 ± 0.0010	0.131 ± 0.0012
	37.5%	0.043 ± 0.0006	0.136 ± 0.0006
	50%	0.047 ± 0.0006	0.142 ± 0.0012

Table 11: **TOTEM - Generalist Imputation** (↓)

Metric		MSE	MAE
W	12.5%	0.029 ± 0.0012	0.060 ± 0.0047
	25%	0.030 ± 0.0006	0.060 ± 0.0047
	37.5%	0.032 ± 0.0006	0.062 ± 0.0030
	50%	0.036 ± 0.0006	0.067 ± 0.00036
E	12.5%	0.065 ± 0.0020	0.171 ± 0.0032
	25%	0.071 ± 0.0015	0.179 ± 0.0031
	37.5%	0.080 ± 0.0025	0.189 ± 0.0032
	50%	0.095 ± 0.0026	0.205 ± 0.0032
m1	12.5%	0.041 ± 0.0006	0.132 ± 0.0015
	25%	0.044 ± 0.0000	0.135 ± 0.0010
	37.5%	0.048 ± 0.0006	0.139 ± 0.0040
	50%	0.058 ± 0.0010	0.152 ± 0.0000
m2	12.5%	0.040 ± 0.0020	0.125 ± 0.0067
	25%	0.041 ± 0.0015	0.126 ± 0.0058
	37.5%	0.043 ± 0.0015	0.129 ± 0.0049
	50%	0.048 ± 0.0010	0.136 ± 0.0038
h1	12.5%	0.100 ± 0.0049	0.201 ± 0.0049
	25%	0.108 ± 0.0049	0.209 ± 0.0038
	37.5%	0.122 ± 0.0064	0.220 ± 0.0044
	50%	0.144 ± 0.0078	0.237 ± 0.0049
h2	12.5%	0.075 ± 0.0012	0.175 ± 0.0053
	25%	0.076 ± 0.0006	0.177 ± 0.0036
	37.5%	0.093 ± 0.0222	0.195 ± 0.0200
	50%	0.089 ± 0.0010	0.192 ± 0.0035
Zero-Shot			
N2	12.5%	0.029 ± 0.0015	0.120 ± 0.0045
	25%	0.033 ± 0.0010	0.127 ± 0.0035
	37.5%	0.041 ± 0.0006	0.139 ± 0.0025
	50%	0.056 ± 0.0006	0.160 ± 0.0012
N5	12.5%	0.017 ± 0.0010	0.085 ± 0.0030
	25%	0.019 ± 0.0010	0.090 ± 0.0030
	37.5%	0.022 ± 0.0006	0.098 ± 0.0025
	50%	0.029 ± 0.0006	0.110 ± 0.0025
R	12.5%	0.071 ± 0.0070	0.109 ± 0.0040
	25%	0.087 ± 0.0064	0.117 ± 0.0031
	37.5%	0.112 ± 0.0050	0.129 ± 0.0035
	50%	0.148 ± 0.0032	0.147 ± 0.0023
B	12.5%	0.632 ± 0.0087	0.642 ± 0.0068
	25%	0.693 ± 0.0070	0.665 ± 0.0047
	37.5%	0.761 ± 0.0055	0.692 ± 0.0023
	50%	0.827 ± 0.0044	0.718 ± 0.0000
S	12.5%	0.057 ± 0.0012	0.160 ± 0.0023
	25%	0.061 ± 0.0006	0.168 ± 0.0021
	37.5%	0.069 ± 0.0006	0.178 ± 0.0021
	50%	0.082 ± 0.0010	0.193 ± 0.0015

Table 12: **GPT2 - Generalist Imputation** (↓)

Metric		MSE	MAE
W	12.5%	0.029 ± 0.0000	0.045 ± 0.0006
	25%	0.033 ± 0.0006	0.048 ± 0.0006
	37.5%	0.037 ± 0.0006	0.054 ± 0.0012
	50%	0.043 ± 0.0012	0.061 ± 0.00017
E	12.5%	0.008 ± 0.0020	0.186 ± 0.0035
	25%	0.091 ± 0.0020	0.197 ± 0.0025
	37.5%	0.108 ± 0.0021	0.213 ± 0.0026
	50%	0.132 ± 0.0026	0.236 ± 0.0026
m1	12.5%	0.052 ± 0.0012	0.141 ± 0.0016
	25%	0.065 ± 0.0021	0.154 ± 0.0021
	37.5%	0.085 ± 0.0038	0.171 ± 0.0026
	50%	0.117 ± 0.0052	0.196 ± 0.0026
m2	12.5%	0.029 ± 0.0000	0.095 ± 0.0006
	25%	0.033 ± 0.0006	0.101 ± 0.0006
	37.5%	0.038 ± 0.0006	0.110 ± 0.0012
	50%	0.045 ± 0.0006	0.121 ± 0.0012
h1	12.5%	0.113 ± 0.0012	0.217 ± 0.0021
	25%	0.131 ± 0.0010	0.231 ± 0.0015
	37.5%	0.153 ± 0.0012	0.247 ± 0.0017
	50%	0.182 ± 0.0006	0.266 ± 0.0012
h2	12.5%	0.067 ± 0.0010	0.155 ± 0.0015
	25%	0.071 ± 0.0006	0.160 ± 0.0015
	37.5%	0.077 ± 0.010	0.167 ± 0.0015
	50%	0.086 ± 0.0032	0.179 ± 0.0038
Zero-Shot			
N2	12.5%	0.047 ± 0.0006	0.145 ± 0.0015
	25%	0.064 ± 0.0017	0.164 ± 0.0015
	37.5%	0.090 ± 0.0036	0.191 ± 0.0032
	50%	0.131 ± 0.0051	0.228 ± 0.0044
N5	12.5%	0.021 ± 0.0006	0.095 ± 0.0012
	25%	0.028 ± 0.0006	0.107 ± 0.0010
	37.5%	0.039 ± 0.0015	0.123 ± 0.0015
	50%	0.055 ± 0.0015	0.145 ± 0.0023
R	12.5%	0.093 ± 0.0010	0.119 ± 0.0015
	25%	0.125 ± 0.0006	0.134 ± 0.0026
	37.5%	0.167 ± 0.0021	0.154 ± 0.0042
	50%	0.220 ± 0.0045	0.182 ± 0.0057
B	12.5%	0.392 ± 0.0064	0.496 ± 0.0023
	25%	0.444 ± 0.0071	0.523 ± 0.0029
	37.5%	0.498 ± 0.0080	0.553 ± 0.0023
	50%	0.591 ± 0.0700	0.599 ± 0.0275
s	12.5%	0.070 ± 0.0012	0.173 ± 0.0017
	25%	0.084 ± 0.0010	0.189 ± 0.0015
	37.5%	0.103 ± 0.0010	0.209 ± 0.0021
	50%	0.128 ± 0.0015	0.234 ± 0.0021

I.2 Anomaly Detection Results - Means and Standard Deviations

Table 13: **TOTEM - Specialist Anomaly Detection** (↑)

		Mean ± Std
F1	SMD	0.796 ± 0.0137
	MSL	0.826 ± 0.0052
	SMAP	0.940 ± 0.0008
	SWAT	0.943 ± 0.0006
	PSM	0.959 ± 0.0008
R	SMD	0.761 ± 0.0207
	MSL	0.829 ± 0.0071
	SMAP	0.940 ± 0.0013
	SWAT	0.959 ± 0.0012
	PSM	0.942 ± 0.0004
P	SMD	0.835 ± 0.0054
	MSL	0.823 ± 0.0033
	SMAP	0.940 ± 0.0004
	SWAT	0.927 ± 0.0003
	PSM	0.976 ± 0.0012

Table 14: **TOTEM - Generalist Anomaly Detection** (↑)

		Mean ± Std
F1	SMD	0.786 ± 0.0386
	MSL	0.833 ± 0.0020
	SMAP	0.925 ± 0.0014
	SWAT	0.944 ± 0.0005
	PSM	0.958 ± 0.0002
	N2	0.513 ± 0.0397
	N5	0.513 ± 0.0390
	R	0.494 ± 0.0625
	B	0.492 ± 0.0229
	S	0.522 ± 0.0418
R	SMD	0.721 ± 0.0565
	MSL	0.830 ± 0.0046
	SMAP	0.915 ± 0.0020
	SWAT	0.961 ± 0.0010
	PSM	0.939 ± 0.0004
	N2	0.769 ± 0.0594
	N5	0.768 ± 0.0582
	R	0.705 ± 0.0825
	B	0.737 ± 0.0340
	S	0.774 ± 0.0581
P	SMD	0.867 ± 0.0114
	MSL	0.836 ± 0.0014
	SMAP	0.936 ± 0.0009
	SWAT	0.927 ± 0.0001
	PSM	0.977 ± 0.0002
	N2	0.385 ± 0.0299
	N5	0.385 ± 0.0294
	R	0.380 ± 0.0502
	B	0.369 ± 0.0172
	S	0.394 ± 0.0325

Table 15: **GPT2 - Generalist Anomaly Detection** ($\uparrow$)

		Mean $\pm$ Std
F1	SMD	0.797 ± 0.0326
	MSL	0.802 ± 0.0205
	SMAP	0.671 ± 0.0041
	SWAT	0.896 ± 0.0016
	PSM	0.905 ± 0.0759
	N2	0.390 ± 0.0596
	N5	0.422 ± 0.0047
	R	0.361 ± 0.0204
	B	0.208 ± 0.0462
	S	0.381 ± 0.0621
R	SMD	0.734 ± 0.0559
	MSL	0.785 ± 0.0277
	SMAP	0.534 ± 0.0051
	SWAT	0.875 ± 0.0033
	PSM	0.878 ± 0.0624
	N2	0.337 ± 0.0592
	N5	0.368 ± 0.0498
	R	0.297 ± 0.0218
	B	0.177 ± 0.0426
	S	0.318 ± 0.0648
P	SMD	0.874 ± 0.0029
	MSL	0.820 ± 0.0130
	SMAP	0.900 ± 0.0007
	SWAT	0.918 ± 0.0006
	PSM	0.934 ± 0.0925
	N2	0.464 ± 0.0561
	N5	0.496 ± 0.0396
	R	0.463 ± 0.0139
	B	0.253 ± 0.0498
	S	0.477 ± 0.5000

I.3 FORECASTING RESULTS - MEANS AND STANDARD DEVIATIONS

Table 16: **TOTEM - Specialist Forecasting** ($\downarrow$)

Metric		Mean ± Std	
		MSE	MAE
W	96	0.165 ± 0.0015	0.208 ± 0.0012
	192	0.207 ± 0.0006	0.250 ± 0.0012
	336	0.257 ± 0.0002	0.291 ± 0.0006
	720	0.326 ± 0.0035	0.340 ± 0.0023
E	96	0.178 ± 0.0015	0.263 ± 0.0010
	192	0.187 ± 0.0015	0.272 ± 0.0015
	336	0.199 ± 0.0012	0.285 ± 0.0012
	720	0.236 ± 0.0035	0.318 ± 0.0031
T	96	0.523 ± 0.0010	0.303 ± 0.0006
	192	0.530 ± 0.0030	0.303 ± 0.0017
	336	0.549 ± 0.0017	0.311 ± 0.0021
	720	0.598 ± 0.0095	0.331 ± 0.0062
m1	96	0.320 ± 0.0006	0.347 ± 0.0006
	192	0.379 ± 0.0017	0.382 ± 0.0012
	336	0.406 ± 0.0040	0.402 ± 0.0026
	720	0.471 ± 0.0006	0.438 ± 0.0010
m2	96	0.176 ± 0.0006	0.253 ± 0.0010
	192	0.247 ± 0.0012	0.302 ± 0.0015
	336	0.317 ± 0.0046	0.348 ± 0.0031
	720	0.426 ± 0.0085	0.410 ± 0.0062
h1	96	0.380 ± 0.0006	0.394 ± 0.0000
	192	0.434 ± 0.0010	0.427 ± 0.0006
	336	0.490 ± 0.0023	0.459 ± 0.0015
	720	0.539 ± 0.0031	0.513 ± 0.0020
h2	96	0.293 ± 0.0015	0.338 ± 0.0006
	192	0.375 ± 0.0031	0.390 ± 0.0026
	336	0.422 ± 0.0046	0.431 ± 0.0031
	720	0.610 ± 0.0095	0.567 ± 0.0081

Table 17: **GPT2 - Specialist Forecasting, Lookback Window of 96** ($\downarrow$)

Metric		Mean ± Std	
		MSE	MAE
W	96	0.184 ± 0.0013	0.224 ± 0.0014
	192	0.231 ± 0.0012	0.263 ± 0.0009
	336	0.285 ± 0.0015	0.302 ± 0.0013
	720	0.362 ± 0.0016	0.351 ± 0.0008
E	96	0.186 ± 0.0004	0.272 ± 0.0005
	192	0.190 ± 0.0007	0.278 ± 0.0008
	336	0.204 ± 0.0003	0.291 ± 0.0005
	720	0.245 ± 0.0012	0.324 ± 0.0014
T	96	0.471 ± 0.0016	0.311 ± 0.0016
	192	0.479 ± 0.0017	0.312 ± 0.0010
	336	0.490 ± 0.0009	0.317 ± 0.0010
	720	0.524 ± 0.0019	0.336 ± 0.0018
m1	96	0.328 ± 0.0022	0.363 ± 0.0014
	192	0.368 ± 0.0006	0.382 ± 0.0004
	336	0.400 ± 0.0013	0.404 ± 0.0011
	720	0.462 ± 0.0010	0.440 ± 0.0009
m2	96	0.178 ± 0.0000	0.263 ± 0.0000
	192	0.245 ± 0.0000	0.307 ± 0.0000
	336	0.307 ± 0.0000	0.346 ± 0.0000
	720	0.410 ± 0.0000	0.409 ± 0.0000
h1	96	0.379 ± 0.0032	0.397 ± 0.0007
	192	0.438 ± 0.0037	0.427 ± 0.0004
	336	0.474 ± 0.0045	0.448 ± 0.0004
	720	0.496 ± 0.0066	0.475 ± 0.0033
h2	96	0.295 ± 0.0000	0.348 ± 0.0000
	192	0.384 ± 0.0000	0.402 ± 0.0000
	336	0.418 ± 0.0000	0.432 ± 0.0000
	720	0.423 ± 0.0000	0.446 ± 0.0000

Table 18: **TOTEM - Generalist and Zero-Shot Forecasting** (↓)

Metric		Mean ± Std MSE	MAE
W	96	0.172 ± 0.0010	0.216 ± 0.0006
	192	0.217 ± 0.0006	0.256 ± 0.0006
	336	0.266 ± 0.0015	0.295 ± 0.0015
	720	0.334 ± 0.0010	0.342 ± 0.0012
E	96	0.179 ± 0.0006	0.264 ± 0.0012
	192	0.181 ± 0.0006	0.267 ± 0.0000
	336	0.196 ± 0.0020	0.283 ± 0.0015
	720	0.230 ± 0.0035	0.314 ± 0.0029
T	96	0.507 ± 0.0020	0.284 ± 0.0006
	192	0.511 ± 0.0030	0.282 ± 0.0006
	336	0.535 ± 0.0076	0.292 ± 0.0012
	720	0.580 ± 0.0046	0.309 ± 0.0006
m1	96	0.374 ± 0.0000	0.384 ± 0.0006
	192	0.400 ± 0.0015	0.399 ± 0.0023
	336	0.432 ± 0.0040	0.424 ± 0.0015
	720	0.487 ± 0.0081	0.460 ± 0.0017
m2	96	0.198 ± 0.0006	0.275 ± 0.0012
	192	0.266 ± 0.0035	0.319 ± 0.0021
	336	0.365 ± 0.0115	0.377 ± 0.0038
	720	0.588 ± 0.0699	0.511 ± 0.0281
h1	96	0.382 ± 0.0364	0.404 ± 0.0012
	192	0.463 ± 0.0025	0.435 ± 0.0006
	336	0.507 ± 0.0025	0.463 ± 0.0010
	720	0.517 ± 0.0010	0.500 ± 0.0017
h2	96	0.307 ± 0.0012	0.345 ± 0.0015
	192	0.406 ± 0.0038	0.403 ± 0.0023
	336	0.505 ± 0.0114	0.460 ± 0.0035
	720	0.661 ± 0.0514	0.557 ± 0.0215
Zero-Shot			
N2	96	1.138 ± 0.0032	0.777 ± 0.0012
	192	1.149 ± 0.0026	0.785 ± 0.0012
	336	1.092 ± 0.0062	0.770 ± 0.0026
	720	1.045 ± 0.0040	0.754 ± 0.0023
N5	96	0.483 ± 0.0012	0.484 ± 0.0012
	192	0.495 ± 0.0021	0.491 ± 0.0015
	336	0.468 ± 0.0035	0.483 ± 0.0029
	720	0.451 ± 0.0023	0.477 ± 0.0023
R	96	1.120 ± 0.0081	0.582 ± 0.0036
	192	1.242 ± 0.0151	0.635 ± 0.0074
	336	1.237 ± 0.0153	0.626 ± 0.0076
	720	1.182 ± 0.0151	0.604 ± 0.0050
B	96	0.805 ± 0.0070	0.739 ± 0.0035
	192	0.836 ± 0.0040	0.752 ± 0.0021
	336	0.809 ± 0.0038	0.748 ± 0.0021
	720	0.896 ± 0.0137	0.794 ± 0.0085
S	96	0.446 ± 0.0032	0.482 ± 0.0017
	192	0.462 ± 0.0015	0.491 ± 0.0010
	336	0.521 ± 0.0122	0.525 ± 0.0068
	720	0.717 ± 0.0096	0.625 ± 0.0040

Table 19: **GPT2 - Generalist and Zero-Shot Forecasting** (↓)

Metric		Mean ± Std	
		MSE	MAE
W	96	0.201 ± 0.0017	0.237 ± 0.0012
	192	0.247 ± 0.0020	0.275 ± 0.0015
	336	0.298 ± 0.0006	0.311 ± 0.0006
	720	0.372 ± 0.0010	0.360 ± 0.0006
E	96	0.194 ± 0.0012	0.278 ± 0.0021
	192	0.199 ± 0.0006	0.284 ± 0.0006
	336	0.214 ± 0.0012	0.300 ± 0.0015
	720	0.255 ± 0.0006	0.331 ± 0.0012
T	96	0.484 ± 0.0046	0.320 ± 0.0042
	192	0.488 ± 0.0006	0.320 ± 0.0006
	336	0.502 ± 0.0020	0.326 ± 0.0021
	720	0.534 ± 0.0021	0.343 ± 0.0021
m1	96	0.487 ± 0.0106	0.468 ± 0.0035
	192	0.516 ± 0.0071	0.480 ± 0.0021
	336	0.548 ± 0.0015	0.499 ± 0.0015
	720	0.581 ± 0.0031	0.511 ± 0.0012
m2	96	0.243 ± 0.0021	0.315 ± 0.0021
	192	0.297 ± 0.0012	0.346 ± 0.0010
	336	0.349 ± 0.0025	0.376 ± 0.0020
	720	0.439 ± 0.0010	0.423 ± 0.0010
h1	96	0.421 ± 0.0058	0.408 ± 0.0010
	192	0.480 ± 0.0026	0.436 ± 0.0020
	336	0.518 ± 0.0161	0.453 ± 0.0070
	720	0.517 ± 0.0036	0.467 ± 0.0035
h2	96	0.298 ± 0.0090	0.343 ± 0.0049
	192	0.381 ± 0.0153	0.392 ± 0.0072
	336	0.406 ± 0.0271	0.419 ± 0.0144
	720	0.423 ± 0.0078	0.438 ± 0.0051
Zero-Shot			
N2	96	1.332 ± 0.0012	0.830 ± 0.0010
	192	1.416 ± 0.0080	0.863 ± 0.0025
	336	1.358 ± 0.0123	0.851 ± 0.0042
	720	1.308 ± 0.0026	0.840 ± 0.0010
N5	96	0.528 ± 0.0006	0.499 ± 0.0010
	192	0.578 ± 0.0015	0.524 ± 0.0006
	336	0.548 ± 0.0040	0.515 ± 0.0015
	720	0.537 ± 0.0006	0.511 ± 0.0006
R	96	1.465 ± 0.0185	0.725 ± 0.0031
	192	1.638 ± 0.0280	0.785 ± 0.0078
	336	1.601 ± 0.0244	0.769 ± 0.0060
	720	1.552 ± 0.0110	0.760 ± 0.0035
B	96	0.838 ± 0.0149	0.762 ± 0.0071
	192	0.837 ± 0.0095	0.752 ± 0.0040
	336	0.792 ± 0.0104	0.738 ± 0.0050
	720	0.927 ± 0.0066	0.806 ± 0.0038
S	96	0.443 ± 0.0010	0.478 ± 0.0006
	192	0.481 ± 0.0006	0.499 ± 0.0006
	336	0.541 ± 0.0010	0.533 ± 0.0006
	720	0.773 ± 0.0020	0.643 ± 0.0010

I.4 ADDITIONAL ABLATIONS

Table 20: **TimeTOTEM Ablation - Specialist Forecasting**

		Mean ± Std	
		MSE	MAE
W	96	0.164 ± 0.0006	0.209 ± 0.0006
	192	0.209 ± 0.0017	0.251 ± 0.0023
	336	0.261 ± 0.0012	0.293 ± 0.0017
	720	0.332 ± 0.0023	0.340 ± 0.0006
E	96	0.179 ± 0.0015	0.262 ± 0.0015
	192	0.185 ± 0.0006	0.269 ± 0.0000
	336	0.204 ± 0.0055	0.289 ± 0.0061
	720	0.244 ± 0.0040	0.325 ± 0.0036
T	96	0.528 ± 0.0081	0.310 ± 0.0092
	192	0.500 ± 0.0606	0.349 ± 0.0699
	336	0.531 ± 0.0424	0.365 ± 0.0852
	720	0.578 ± 0.0361	0.398 ± 0.1103
m1	96	0.326 ± 0.0006	0.355 ± 0.0006
	192	0.377 ± 0.0023	0.386 ± 0.0012
	336	0.409 ± 0.0006	0.409 ± 0.0006
	720	0.469 ± 0.0015	0.441 ± 0.0000
m2	96	0.176 ± 0.0010	0.254 ± 0.0006
	192	0.247 ± 0.0031	0.303 ± 0.0026
	336	0.318 ± 0.0006	0.350 ± 0.0021
	720	0.419 ± 0.0067	0.411 ± 0.0044
h1	96	0.377 ± 0.0010	0.395 ± 0.0006
	192	0.428 ± 0.0015	0.428 ± 0.0015
	336	0.480 ± 0.0021	0.462 ± 0.0012
	720	0.530 ± 0.0110	0.522 ± 0.0108
h2	96	0.294 ± 0.0021	0.338 ± 0.0010
	192	0.373 ± 0.0023	0.389 ± 0.0032
	336	0.423 ± 0.0031	0.433 ± 0.0025
	720	0.591 ± 0.0145	0.556 ± 0.0051

Table 21: **TimeTOTEM Ablation - Generalist and Zero-Shot Forecasting**

Metric		Mean ± Std	
		MSE	MAE
W	96	0.173 ± 0.0012	0.218 ± 0.0006
	192	0.218 ± 0.0006	0.261 ± 0.0006
	336	0.267 ± 0.0006	0.299 ± 0.0006
	720	0.337 ± 0.0010	0.347 ± 0.0006
E	96	0.183 ± 0.0012	0.267 ± 0.0012
	192	0.189 ± 0.0006	0.275 ± 0.0000
	336	0.204 ± 0.0010	0.291 ± 0.0010
	720	0.242 ± 0.0006	0.325 ± 0.0006
T	96	0.517 ± 0.0000	0.293 ± 0.0029
	192	0.526 ± 0.0030	0.296 ± 0.0006
	336	0.552 ± 0.0015	0.304 ± 0.0015
	720	0.602 ± 0.0046	0.326 ± 0.0015
m1	96	0.428 ± 0.0090	0.420 ± 0.0040
	192	0.438 ± 0.0015	0.427 ± 0.0010
	336	0.469 ± 0.0062	0.447 ± 0.0042
	720	0.546 ± 0.0081	0.493 ± 0.0017
m2	96	0.207 ± 0.0015	0.286 ± 0.0020
	192	0.269 ± 0.0015	0.325 ± 0.0010
	336	0.358 ± 0.0199	0.377 ± 0.0091
	720	0.521 ± 0.0165	0.482 ± 0.0026
h1	96	0.401 ± 0.0006	0.410 ± 0.0006
	192	0.453 ± 0.0010	0.441 ± 0.0010
	336	0.496 ± 0.0017	0.468 ± 0.0006
	720	0.518 ± 0.0020	0.510 ± 0.0017
h2	96	0.305 ± 0.0006	0.346 ± 0.0006
	192	0.396 ± 0.0015	0.402 ± 0.0001
	336	0.492 ± 0.0310	0.458 ± 0.0131
	720	0.599 ± 0.0105	0.531 ± 0.0026
N2	96	1.127 ± 0.0017	0.773 ± 0.0006
	192	1.169 ± 0.0032	0.793 ± 0.0010
	336	1.115 ± 0.0010	0.780 ± 0.0006
	720	1.070 ± 0.0035	0.766 ± 0.0010
N5	96	0.481 ± 0.0015	0.483 ± 0.0006
	192	0.508 ± 0.0012	0.500 ± 0.0000
	336	0.481 ± 0.0006	0.491 ± 0.0006
	720	0.467 ± 0.0010	0.488 ± 0.0010
R	96	1.102 ± 0.0031	0.578 ± 0.0021
	192	1.207 ± 0.0036	0.628 ± 0.0017
	336	1.190 ± 0.0021	0.613 ± 0.0010
	720	1.149 ± 0.0017	0.596 ± 0.0020
B	96	0.825 ± 0.0079	0.751 ± 0.0076
	192	0.847 ± 0.0021	0.761 ± 0.0012
	336	0.831 ± 0.0066	0.764 ± 0.0042
	720	0.928 ± 0.0131	0.813 ± 0.0050
S	96	0.446 ± 0.0015	0.481 ± 0.0010
	192	0.478 ± 0.0015	0.499 ± 0.0000
	336	0.535 ± 0.0012	0.532 ± 0.0006
	720	0.736 ± 0.0025	0.631 ± 0.0006

Table 22: **Detailed Codebook Ablation** (↓)

	K	Mean ± Std MSE	Mean ± Std MAE
W	256	0.016 ± 0.0004	0.067 ± 0.0011
	512	0.013 ± 0.0011	0.061 ± 0.0032
	32	0.039 ± 0.0005	0.112 ± 0.0064
E	256	0.021 ± 0.0012	0.103 ± 0.0029
	512	0.015 ± 0.0005	0.088 ± 0.0014
	32	0.046 ± 0.0007	0.152 ± 0.0016
T	256	0.012 ± 0.0003	0.075 ± 0.0007
	512	0.010 ± 0.0012	0.069 ± 0.0044
	32	0.031 ± 0.0007	0.120 ± 0.0008
All	256	0.019 ± 0.0003	0.094 ± 0.0007
	512	0.018 ± 0.0025	0.091 ± 0.0062
	32	0.045 ± 0.0014	0.146 ± 0.0030

I.5 EXPLORATORY RESULTS

Table 23: **Mixed Models - Forecasting** (↓)

Metric		Mean ± Std	
		MSE	MAE
W	96	0.164 ± 0.0010	0.208 ± 0.0012
	192	0.208 ± 0.0010	0.251 ± 0.0015
	336	0.258 ± 0.0012	0.290 ± 0.0015
	720	0.329 ± 0.0021	0.338 ± 0.0015
E	96	0.178 ± 0.0006	0.263 ± 0.0010
	192	0.187 ± 0.0021	0.273 ± 0.0017
	336	0.199 ± 0.0012	0.285 ± 0.0017
	720	0.238 ± 0.0012	0.320 ± 0.0012
T	96	0.521 ± 0.0010	0.301 ± 0.0010
	192	0.530 ± 0.0023	0.303 ± 0.0012
	336	0.555 ± 0.0080	0.313 ± 0.0072
	720	0.605 ± 0.0097	0.337 ± 0.0075
m1	96	0.328 ± 0.0036	0.352 ± 0.0006
	192	0.377 ± 0.0021	0.383 ± 0.0012
	336	0.408 ± 0.0035	0.404 ± 0.0021
	720	0.470 ± 0.0035	0.440 ± 0.0021
m2	96	0.175 ± 0.0006	0.253 ± 0.0010
	192	0.247 ± 0.0006	0.302 ± 0.0010
	336	0.318 ± 0.0006	0.348 ± 0.0031
	720	0.427 ± 0.0012	0.410 ± 0.0067
h1	96	0.382 ± 0.0025	0.395 ± 0.0015
	192	0.437 ± 0.0012	0.427 ± 0.0006
	336	0.490 ± 0.0015	0.460 ± 0.0021
	720	0.536 ± 0.0031	0.512 ± 0.0032
h2	96	0.294 ± 0.0010	0.339 ± 0.0012
	192	0.375 ± 0.0025	0.391 ± 0.0023
	336	0.421 ± 0.0050	0.431 ± 0.0031
	720	0.610 ± 0.0089	0.567 ± 0.0075

Table 24: **Traffic Only - Specialist Zero-Shot Performance** (↓)

Metric		Mean ± Std	
		MSE	MAE
N2	96	1.194 ± 0.0062	0.798 ± 0.0020
	192	1.218 ± 0.0074	0.808 ± 0.0023
	336	1.190 ± 0.0153	0.804 ± 0.0052
	720	1.117 ± 0.0137	0.784 ± 0.0056
N5	96	0.515 ± 0.0026	0.505 ± 0.0012
	192	0.535 ± 0.0051	0.514 ± 0.0028
	336	0.524 ± 0.0071	0.513 ± 0.0030
	720	0.500 ± 0.0064	0.507 ± 0.0032
R	96	1.171 ± 0.0023	0.635 ± 0.0019
	192	1.273 ± 0.0090	0.673 ± 0.0042
	336	1.232 ± 0.0055	0.653 ± 0.0022
	720	1.198 ± 0.0057	0.642 ± 0.0041
B	96	0.812 ± 0.0037	0.749 ± 0.0025
	192	0.858 ± 0.0025	0.767 ± 0.0015
	336	0.826 ± 0.0041	0.759 ± 0.0030
	720	0.919 ± 0.0063	0.803 ± 0.0037
S	96	0.476 ± 0.0012	0.508 ± 0.0012
	192	0.511 ± 0.0005	0.528 ± 0.0005
	336	0.576 ± 0.0024	0.568 ± 0.0009
	720	0.795 ± 0.0017	0.685 ± 0.0012

Table 25: **Electricity Only - Specialist Zero-Shot Performance** (↓)

Metric		Mean ± Std	
		MSE	MAE
N2	96	1.193 ± 0.0059	0.802 ± 0.0020
	192	1.300 ± 0.0016	0.845 ± 0.0003
	336	1.260 ± 0.0162	0.837 ± 0.0055
	720	1.234 ± 0.0054	0.832 ± 0.0016
N5	96	0.489 ± 0.0024	0.490 ± 0.0011
	192	0.555 ± 0.0012	0.527 ± 0.0007
	336	0.538 ± 0.0064	0.525 ± 0.0033
	720	0.533 ± 0.0010	0.527 ± 0.0006
R	96	1.141 ± 0.0056	0.579 ± 0.0028
	192	1.297 ± 0.0162	0.652 ± 0.0079
	336	1.247 ± 0.0108	0.628 ± 0.0059
	720	1.236 ± 0.0053	0.633 ± 0.0070
B	96	0.820 ± 0.0065	0.756 ± 0.0034
	192	0.843 ± 0.0042	0.759 ± 0.0022
	336	0.791 ± 0.0023	0.741 ± 0.0019
	720	0.886 ± 0.0059	0.790 ± 0.0020
S	96	0.460 ± 0.0017	0.487 ± 0.0010
	192	0.505 ± 0.0017	0.511 ± 0.0008
	336	0.569 ± 0.0020	0.545 ± 0.0011
	720	0.764 ± 0.0046	0.641 ± 0.0014