STRUCTURAL MODES OF CONTEXT-TO-CONTEXT ATTENTION IN CHRONOS: A MECHANISTIC AUTOPSY OF A TIME SERIES FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models for time series forecasting are trained on vast, high-dimensional real-world data and synthetic corpora, yet they can fail on zero-shot forecasting of sequences generated by simple, deterministic functions. In this work, we investigate the internal mechanism behind this shortcoming of the original Chronos, a time series foundation model whose pointwise tokenization makes its attention directly interpretable. Following techniques in explainable AI, we probe Chronos with the controlled stimuli of Jacobi elliptic functions, autoregressive iterative series, and the logistic map as inputs, where we vary the parameters that modulate periodicity, locality, and chaoticity. We extract the average attention matrices of each encoder block. These matrices resolve into a small set of canonical structural modes. Early blocks are dominated by value-based modes, in which tokens with equal values attend to each other and all tokens attend to extrema, while later blocks concentrate attention on local neighborhoods and off-diagonal bands. We quantify these modes with locality and periodicity metrics defined on the attention matrices, and we relate the behavior of the encoder to forecast accuracy through rank correlations and layer pruning. For several stimulus families, the forecast accuracy improves when the encoder resolves the input into at least one interpretable structural mode of strong locality or clear periodicity, and degrades when it resolves neither, as for near-aperiodic or chaotic signals. We also address how our analysis can be extended to other time series foundation models to diagnose how models encode using other architectural strategies such as patch embedding and causal masking. This suggests a promising direction for interpreting how deep learning architectures structure data processing.

1 INTRODUCTION

Since the conception of the transformer (Vaswani et al., 2017) and large language models (Devlin et al., 2019; Radford et al., 2018), foundation models have been developed in many disciplines. Trained with billions of data (real and synthetic) and equipped with millions to trillions of parameters, foundation models are constructed with the aim of being state-of-the-art for zero-shot prediction or fine-tuned for downstream tasks (Bommasani et al., 2021; Awais et al., 2025; Liang et al., 2024; Zhou et al., 2025). In recent years, this paradigm has found its way to time series forecasting through models such as Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), and TimesFM (Das et al., 2024), which reportedly surpass classic models such as ARIMA and other neural network models (Salinas et al., 2020; Zhou et al., 2021; Wu et al., 2021; Lim et al., 2021; Nie et al., 2022) trained for time series forecasting.

Despite their zero-shot capabilities, these foundation models have been shown to struggle on certain types of synthetic data and shifted-domain data, where simpler methods are competitive (Toner et al., 2025; Zhang et al., 2025). In particular, Gupta (2025) shows that foundation models perform well on bounded periodic functions but their performance deteriorates with noise. Figure 1a demonstrates one example where foundation models fail on the Jacobi elliptic function (Lawden, 2013), a simple, deterministic function $cn(t; k)$, which is a cosine function when $k = 0$ and smoothly transforms into a hyperbolic secant as $k \to 1$. While foundation models (Chronos, Moirai, TimesFM) can produce quality forecasts when $k = 0$, the forecast error noticeably rises as $k \to 1$. Seemingly, models

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

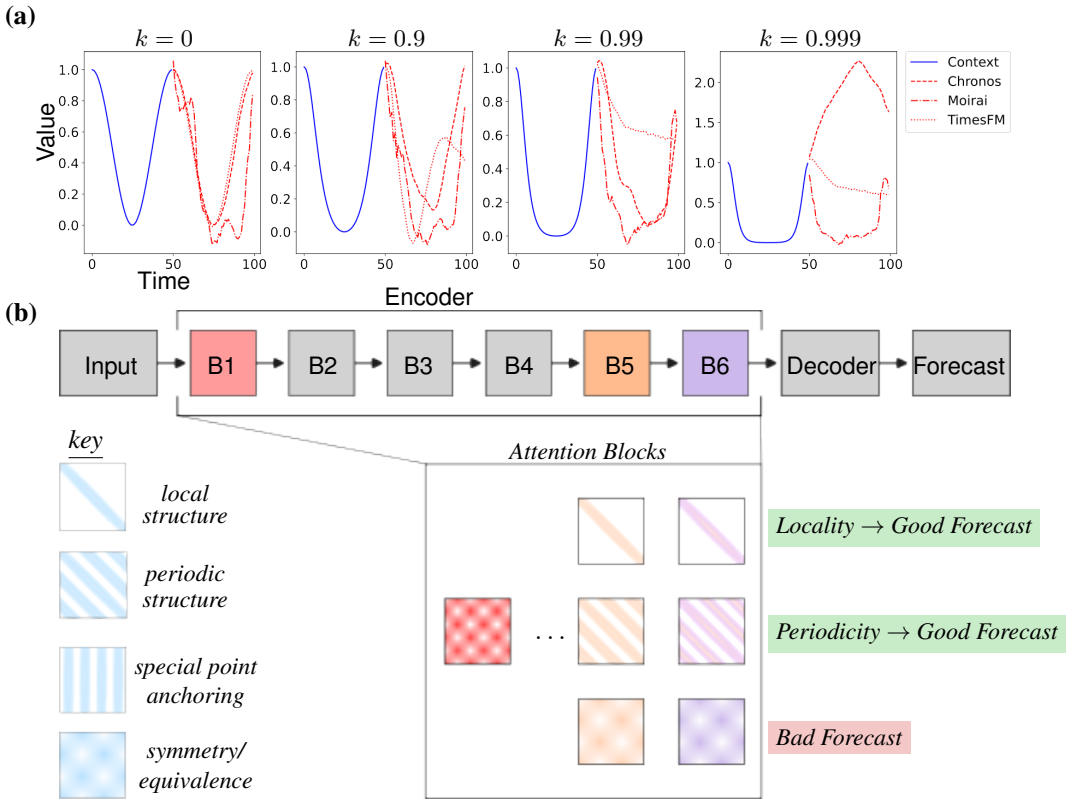

Figure 1: **(a)** Foundation models (Chronos, Moirai, TimesFM) forecast on Jacobi elliptic functions $cn(t; k)$. Forecasts are accurate for $k = 0$, where cn is a cosine, and degrade as $k \to 1$. **(b)** Summary of our findings for the Chronos encoder. Its attention matrices resolve into canonical structural modes (key): a band along the main diagonal (**locality**: neighboring points attend to each other, as also observed in BERT (Kovaleva et al., 2019)); bands parallel to the diagonal (**periodicity**: points one period apart attend to each other); vertical bands (**special-point anchoring**: all points attend to a few anchors, such as extrema); and anti-diagonal or cross-shaped activity (**symmetry/equivalence**: points with equal values attend to each other). Early blocks are dominated by value-based modes, and later blocks by locality and periodicity. Forecast accuracy tends to be higher when the encoder resolves the input into strong locality or clear periodicity.

trained on large, diverse real-world and synthetic corpora fail on simple one-parameter variation time sequences. One explanation for these failures is the sensitivity of zero-shot performance to the statistical alignment of the training and target domains (Karaouli et al., 2025). This raises the question of what the models are learning, and what internal mechanism drives the successes and failures. In response, we systematically investigate how a foundation model processes time series data and the division of labor in the attention layers.

The black box nature of large neural network models has driven a large body of interpretability research on transformers. This line of work identifies the features the model captures from its input (Clark et al., 2019; Belinkov, 2022; Vig & Belinkov, 2019; Bricken et al., 2023; Cunningham et al., 2023) and the components responsible for its output (Elhage et al., 2021; Olsson et al., 2022; Serrano & Smith, 2019) (see Rai et al. (2024) for a detailed taxonomy). Ablation has especially been revealing about the division of labor across depth. For large language models, layer pruning reveals that they use a two-stage decision process in which pruning early layers causes a performance collapse but deeper layers are largely redundant (Shi et al., 2026; Lu et al., 2026). Bao et al. (2026) establish a similar redundancy for time series foundation models. Other works visualize values of the attention matrices themselves and observe clustering effects. Attention weights are analyzed with respect to their task-specific usefulness and interpretability (Raghu et al., 2021; Zhang et al., 2021). Emerging patterns can be understood as learned similarity between tokens, and canonical structures

have been identified (Tsai et al., 2019; Kovaleva et al., 2019). Recent theoretical work reduces self-attention to an interacting particle system and analyzes clustering dynamics at the mean-field limit (Geshkovski et al., 2023; 2025).

We apply pruning and visualization of patterns on the attention matrices to study time series foundation models. To our knowledge, time series foundation models have not been investigated as rigorously as large language models. Kalnāre et al. (2025) transfer language-model techniques (attention saliency, sparse autoencoder) to time series classifiers, but not to foundation models. Mishra (2026) applies sparse autoencoders to Chronos and correlates layer depth with semantic features such as frequency and seasonality but does not connect their results to forecast accuracy.

In this work, we quantify the behavior of Chronos's internal mechanisms and connect them to forecast accuracy. We focus on Chronos (Ansari et al., 2024) because its T5 encoder-decoder architecture offers tractable layer counts and sizes, and because its pointwise tokenization (one token per time step) makes the latent representation directly interpretable. In particular, throughout this work, we show results with the small sized model of Chronos. Results for a larger sized model are available in Appendix C. We generate synthetic time series data from parametric functions (Jacobi elliptic functions, autoregressive iterative series, and the logistic map) where we vary a single parameter to systematically modulate properties such as periodicity, decay, and chaoticity. By recording and quantifying internal activity (attention weight matrices) at each encoder block in response to these stimuli, and correlating those responses with behavioral output (forecast accuracy), we construct a direct mechanistic account of what drives success and failure. Although this work focuses on Chronos, the methodology can be applied to other foundation model architectures. In Appendices D and E, we apply it to Moirai-1, which uses patch embedding, and to Moirai-2, a decoder-only model with causal attention.

Visualizing attention on time series reveals a small set of canonical structural modes in Chronos's encoder (figure 1). We define locality and periodicity metrics on the attention weight matrices to discern these modes. We also perform block lesion studies (layer pruning) and map the functional role of each encoder block. An interactive HTML viewer of our results is available in our GitHub repository[1]. Our findings, summarized in figure 1, are as follows:

- Chronos's encoder resolves time series into a small set of canonical structural modes that change with depth. Early blocks are dominated by value-based modes: tokens with equal values attend to each other, and all tokens attend to special points such as peaks and troughs. Later blocks concentrate attention on local neighborhoods and, for periodic inputs, on points one period apart. We quantify these modes with locality and periodicity metrics defined directly on the attention matrices.
- The encoder recognizes the periodic structure in the input. The detected period aligns with the true period of the input signal, and the strength of this periodicity signal increases in later blocks, peaking around block 5.
- For several stimulus families, forecast accuracy correlates with whether the encoder successfully resolves the input into strong locality (local dependencies the decoder can exploit) or clear periodicity (a repeating pattern the encoder maintains through its layers). When neither mode is achieved (near-aperiodic elliptic functions or chaotic dynamics), the accuracy of the forecast degrades. The relationship does not hold for the iterative series, where changes in amplitude confound the forecast.
- Layer pruning shows that adding block 5 yields the largest gain in forecast accuracy, while the final block can hurt performance on difficult signals. The collapse caused by removing the first block is explained by the loss of relative position bias, which Chronos, following the T5 architecture, computes only in the first block and shares with all later blocks. Layer-pruning results for such architectures therefore combine the function of the first block with the positional information it carries.
- The analysis extends to Moirai-1 and Moirai-2. In both, locality persists but periodic structure is blurred and only weakly related to forecast accuracy, suggesting that pointwise tokenization is what allows Chronos to represent periodicity.

[1]Our repository is currently anonymous at `https://anonymous.4open.science/r/foundation-model-forecasting-9133`

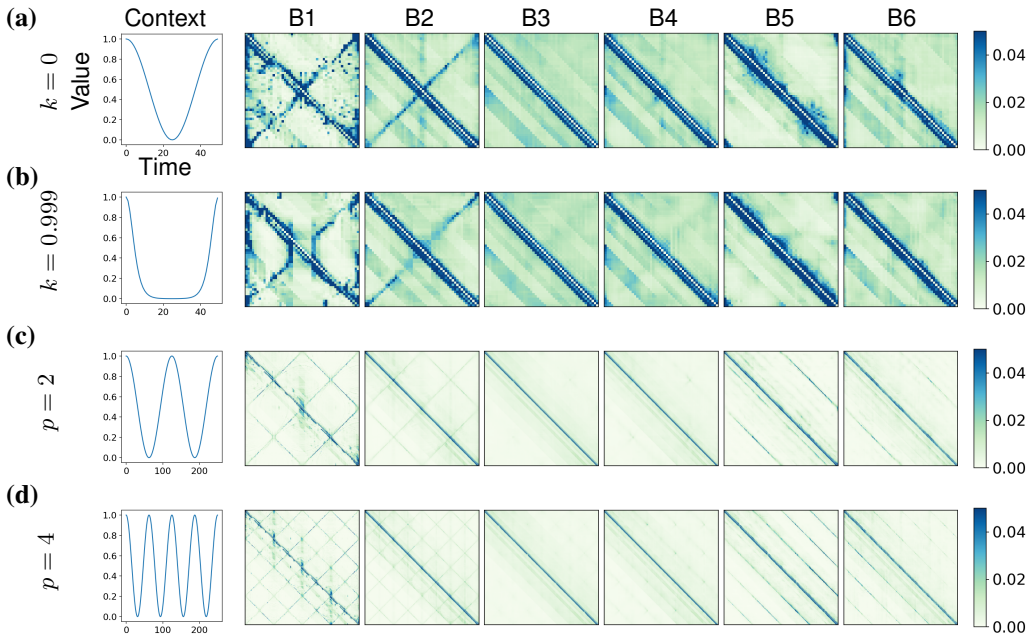

Figure 2: This figure shows the full average attention matrices of sequences generated from the Jacobi elliptic function where **(a)** $k = 0$ and **(b)** $k = 0.999$ and from a cosine function with **(c)** two cycles and **(d)** four cycles. Attention is computed from *softmax*$(QK^T + b)$, averaged across attention heads. Early blocks show value-based structure: equal-value attention, which produces symmetric and cross-shaped patterns, and vertical bands at extrema. Later blocks attend to local neighborhoods, and for periodic inputs, bands parallel to the diagonal form in the later blocks.

2 Attention Analysis

The small sized model of Chronos has six attention blocks in its encoder. Each attention block consists of a multihead attention layer with eight heads, followed by a shallow 2-layer feedforward network. Details on the architecture can be found in Ansari et al. (2024). We extract the query and key matrices, denoted $Q^{(h)}, K^{(h)}$ respectively, from each attention head h. The attention weight matrix for the head is computed as $X^{(h)} = \text{softmax}\left(Q^{(h)}{K^{(h)}}^T + b^{(h)}\right)$ where b is the relative position bias determined only by the length of the sequence. The bias is computed once in the first encoder block and passed to every later block. An example of these attention matrices for each head and each block is shown in figure 5 for a standard cosine function. In Chronos, each $X^{(h)}$ is then applied to the value matrix before rejoining into one output. For simplicity, we compute an average attention weight matrix by averaging across H attention heads in a layer, i.e. $X = \frac{1}{H}\sum_{h=1}^{H} X^{(h)}$.

Chronos uses pointwise embedding to tokenize the input sequence, which enables direct interpretation of the attention matrices. We take the input sequence (context) to be generated by a function f at time steps $t_1, \ldots, t_n$. As a result, component X_{ij} corresponds precisely to the attention between the tokenized values of $f(t_i)$ and $f(t_j)$. Two features of this setup enable our analysis. Firstly, pointwise tokenization maps inputs with equal values (up to quantization) to the same token. At the first block, whose only inputs are the token embeddings and the position bias, such tokens have identical queries and keys and therefore identical rows and columns in $Q^{(h)}{K^{(h)}}^T$. Symmetric, value-based structure in the early blocks is therefore partly a consequence of the tokenization. What the model learns is how strongly equal-valued tokens attend to each other. Secondly, because b is identical for every input, attention structure that stays the same across very different inputs can be attributed to position, whereas structure that changes with the input reflects its content.

Across the average attention matrices, we observe a small set of recurring structural modes. In particular, token i attends to token j if $f(t_i) = f(t_j)$ (**equal-value attention**), $|i - j| \leq \epsilon$ (**locality**),

$f^{(d)}(t_i) = f^{(d)}(t_j)$ (**higher-order similarity**), or $i-j \in [a, b]$ (**off-diagonal attention**, i.e. i attends to j at a fixed range of distances from j). There is also **special-point anchoring** where every token i attends to a small set of anchor tokens $\{j_1, j_2 \ldots, j_k\}$, independently of i. Locality patterns were previously observed in attention matrices of the language model BERT (Kovaleva et al., 2019). An attention matrix exhibits a local pattern when attention weights are higher between tokens that are closer to each other; this appears as a strong main diagonal. Meanwhile, activity away from the diagonal links tokens that are far apart in time.

For the Jacobi elliptic functions (Figure 2a-b), blocks 1 and 2 link tokens with equal values, while later blocks attend to local neighborhoods. In addition, the attention matrices also preserve periodicity when a time series contains several cycles. On cosine functions with varying number of cycles, figures 2c-d reveal bands parallel to the diagonal, spaced by the period, at blocks 5 and 6; this is the signature of attention at higher-order similarity. We also notice vertical bands in blocks 1 and 2 where peaks and troughs are, suggesting that these early blocks are attending to special values such as the minima and the maxima. We find these patterns in other time series, which we show in figures 8, 9, and 10 in the supplementary material (section B.2). We also note that there is a consistent off-diagonal band for each time series, around time step 20, which is likely an artifact of the model and the learned position bias.

3 Locality and Periodicity Metrics

In this section, we define locality and periodicity strength metrics to quantify our observations in the attention matrices. Let $X \in \mathbb{R}^{n \times n}$ be an attention matrix with entries X_{ij}. We adapt locality metrics from Raghu et al. (2021) and Zhang et al. (2021). To our knowledge, no prior work defines a periodicity metric directly on attention weight matrices.

3.1 Expectation over Distance

We define the distance-to-diagonal matrix with entries $D_{ij} = |i - j|$. Since each row of X is a probability distribution, $\sum_{j=1}^{n} X_{ij} D_{ij}$ is a weighted average of the distance to the diagonal in row i. We compute this expectation value for each row and take the average:

$$\text{EOD}(X) = \frac{1}{n} \sum_{i,j=1}^{n} X_{ij} D_{ij} = \frac{1}{n} \sum_{i,j=1}^{n} X_{ij} |i - j| \tag{1}$$

which we call expectation over distance (EOD). We expect that if there is more attention on local neighborhoods, EOD should be small. We can compare against the uniform case where each entry of X is $1/n$. Under the uniform case, every value in the time series sequence is equally important to every other one, and the further EOD falls below the uniform baseline, the stronger the locality. EOD for the uniform case is computed in Appendix A as $(n^2 - 1)/3n$.

Expectation over distance is related to the mean distance in Raghu et al. (2021), which computes a spatially weighted average of attention in vision transformers and averages over a dataset. EOD adapts the same distance-weighted intuition to the strictly sequential structure of time series, where locality is one-dimensional and input-specific.

3.2 Main Diagonal Strength

We also quantify locality by focusing on the attention along the main diagonal. Defining the diagonal band, $B_{\text{diag}} = \{(i, j) : D_{ij} \leq w\}, w \leq n$, we compute the fraction of total mass within a width-w band around the diagonal:

$$\text{MDS}(X) = \frac{\sum_{D_{ij} \leq w} X_{ij}}{\sum_{i,j} X_{ij}} \tag{2}$$

For this metric, which we call the main diagonal strength (MDS), we expect greater diagonal strength when there is greater locality. For the uniform attention matrix, MDS is $(n + 2nw - w(w + 1))/n^2$ (Appendix A). We have stronger locality when diagonal strength is higher than the uniform baseline.

Main diagonal strength is independently derived from the diagonality metric of Zhang et al. (2021), who define it with distance weighted average while we consider a fraction of attention mass within

a fixed bandwidth of the diagonal. The fixed bandwidth introduces a hard threshold of what we quantify as local. In this sense, MDS measures how much attention is local in the absolute sense. Throughout our work, we fix $w = 0.2n$. We display the uniform attention baseline alongside results and interpret locality relative to the uninformative case.

3.3 Periodicity Strength

We develop a metric capturing the repeated patterns on the off-diagonal bands of the attention matrix. For each diagonal offset d, we compute the mean of the d and $-d$ offset diagonals to obtain a sequence of values we call the band strength profile β. It is expected that we would see repeating peaks with the same lag as the periodicity of the time series. We then compute autocorrelation

$$R_\tau = \sum_{i=0}^{\lfloor n/2 \rfloor - \tau - 1} \beta_i \beta_{i+\tau}, \quad \tau = 0, \dots \lfloor n/2 \rfloor \tag{3}$$

We normalize R as $R_{norm} = R/R_0$, which measures how well the profile matches itself after shifting by τ. The periodicity strength is then computed by selecting the most topographic prominent peak of R_{norm}, $R_{norm}[\tau_{peak}]$ scaled by $\beta[\tau_{peak}]$, i.e. periodicity strength measures frequency scaled by the attention on the diagonal band. We use a prominence threshold of 0.15, so local maxima with prominence below 0.15 are discarded as noise. Our methodology ensures that if the matrix is uniform, periodicity strength is 0.

4 Results

4.1 Locality and Periodicity Profile

Locality profiles for various time series are shown in figure 3. We note that locality generally peaks at later blocks, in particular block 5 for Chronos-small, with differences across time series and parameters. The locality profile for Jacobi elliptic functions (figure 3a) begins to reveal the features observed by the encoder. As $k \to 1$, locality decreases, perhaps suggesting that failing to achieve "strong" locality is detrimental to the forecast. Time series with higher frequency patterns also have lower locality (figure 3b).

In addition to the Jacobi elliptic function, we consider an iterative series defined by

$$x_{n+1} = \sum_{i=0}^{l} c_i x_{n-i}, \quad c_i = d^{i+1}, \quad d \in (0, 1) \tag{4}$$

where l is a lag parameter, and d is a decay parameter used to keep the time series from growing to infinity[2]. As the lag increases, we induce time series with lower locality and periodicity (figures 3c, 8). A large d renders sequences with less locality as the decay factor causes the process to vanish and creates attention matrices with little activity (figures 3d, 9). Meanwhile, a very small decay factor leaves the time series oscillating with a high frequency, rendering low locality and high periodicity.

Lastly, we test on the logistic map, $x_{n+1} = r x_n (1 - x_n)$. The attention matrices in figure 10 show clean periodic structures in the period-doubling regime ($3 \leq r < 3.57$) but not in the chaotic regime ($r \geq 3.57$). The locality profile with the position bias included (Figure 3e), however, shows similar locality for all growth rates r. This similarity is likely an artifact of the position bias. When position bias is removed, the chaotic regime shows higher locality than the period-doubling regime, suggesting that the encoder extracts more local information when the series is chaotic.

The periodicity profiles in figure 3 show that the periodicity strength of the attention tracks the frequency content of the input. For example, a simple cosine function with higher frequency is shown to have greater periodicity strength, and the chaotic regime of a logistic map is aperiodic. Notably, most time series see an increase in periodicity strength at block 5, though periodicity can emerge as early as block 1 due to value equivalence.

[2]We restrict the time series from growing as Chronos is better suited for forecasts within the range of the context given.

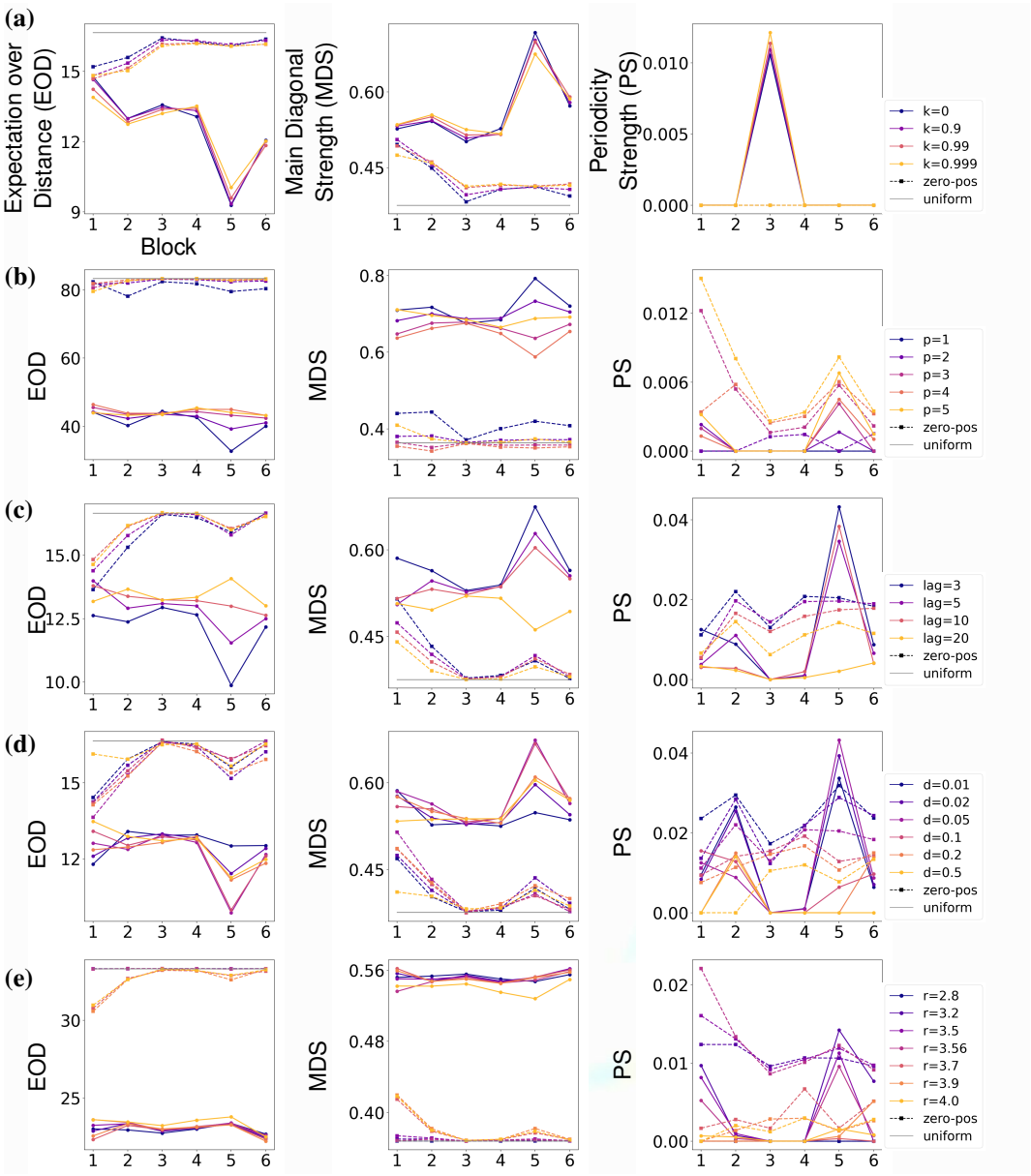

Figure 3: Locality and periodicity profiles for **(a)** Jacobi elliptic functions (for various modulus k), **(b)** cosine functions with different number of cycles (p), **(c)-(d)** the iterative series $x_{n+1} = \sum_{i=0}^{l} c_i x_{n-i}$ 4, and **(e)** logistic map $x_{n+1} = r x_n(1 - x_n)$ for varied r. In **(c)**, we vary values of l with fixed $d = 0.05$, and in **(d)**, we vary d and fix $l = 3$. Expectation over distance (EOD) and main diagonal strength (MDS) are computed across encoder blocks with and without position bias (solid vs dashed line respectively). Locality generally peaks at block 5 except for the logistic map.

4.2 Effects of Position Encoding on Locality

In figure 3, we also show the locality profiles when position encoding is removed and consider the effects of position encoding on locality. Chronos uses a relative position bias computed in the first attention block, passed to all later blocks, and dependent only on the length of the context. To see the impact of position encoding, we ablate the position bias imposed in the first attention block. The resulting attention matrices can be found in Appendix B.2. Compared against the attention matrices with position encoding, these matrices retain patterns from the early blocks.

Since position encoding provides information on what local neighborhoods mean for the model, we expected that the computed locality without position encoding would fall towards the uniform line. While that is generally true, not all blocks present uniform locality, as block 1 generally has the greatest locality, and locality at later blocks tends to peak around block 5. These traits are especially noticeable for the iterative series and the chaotic regime of the logistic map. This result suggests that despite the lack of position encoding, these blocks are able to discover some local information, which is supported by other work showing that models with no position encoding can still learn position information (Haviv et al., 2022). Furthermore, this suggests that block 5 of our Chronos model drives the increase in locality in the later blocks.

4.3 Layer Pruning

The locality and periodicity profiles indicate block 5's significance. In this section, we perform layer pruning, that is, we remove attention blocks in the encoder and observe the changes to the mean absolute error. The results in Appendix B.3 and figure 4a,b suggest that block 1 and block 5 of Chronos-small are the most important for the quality of the forecast. Removing block 1 creates a constant-valued forecast. Further experiments reveal that this forecast error is largely due to the loss of the position encoding bias (Figure 28). The effect of removing block 1 therefore reflects the role of the shared position bias rather than the function of block 1 alone. This caveat applies to layer-pruning studies of any architecture in which a component shared across layers resides in a single layer. On the other hand, when we incrementally add blocks, adding up to block 5 substantially reduces the forecast's mean absolute error. The large gain from an intermediate block and the small or negative contribution of the final block are consistent with layer-pruning studies of large language models and time series foundation models, which find the deepest layers to be largely redundant (Bao et al., 2026; Lu et al., 2026; Shi et al., 2026).

Given the importance of block 5, along with its high activity in locality and periodicity, we focus on block 5 in Figure 4c, which shows how locality and periodicity computed at this block relate to the quality of the forecast made by the model truncated at the same block. Table 1 shows the Spearman rank correlations between the metrics and two error measures: the mean absolute error (MAE) and skill (MAE divided by the MAE of a naive forecast). Correlation across all encoding blocks can be found in table 5. Strong correlations are not specific to block 5. At block 2, for example, MAE is also strongly negatively correlated with MDS for four of the five families.

For some families, lower forecast error is significantly associated with higher locality or periodicity. Consider, for example, the Jacobi elliptics (varying k) and the cosine functions with increasing frequencies; decreasing error is highly correlated with increasing locality and increasing periodicity respectively. For the logistic map, there does not appear to be a definitive alignment to only locality or periodicity. In the period-doubling regime, periodicity aligns with forecast accuracy. Meanwhile, in the chaotic regime, periodicity is low and locality can be high. One interpretation is that the model relies on locality when periodicity is not clear.

The iterative series do not follow this trend. For both the lag and the decay families, higher locality and higher periodicity at block 5 are associated with higher skill. For the decay family, the correlation between MAE and MDS has the opposite sign to the correlation between skill and MDS, and the same holds for periodicity. We attribute this to changes in amplitude. Changing d changes how quickly the series decays toward zero, which changes both the MAE and the error of the naive forecast. When the series is close to zero, the naive forecast becomes competitive. For a fixed d, the amplitude also decays faster when the lag is small, which explains the dependence on the lag. This suggests the model relies on structural features that a change in amplitude disrupts.

Our ablation study also shows that adding block 6 is sometimes detrimental. The inconsistency of block 6 suggests either that it plays a different functional role for the decoder, or that it applies an inductive bias that is counterproductive when the structural mode is not well established.

5 Discussion

In this paper, we analyzed the patterns in the attention layers of a time series foundation model by probing it with controlled stimuli and quantifying the resulting attention matrices with locality and periodicity metrics. Our study found that the encoder of Chronos resolves the input into a small

Table 1: Spearman ρ (p-value) between the attention metric read at Chronos-small encoder block 5 and the forecast error of the encoder truncated after block 5. **Bold** marks $p < 0.05$.

Series family	MAE vs. MDS	MAE vs. Periodicity	Skill vs. MDS	Skill vs. Periodicity
Elliptic (k)	$\mathbf{-0.92\ (0.00051)}$	–	$\mathbf{-0.88\ (0.0016)}$	–
Elliptic (period)	$+0.50\ (0.12)$	$\mathbf{-0.65\ (0.03)}$	$+0.54\ (0.089)$	$\mathbf{-0.68\ (0.022)}$
Iterative (decay)	$-0.61\ (0.15)$	$-0.57\ (0.18)$	$\mathbf{+0.79\ (0.036)}$	$+0.71\ (0.071)$
Iterative (lag)	$+0.54\ (0.11)$	$+0.62\ (0.054)$	$\mathbf{+0.67\ (0.033)}$	$\mathbf{+0.68\ (0.029)}$
Logistic (r)	$\mathbf{-0.60\ (0.039)}$	$-0.54\ (0.07)$	$-0.56\ (0.071)$	$\mathbf{-0.65\ (0.032)}$

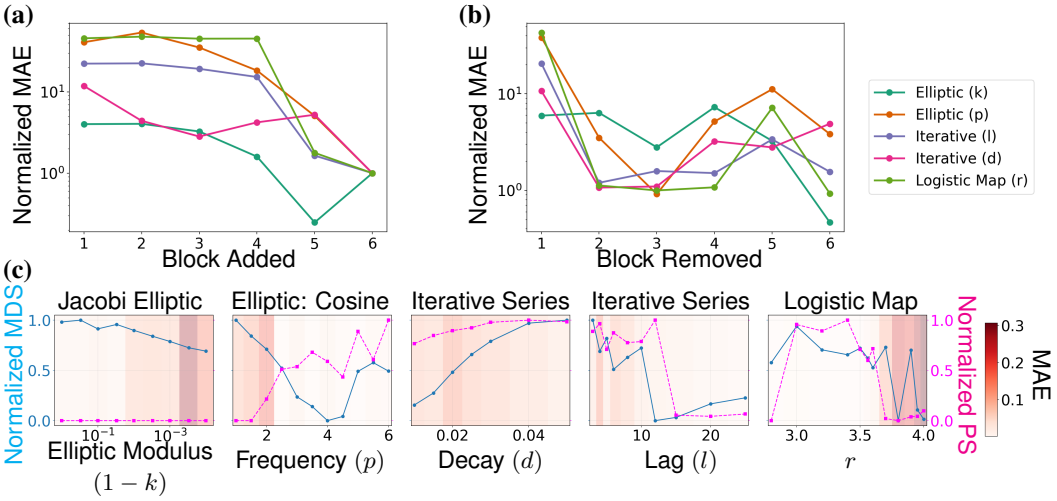

Figure 4: **(a)-(b)** Results from two ablation experiments. For each family of time series, we show the forecast MAE normalized by the MAE of the intact encoder and averaged across the varying parameter (elliptic modulus, frequency, etc.). In **(a)**, we sequentially add blocks and measure the MAE of the forecast. In **(b)**, we remove a block at a time. **(c)** We plot diagonal strength (MDS) and periodicity strength (PS) at block 5 against the ablation MAE when block 5 is added to the encoder. MDS and PS are min-max normalized across blocks and the varying parameter for each family.

set of canonical structural modes that change with depth. In Chronos-small, blocks 1 and 2 are dominated by value-based modes: tokens with equal values attend to each other, and all tokens attend to special points such as minima and maxima. Later attention blocks, especially blocks 5 and 6, then concentrate on local neighborhoods and, when the input is periodic, on points one period apart. Rank correlations and layer pruning connect these modes to forecast accuracy: for several families, forecasts are more accurate when the encoder resolves the input into strong locality or clear periodicity, and less accurate when it resolves neither.

Our conclusions have several limitations. The evidence linking attention modes to forecast accuracy is correlational and based on a small number of parameter values per family, and it does not hold for iterative series, where changes in amplitude increase the forecast error. The encoder's reliance on repeating patterns may also reflect Chronos's training data. Finally, because Chronos computes its relative position bias only in the first block, layer pruning cannot fully separate the role of that block from the role of positional information.

We also report results for Chronos-base (Appendix C) and show how the analysis extends to architectures with patch embedding and causal models, Moirai-1 and Moirai-2 (Appendices D and E). In both Moirai models, locality persists but periodic structure is blurred and only weakly related to forecast error, which suggests that pointwise tokenization is what allows Chronos to represent and exploit periodicity. Prior work has described tokens clustering in self-attention, and our results show explicitly which similarities the tokens of a time series model cluster by: equal values, extrema, temporal neighborhoods, and period shifts. More broadly, they suggest that controlled stimuli are a useful tool for dissecting other foundation models.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data and implement methods via LLM-generated code. We have not used generative AI tools to develop frameworks, propose hypotheses, design experiments, or interpret our results, and the rest of the required disclosure tasks are not applicable to this work. We used generative AI tools to create figures, identify relevant literature, and create the code repository for reproducibility, as well as a website for additional results. We reviewed all AI-assisted work. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The code used for our experiments is currently available in an anonymous repository at `https://anonymous.4open.science/r/foundation-model-forecasting-9133`. We note specific parameter choices in the appendix. Table 2 is a quick guide to the parameters used for our synthetic data.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *TMLR*, November 2024. doi: 10.48550/arXiv.2403.07815.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025. doi: 10.1109/TPAMI.2024.3506283.

Anthony Bao, Venkata Hasith Vattikuti, Jeffrey Lai, and William Gilpin. Universal redundancies in time series foundation models. *arXiv preprint arXiv:2602.01605*, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv*, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6, 2023.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, October 2023.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, April 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *37th Conference on Neural Information Processing Systems*, 2023.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers, August 2025.

Syamantak Datta Gupta. Evaluating time series foundation models on noisy periodic time series, January 2025.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. *arXiv*, 2022.

Matīss Kalnāre, Sofoklis Kitharidis, Thomas Bäck, and Niki van Stein. Mechanistic interpretability for transformer-based time series classification, November 2025.

Nouha Karaouli, Denis Coquenet, Elisa Fromont, Martial Mermillod, and Marina Reyboz. How foundational are foundation models for time series forecasting? In *NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models*, 2025.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4365–4374, 2019.

Derek F Lawden. *Elliptic functions and applications*. Springer Science & Business Media, 2013.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6555–6565, 2024.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

Yao Lu, Hao Cheng, Yujie Fang, Zeyu Wang, Jiaheng Wei, Dongwei Xu, Qi Xuan, and Zhaowei Zhu. Reassessing layer pruning in LLMs: New insights and methods. In *The Fourteenth International Conference on Learning Representations*, 2026.

Anurag Mishra. Dissecting chronos: Sparse autoencoders reveal causal feature hierarchies in time series foundation models, March 2026.

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.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

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

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arxiv*, 2024.

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.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *ICL*, June 2019. doi: 10.48550/arXiv.1906.03731.

Boyu Shi, Chang Liu, ChuanBao Gao, Xu Yang, and Xin Geng. Understanding performance collapse in layer-pruned large language models via decision representation transitions. *arXiv preprint arXiv:2605.07271*, 2026.

William Toner, Thomas L Lee, Artjom Joosen, Rajkarn Singh, and Martin Asenov. Performance of zero-shot time series foundation models on cloud data. In *I Can't Believe It's Not Better ICLR Workshop*, 2025.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4344–4353, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in neural information processing systems*, volume 30. Curran Associates, Inc., 2017.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model, June 2019.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41 st International Conference on Machine Learning*, May 2024. doi: 10.48550/arXiv.2402.02592.

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.

Shucong Zhang, Erfan Loweimi, Peter Bell, and Steve Renals. On the usefulness of self-attention for automatic speech recognition with transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 89–96. IEEE, 2021.

Yunkai Zhang, Qi Zeng, Yawen Zhang, Zhijie Xu, Ming Zheng, Chongyang Gao, Muyan Jiang, and Zeyu Zheng. Are time series foundation models ready for zero-shot forecasting? In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, 16(12):9851–9915, 2025.

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.

A Locality Metrics: Uniform Case

- **Expectation over distance**:

$$\frac{1}{n}\sum_{i,j=1}^{n} X_{ij}|i-j| = \frac{1}{n^2}\sum_{i,j=1}^{n}|i-j| = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(j-i) \tag{5}$$

$$= \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(j-i) = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n-i} j \tag{6}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}(n-i)(n-i+1) = \frac{1}{n^2}\sum_{i=1}^{n}(n^2-2ni+n-i+i^2) \tag{7}$$

$$= \frac{1}{n^2}(n^3+n^2-(2n+1)\sum_{i=1}^{n} i + \sum_{i=1}^{n} i^2) \tag{8}$$

$$= n+1-\frac{(2n+1)(n+1)}{2n} + \frac{(2n+1)(n+1)}{6n} \tag{9}$$

$$= (n+1)[1-\frac{(2n+1)}{3n}] = \frac{n^2-1}{3n} \tag{10}$$

- **Diagonal Strength**: For the uniform attention matrix, $X_{ij} = 1/n$. The denominator is $(1/n) * n^2 = n$. For the numerator, assuming that w is an integer, the diagonal band has $n + 2(n-1) + 2(n-2) + \cdots + 2(n-w)$ many entries. So the numerator is

$$\frac{1}{n}(n + \sum_{i=1}^{w} 2(n-i)) = \frac{n+2nw-w(w+1)}{n} \tag{11}$$

  Thus, the diagonal strength for the uniform attention matrix is $\frac{n+2nw-w(w+1)}{n^2}$. We can also let $w = \lfloor pn \rfloor$ where $p \in [0,1]$.

B Additional Details on Chronos-small

B.1 Experimental Details

For Chronos-small, we generate time series with varying number of timesteps/tokens and sample the context sequence for the model to predict from. We outline these selections in table 2. For all time series generated, the prediction length is fixed at 50 as Chronos performs best limited to a prediction length less than or equal to 64.

Table 2: Reference table for time series data generated for experiments with Chronos-small.

Time Series	Generated Number of Tokens	Context tokens	Fixed parameters
Jacobi elliptic	100	1-50	
Cosine (varying frequencies)	500	1-250	
Iterative Series (4) (varying lag l)	100	1-50	$d = 0.05$
Iterative Series (4) (varying decay d)	100	1-50	$l = 3$
Logistic map	1000	150-250	$x_0 = 0.9$

B.2 Attention Matrices

In this section, we show all the average attention matrices for each of the time series data studied in this paper. See table 3 for a quick reference. Figure 5 shows an example of all heads in each attention block for the Chronos-small model. We show all the average attention matrices in figures

6-10 for our various time series. Figures 11, 12-16 display the attention matrices extracted when position encoding is null (NoPE). We observed that when we remove the position encoding from the model, the main diagonal is significantly reduced and global patterns from early blocks remain in the later blocks.

Table 3: Reference table for Chronos-small attention head figures.

Time Series	Average Attention Heads	NoPE Average Attention Heads
Jacobi elliptic	6	12
Cosine (varying frequencies)	7	13
Iterative Series (4) (varying lag l)	8	14
Iterative Series (4) (varying decay d)	6	12
Logistic map	10	16

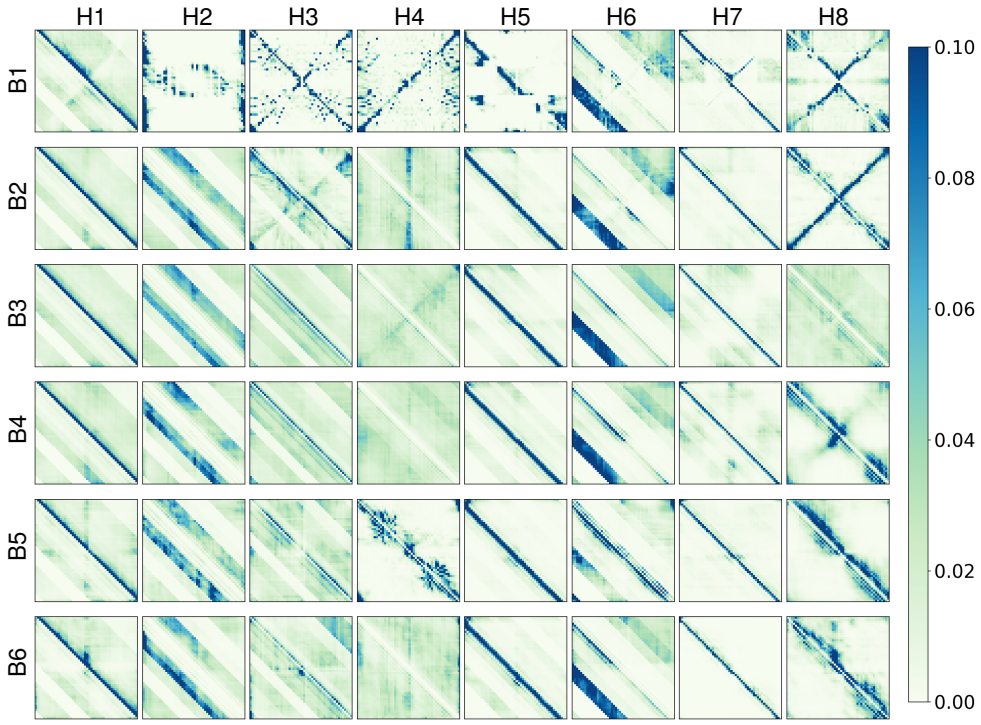

Figure 5: Each attention block (*row*) in Chronos-small has 8 heads (*column*). We show the attention weights for each of these heads in this figure for a standard cosine function.

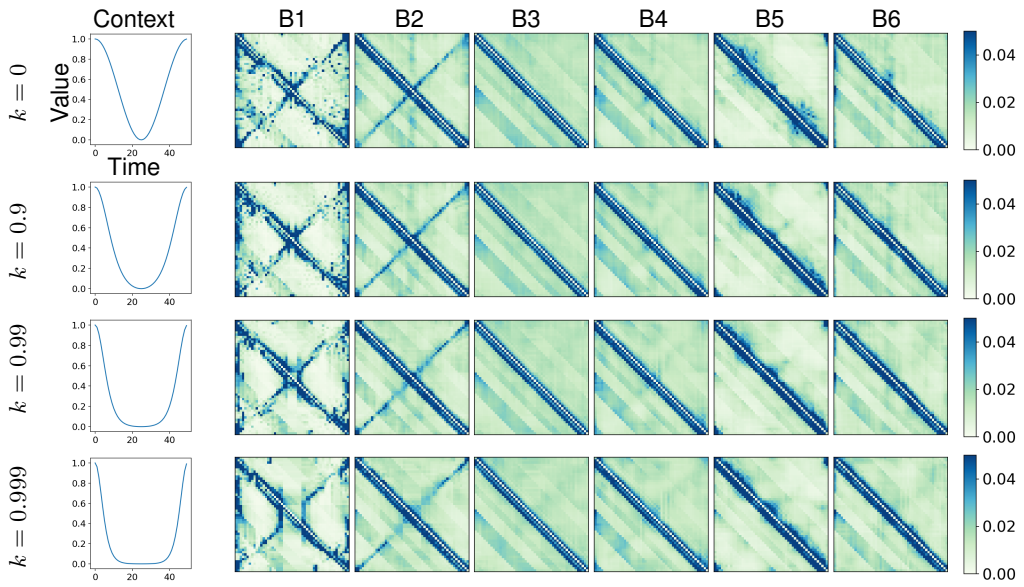

Figure 6: Average attention matrices for Jacobi elliptic functions with varying modulus k. Early blocks capture patterns such as symmetry, while later blocks exhibit increasingly localized attention patterns.

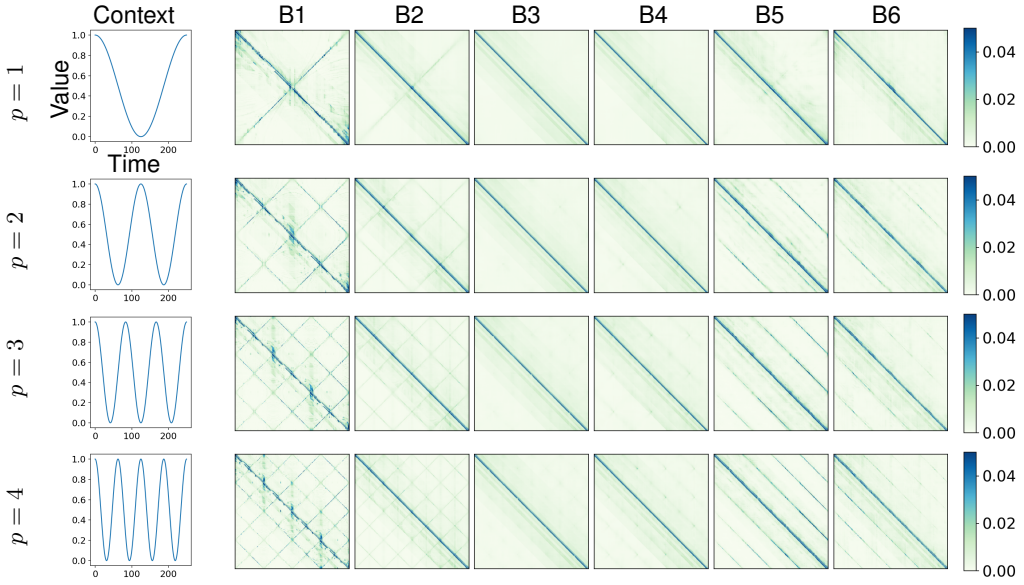

Figure 7: Average attention matrices for a cosine function with varying number of cycles. For sequences with higher frequency, the frequency is captured as off-diagonal bands, noticeable at the later blocks.

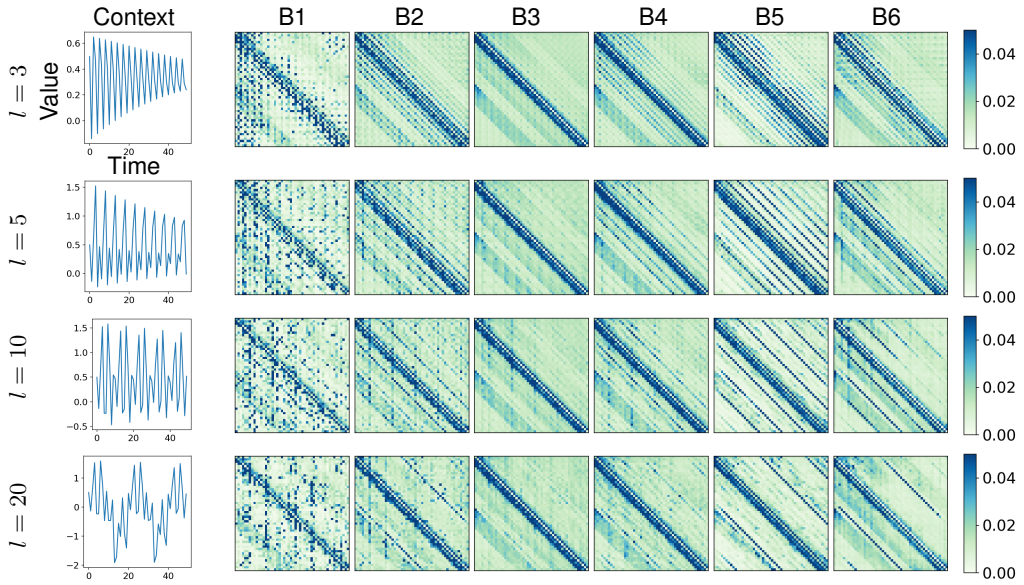

Figure 8: Average attention matrices for the iterative series 4 with varying lag l, decay 0.05. Maximum and minimum values are captured in early layers while periodicity is captured in later layers.

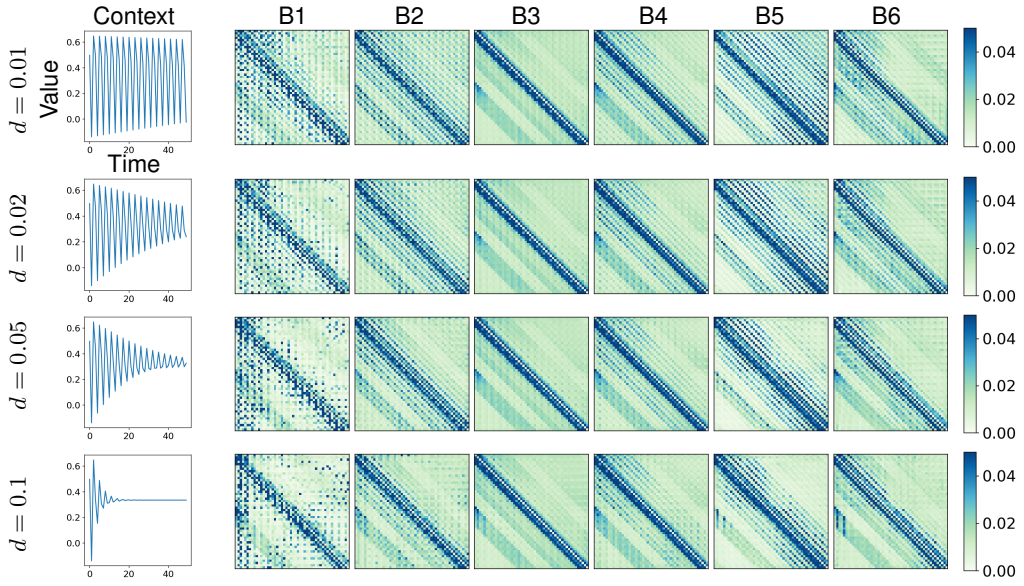

Figure 9: Average attention matrices for the iterative series 4 with varying decay d, lag 3. For time series with greater decay, attention is not concentrated on time steps with similar values.

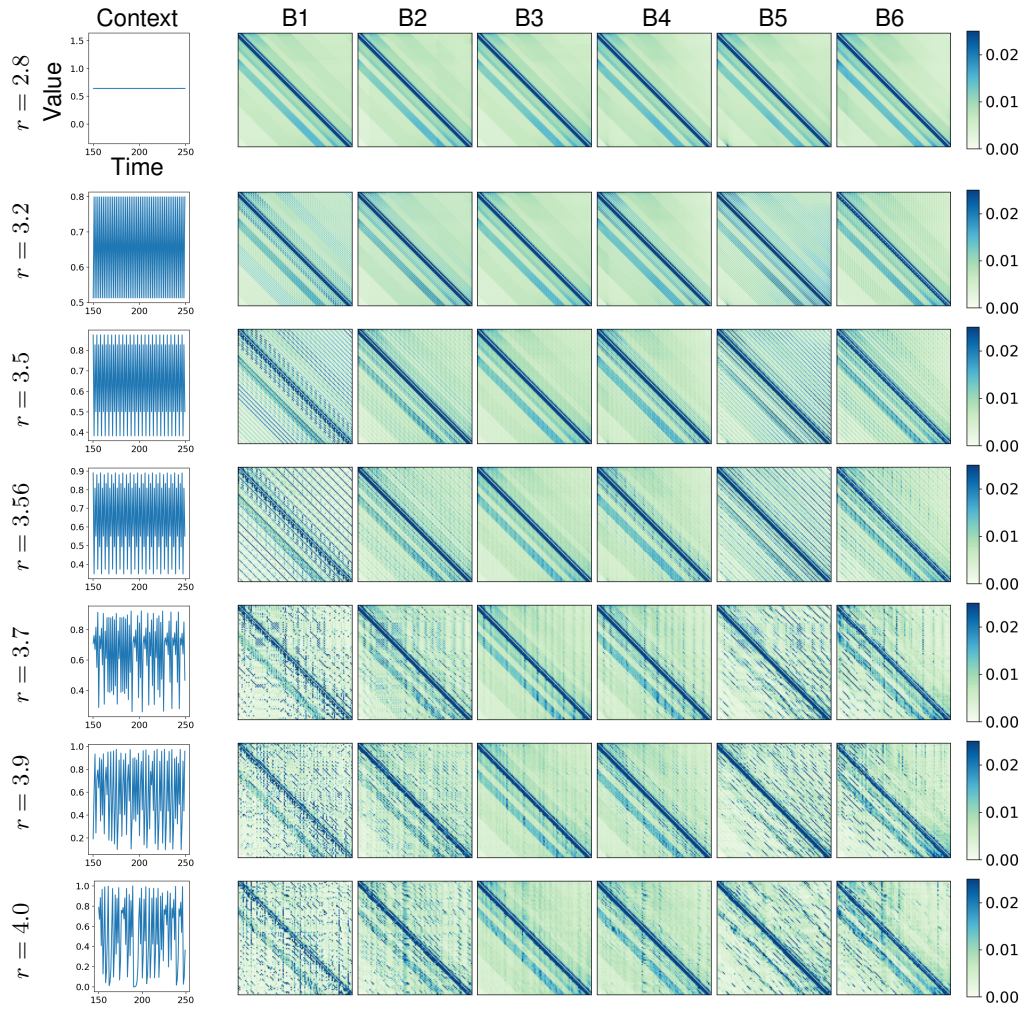

Figure 10: Average attention matrices for the logistic map with varying growth rate r. The transition from periodic to chaotic behavior is reflected in the changing attention patterns.

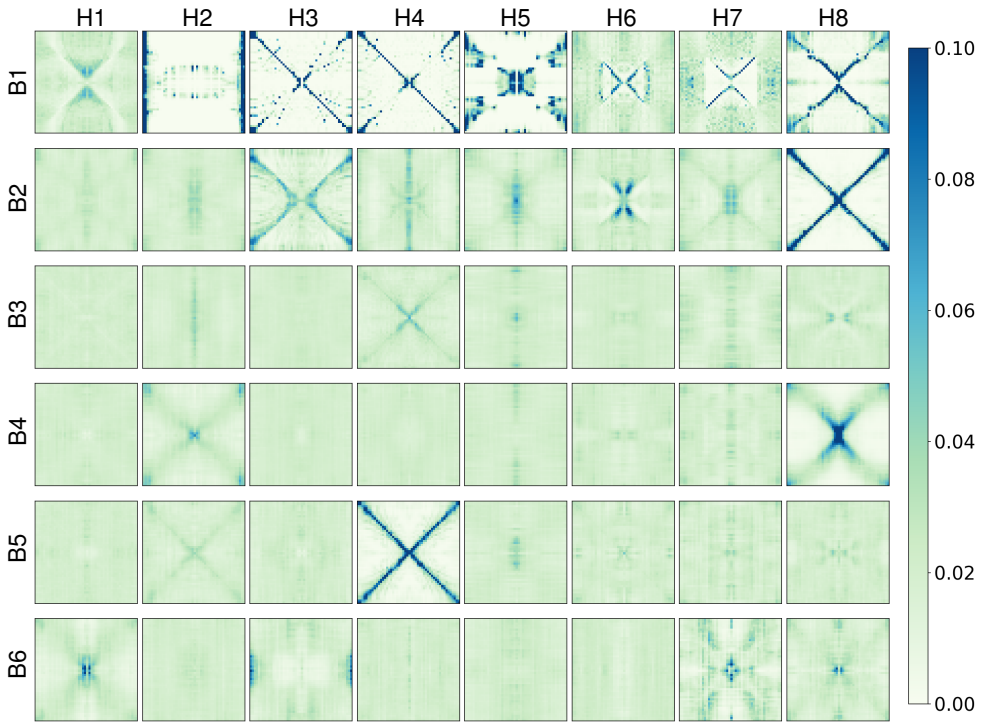

Figure 11: Attention heads in Chronos' encoder for a standard cosine where position encoding is ablated.

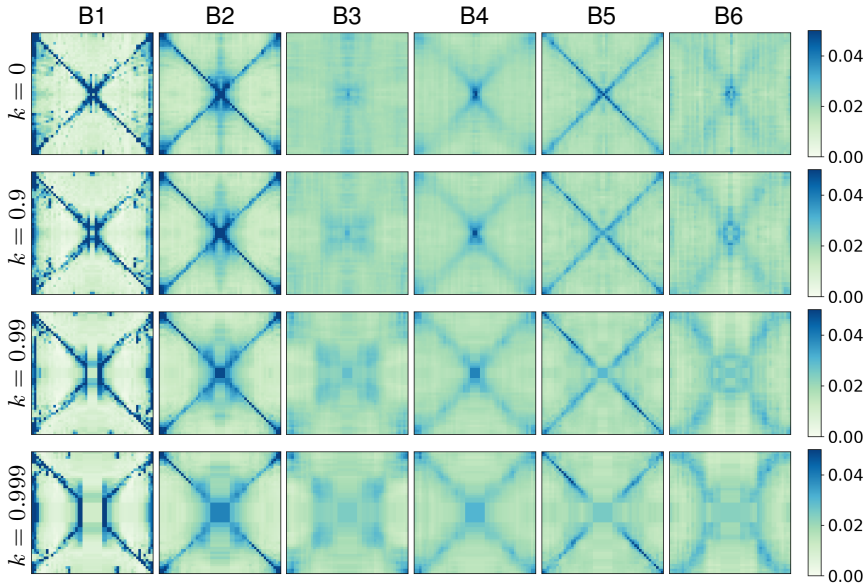

Figure 12: Average attention matrices for Jacobi elliptic functions with varying modulus k and zero position encoding. These matrices show that localized attention patterns viewed in 6 comes from the position encoding.

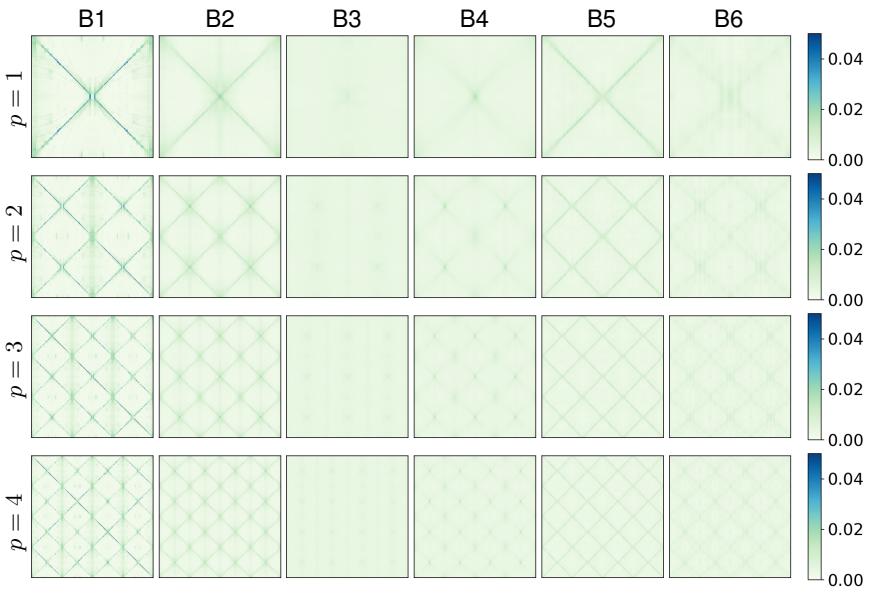

Figure 13: Average attention matrices for cosines with varying frequencies and zero position encoding. These matrices show that localized periodic attention patterns observed in 7 comes from the positonal encoding.

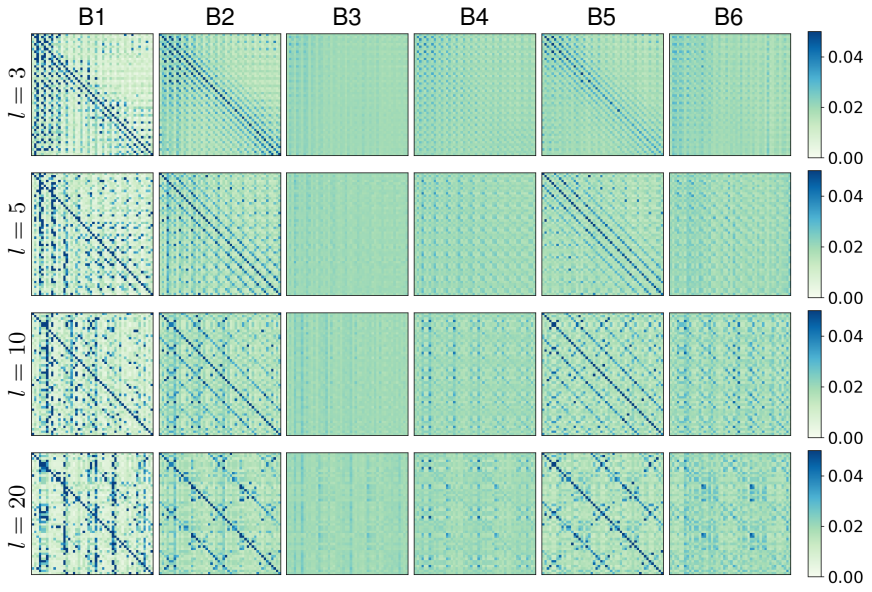

Figure 14: Average attention matrices for iterative series with varying lag and zero position encoding. Unlike 8, the off diagonal bands indicating frequencies are still present despite the zeroed out position encoding, however there are fewer localized patterns.

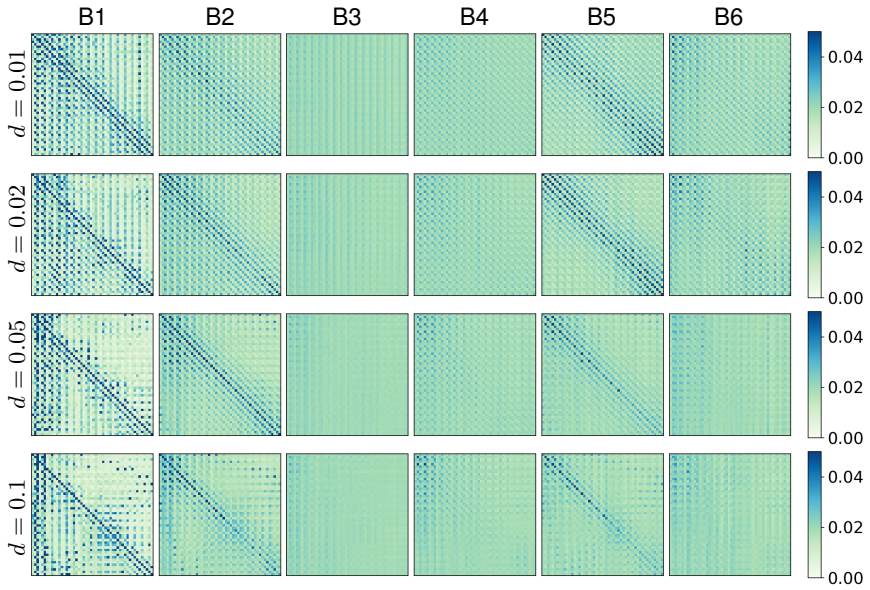

Figure 15: Average attention matrices for iterative series with varying decay and zero position encoding. Patterns are similar to 9 but the diagonal is less prominent.

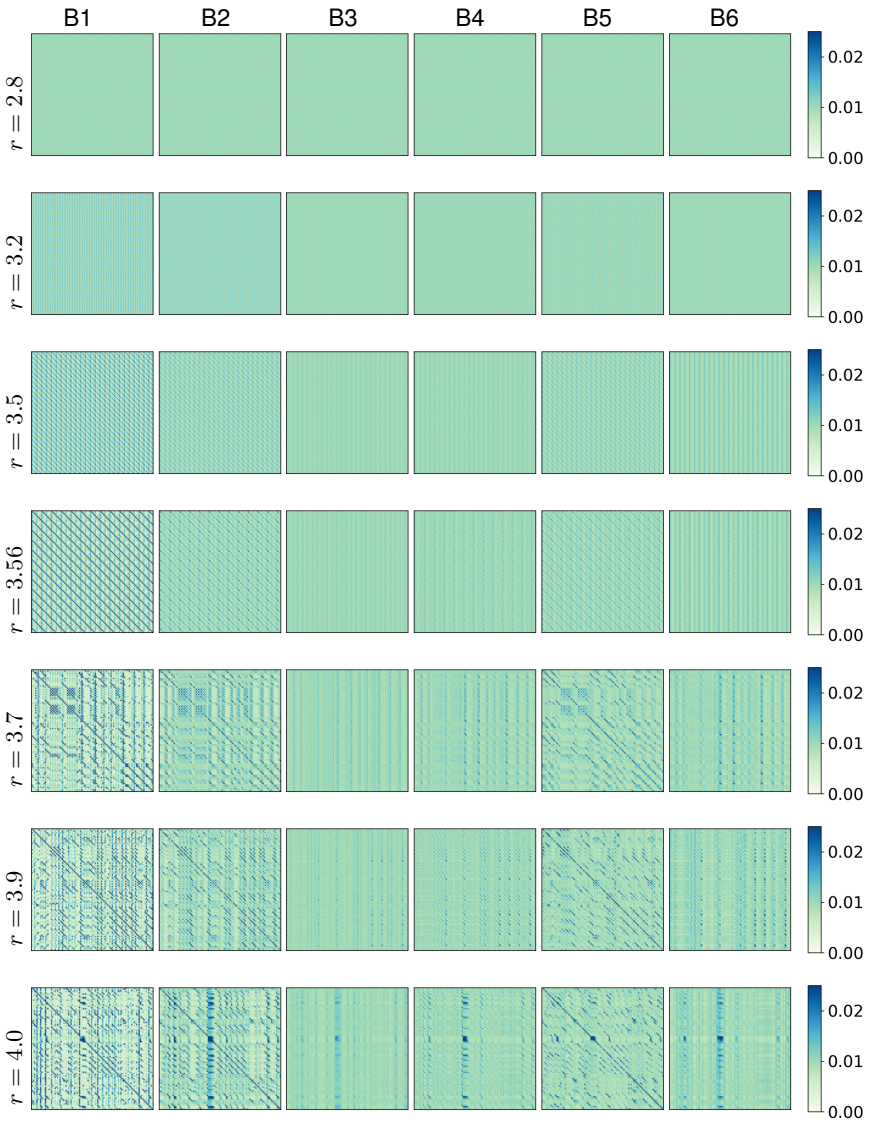

Figure 16: Average attention matrices for the logistic map with varying growth rate r and zero position encoding. Patterns are similar to 10 but the diagonal is less prominent.

B.3 Ablation Study

Detailed results of our ablation study are shown in this section. Figures 17 through 22 display the forecast results for a selected number of time series. We perform two types of ablation: 1) adding blocks sequentially and 2) removing a block one at a time. Figures 23 through 27 show the mean absolute error of the forecasts. See table 4 for a quick reference.

Table 4: Reference table for ablation results on Chronos-small.

Time Series	Adding Blocks Forecasts	Removing Blocks Forecasts	Forecast Error
Jacobi elliptic	17	20	23
Cosine (varying frequencies)	18	21	24
Iterative Series (4) (varying lag l)	19	22	25
Iterative Series (4) (varying decay d)			26
Logistic map			27

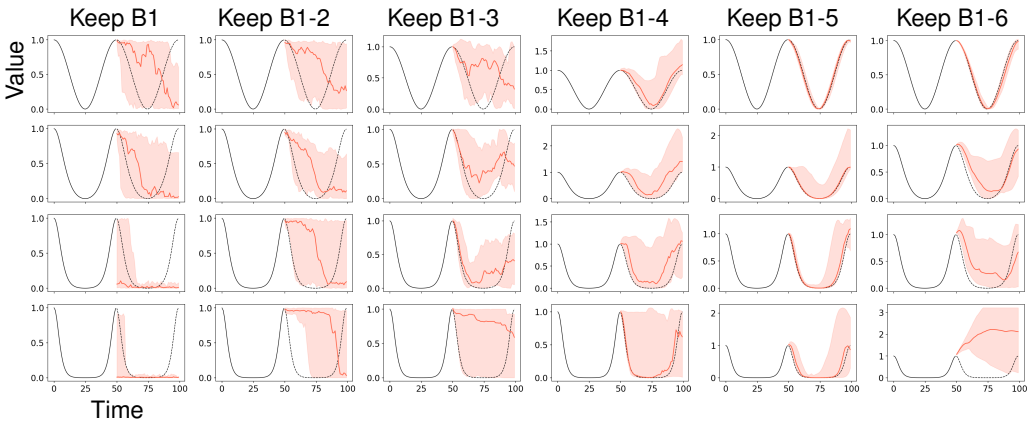

Figure 17: Truncated encoder forecasts for Jacobi elliptic functions ($k = 0, 0.9, 0.99, 0.999$, *top-bottom row*)

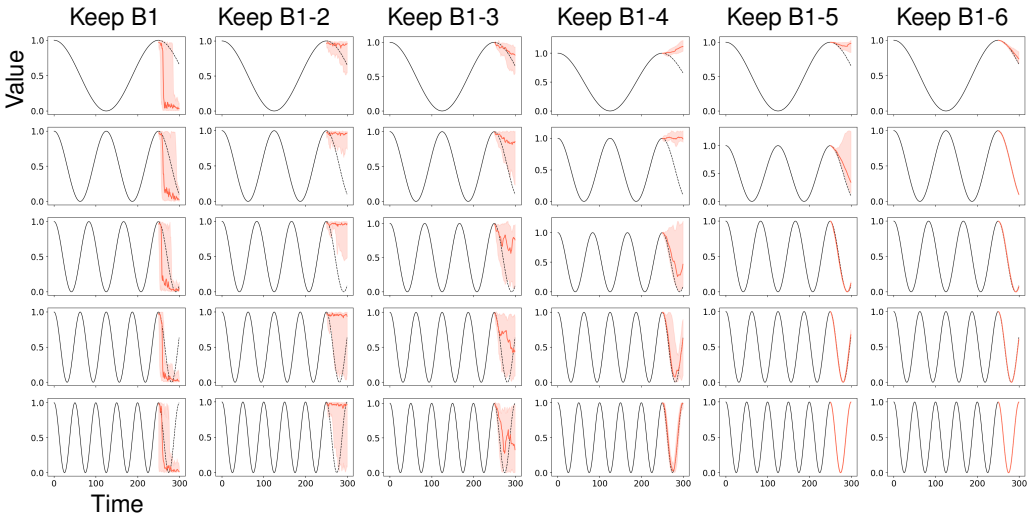

Figure 18: Truncated encoder forecasts for cosine functions (number of cycles 1,2,3,4,5, *top-bottom row*)

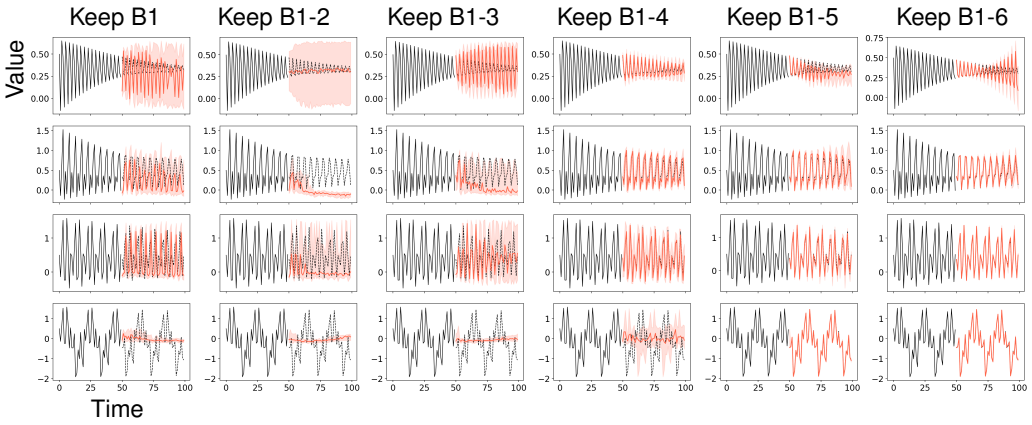

Figure 19: Truncated encoder forecasts for iterative series (4) ($l = 3, 5, 10, 20$, *top-bottom row*)

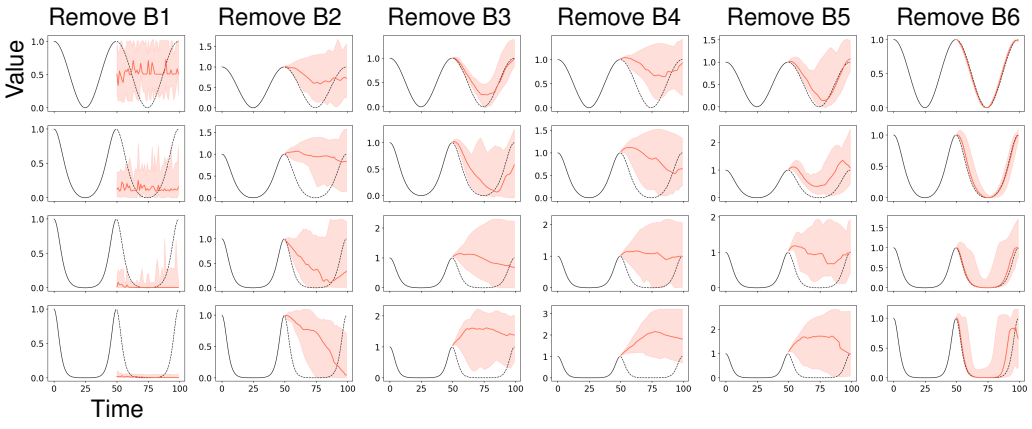

Figure 20: Block removal forecasts for Jacobi elliptic functions ($k = 0, 0.9, 0.99, 0.999$, *top-bottom row*)

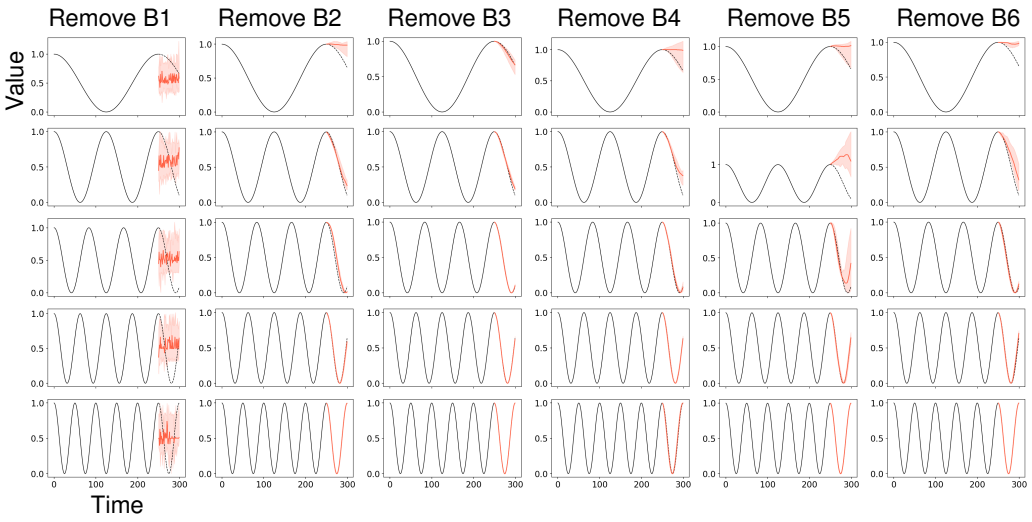

Figure 21: Block removal forecasts for cosine functions (number of cycles 1,2,3,4,5, *top-bottom row*)

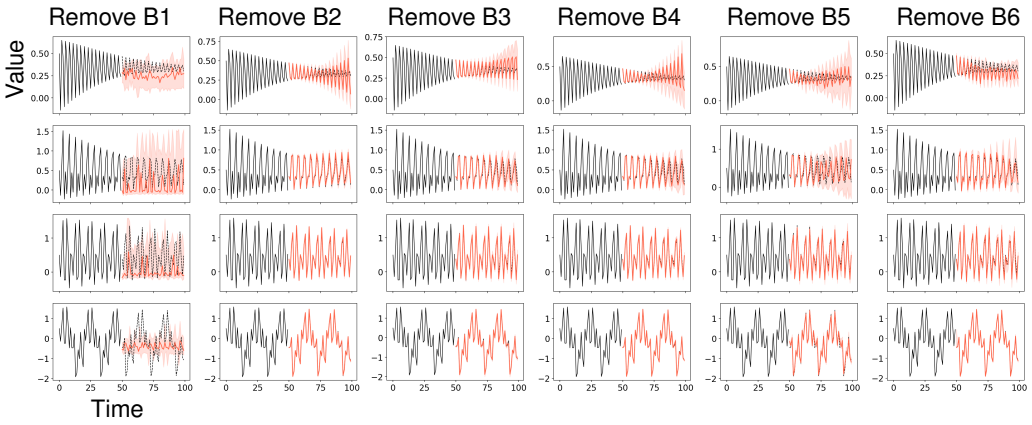

Figure 22: Block removal forecasts for iterative series (4) ($l = 3, 5, 10, 20$, *top-bottom row*)

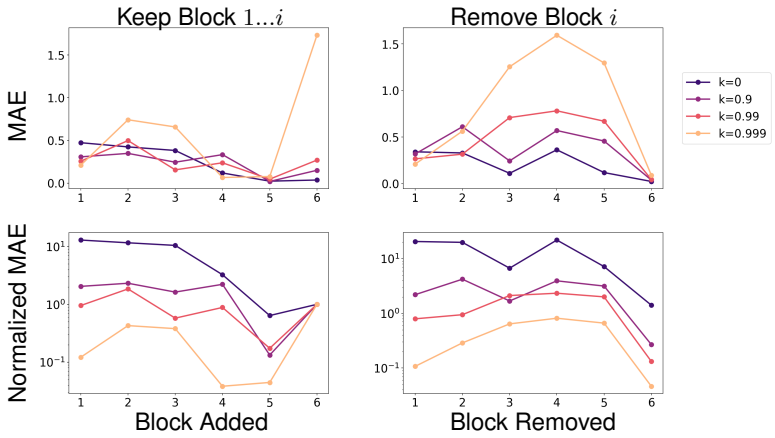

Figure 23: Mean absolute error (MAE) from two ablation studies for the Jacobi elliptic functions (varying modulus k). On the left, we measure MAE as blocks are sequentially added. On the right, we measure MAE after the block has been removed. Normalized MAE is computed by normalizing MAE with respect to the error when all blocks are intact.

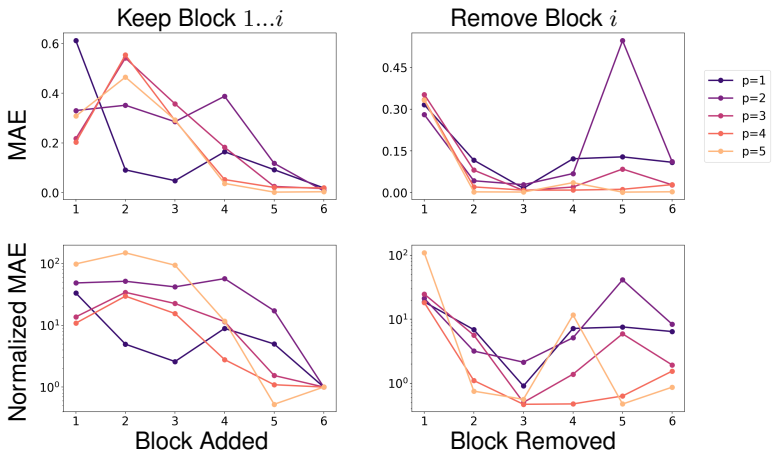

Figure 24: Mean absolute error (MAE) from two ablation studies for the cosine function with various frequencies (number of cycles p). On the left, we measure MAE as blocks are sequentially added. On the right, we measure MAE after the block has been removed. Normalized MAE is computed by normalizing MAE with respect to the error when all blocks are intact.

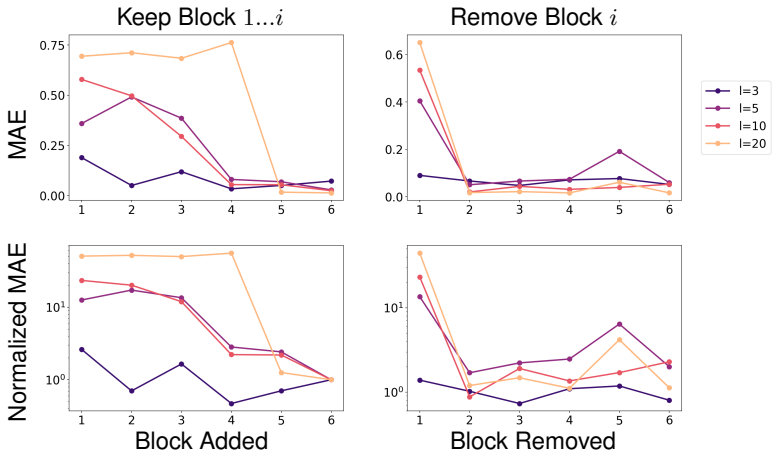

Figure 25: Mean absolute error (MAE) from two ablation studies for the iterative series (4) for various values of lags (l). On the left, we measure MAE as blocks are sequentially added. On the right, we measure MAE after the block has been removed. Normalized MAE is computed by normalizing MAE with respect to the error when all blocks are intact.

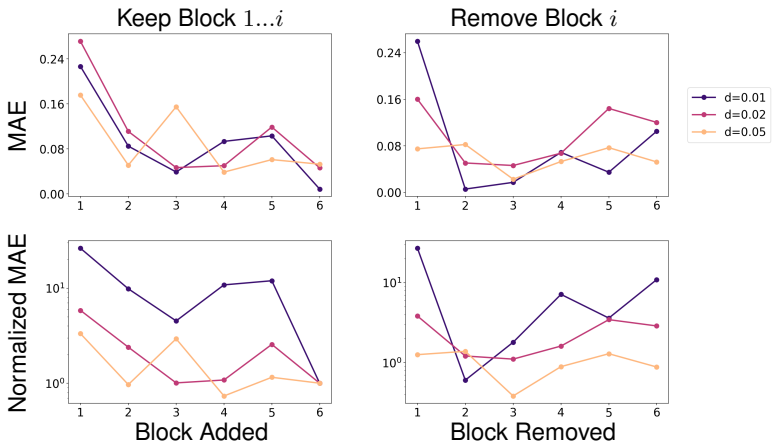

Figure 26: Mean absolute error (MAE) from two ablation studies for the iterative series (4) for various values of d. On the left, we measure MAE as blocks are sequentially added. On the right, we measure MAE after the block has been removed. Normalized MAE is computed by normalizing MAE with respect to the error when all blocks are intact.

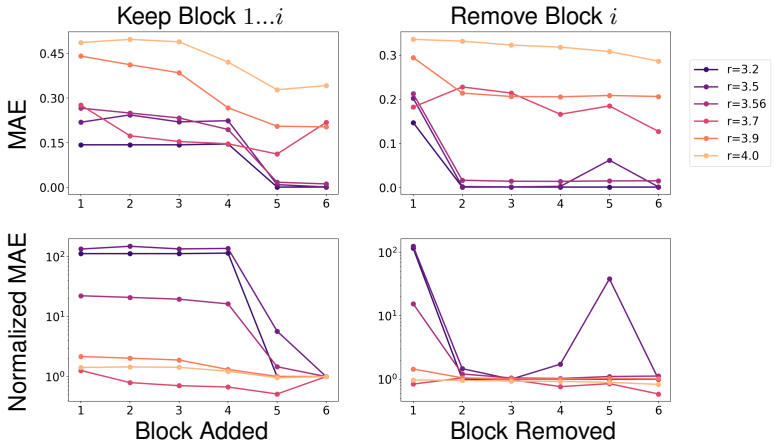

Figure 27: Mean absolute error (MAE) from two ablation studies for the logistic map for various values of r. On the left, we measure MAE as blocks are sequentially added. On the right, we measure MAE after the block has been removed. Normalized MAE is computed by normalizing MAE with respect to the error when all blocks are intact.

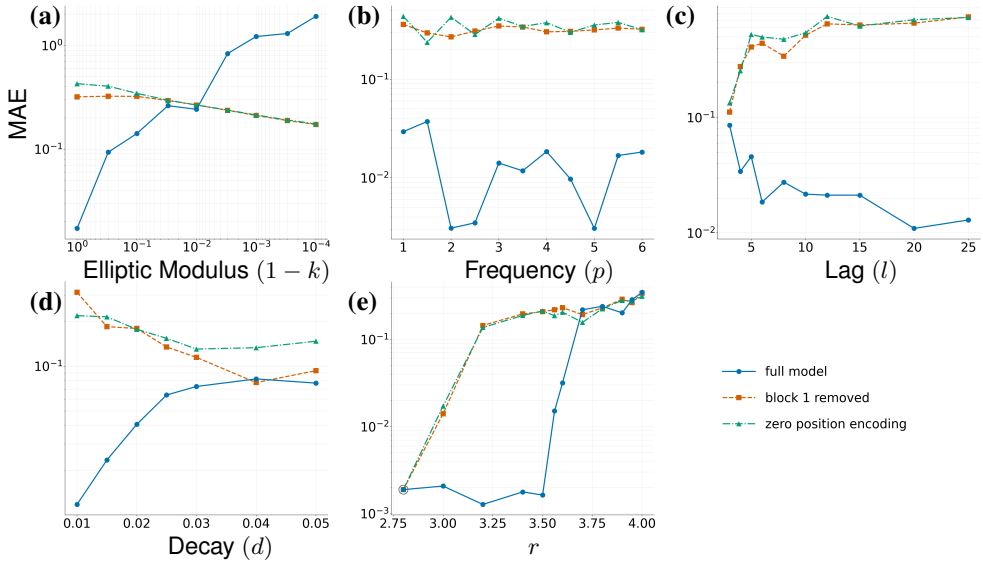

Figure 28: This compares the mean absolute error of ablated models and the full model when given **(a)** Jacobi elliptic functions, **(b)** cosine functions of varying frequencies, **(c)** iterative series 4 varying lags and **(d)** decays, and **(e)** the logistic map. In particular, we can observe similarities between models without block 1 and models without position encoding - suggesting that the increase in error from removing block 1 can be attributed to removing position encoding.

Table 5: Spearman ρ (p-value) between the attention metric read at encoder block i and the forecast error of the model with the encoder truncated after block i, for every depth. **Bold** marks $p < 0.05$.

Series family	MAE vs. MDS	MAE vs. Periodicity	Skill vs. MDS	Skill vs. Periodicity
Block 1				
Elliptic (k)	$+0.13$ (0.73)	–	$+0.08$ (0.83)	–
Elliptic (period)	$+0.11$ (0.75)	-0.37 (0.26)	$+0.21$ (0.54)	-0.32 (0.34)
Iterative (decay)	-0.46 (0.29)	-0.71 (0.071)	-0.07 (0.88)	$+0.46$ (0.29)
Iterative (lag)	-0.53 (0.12)	-0.39 (0.26)	$+0.20$ (0.58)	$+0.18$ (0.63)
Logistic (r)	-0.33 (0.3)	-0.57 (0.055)	-0.48 (0.13)	**-0.68 (0.022)**
Block 2				
Elliptic (k)	$+0.17$ (0.67)	–	-0.27 (0.49)	–
Elliptic (period)	**-0.91 (0.00011)**	–	-0.55 (0.083)	–
Iterative (decay)	**-0.82 (0.023)**	**$+0.82$ (0.023)**	$+0.43$ (0.34)	-0.43 (0.34)
Iterative (lag)	**-0.81 (0.0049)**	-0.32 (0.37)	-0.13 (0.73)	$+0.05$ (0.88)
Logistic (r)	**-0.80 (0.0016)**	-0.01 (0.98)	-0.49 (0.13)	-0.22 (0.52)
Block 3				
Elliptic (k)	$+0.47$ (0.21)	$+0.43$ (0.24)	$+0.33$ (0.38)	$+0.30$ (0.43)
Elliptic (period)	-0.27 (0.42)	–	$+0.31$ (0.36)	–
Iterative (decay)	$+0.75$ (0.052)	–	$+0.39$ (0.38)	–
Iterative (lag)	-0.53 (0.12)	–	-0.14 (0.7)	–
Logistic (r)	**-0.83 (0.00079)**	–	-0.34 (0.31)	–
Block 4				
Elliptic (k)	$+0.07$ (0.86)	–	$+0.30$ (0.43)	–
Elliptic (period)	**$+0.64$ (0.035)**	–	**$+0.80$ (0.0031)**	–
Iterative (decay)	-0.75 (0.052)	**$+0.82$ (0.023)**	-0.14 (0.76)	$+0.21$ (0.64)
Iterative (lag)	-0.60 (0.067)	-0.16 (0.65)	-0.60 (0.067)	-0.02 (0.96)
Logistic (r)	**-0.64 (0.024)**	-0.39 (0.21)	$+0.26$ (0.43)	$+0.10$ (0.77)
Block 5				
Elliptic (k)	**-0.92 (0.00051)**	–	**-0.88 (0.0016)**	–
Elliptic (period)	$+0.50$ (0.12)	**-0.65 (0.03)**	$+0.54$ (0.089)	**-0.68 (0.022)**
Iterative (decay)	-0.61 (0.15)	-0.57 (0.18)	**$+0.79$ (0.036)**	$+0.71$ (0.071)
Iterative (lag)	$+0.54$ (0.11)	$+0.62$ (0.054)	**$+0.67$ (0.033)**	**$+0.68$ (0.029)**
Logistic (r)	**-0.60 (0.039)**	-0.54 (0.07)	-0.56 (0.071)	**-0.65 (0.032)**
Block 6				
Elliptic (k)	$+0.33$ (0.38)	–	$+0.33$ (0.38)	–
Elliptic (period)	$+0.27$ (0.42)	-0.35 (0.3)	$+0.51$ (0.11)	-0.52 (0.1)
Iterative (decay)	**$+0.89$ (0.0068)**	**$+0.96$ (0.00045)**	**$+0.96$ (0.00045)**	**$+1.00$ (<1e-04)**
Iterative (lag)	$+0.55$ (0.098)	$+0.27$ (0.45)	**$+0.79$ (0.0061)**	$+0.42$ (0.23)
Logistic (r)	**-0.79 (0.0022)**	$+0.11$ (0.72)	**-0.74 (0.0098)**	$+0.10$ (0.77)

C CHRONOS-BASE

Using the same input series as Chronos-small (table 2), we compute locality metrics (figure 29), periodicity metrics (figure 30), and ablation results (figures 31, 32) for the Chronos-base model. Chronos-base has 12 attention blocks in its encoder and 109M parameters compared to 35M in the smaller model. 29 shows the locality profiles for our time series data where we can observe a similar trend as the smaller model; locality tends to be higher at the later blocks (blocks 9-12). 30 shows the periodicity profiles and reveals that later blocks, in particular block 10, tend to distinguish between time series with periodicity and those without.

Our ablation results in figures 31 and 32 also show similar trends as the smaller model. Mean absolute error decreases significantly when later blocks like 9 and 10 are added to the encoder. Additionally, strong correlations between forecast error and locality/periodicity are spread throughout blocks 9-12 (table 6). The correlation trends are different for Chronos-base. While, the forecasts on Jacobi elliptic functions are strongly correlated with locality (blocks 9 and 10), there is weak correlation for the cosine functions with both locality and periodicity. Forecasts on the logistic map are more strongly correlated with periodicity.

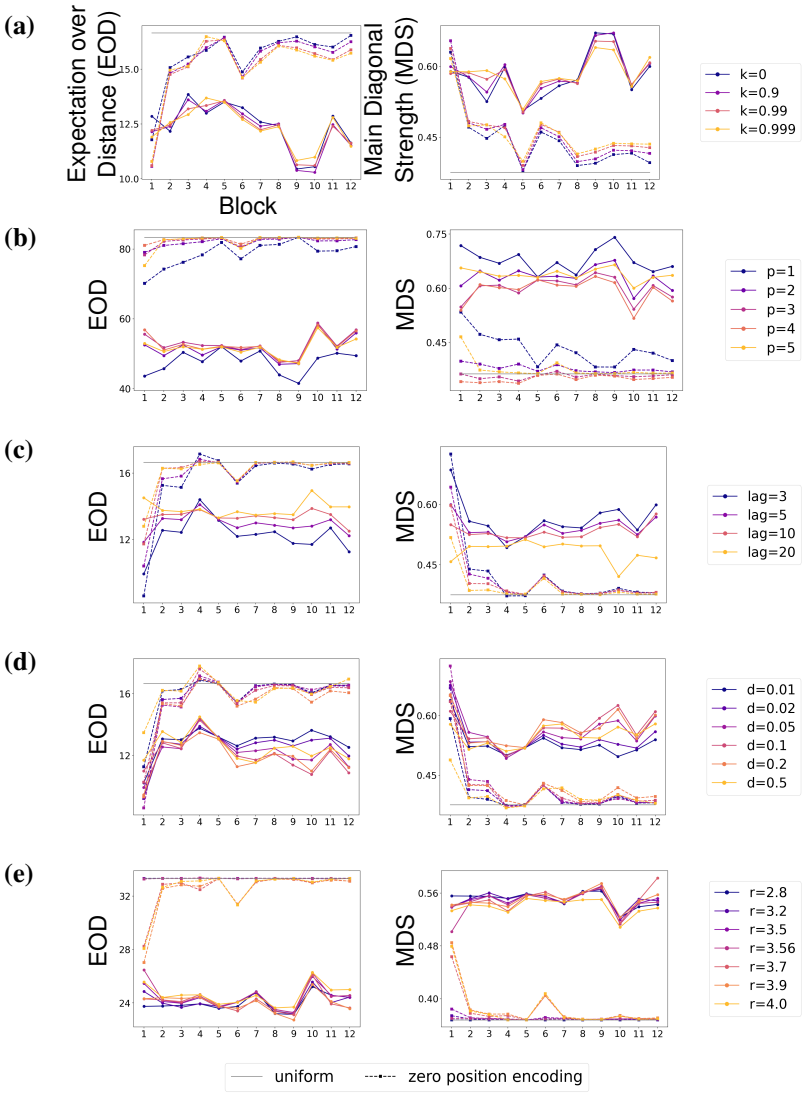

Figure 29: Locality profiles for **(a)** Jacobi elliptic functions (for various modulus k), **(b)** cosine function (Jacobi elliptic function with $k = 0$) with different number of cycles (p), **(c)-(d)** the iterative series $x_{n+1} = \sum_{i=0}^{l} c_i x_{n-i}$ 4, and **(e)** logistic map $x_{n+1} = r x_n(1 - x_n)$ for varied r. In **(c)**, we vary values of l with fixed $d = 0.05$, and in **(d)**, we vary d and fix $l = 3$. Expectation over distance (EOD) and main diagonal strength (MDS) are computed across encoder blocks with and without position bias (solid vs dashed line respectively). Locality generally peaks at block 9-10 except for the logistic map.

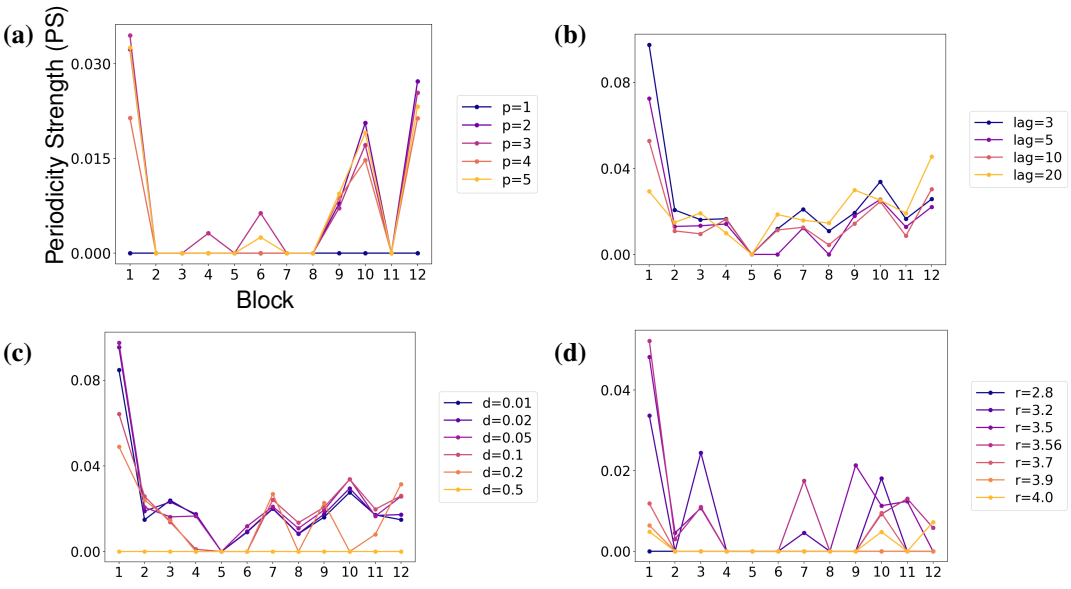

Figure 30: Periodicity strength of **(a)** cosine function with varying frequencies, **(b)** iterative time series with varying lags l and **(c)** decays d, and **(d)** for the logistic map with varying r. Periodicity strength (PS) is computed across encoder blocks for each dynamical system.

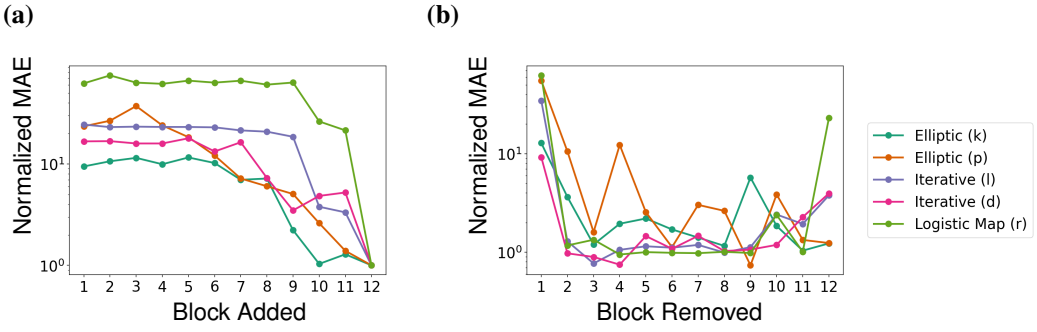

Figure 31: **(a)-(b)** We obtain results from two ablation experiments. We show the normalized MAE averaged across the varying parameter (elliptic modulus k, frequency/number of cycles p, etc.) for each set of time series we consider in this paper. In **(a)** we sequentially add blocks and measure the mean absolute error (MAE) of the forecast. In **(b)**, we remove a block from the encoder and measure MAE. MAE is then normalized by the error when all blocks are intact.

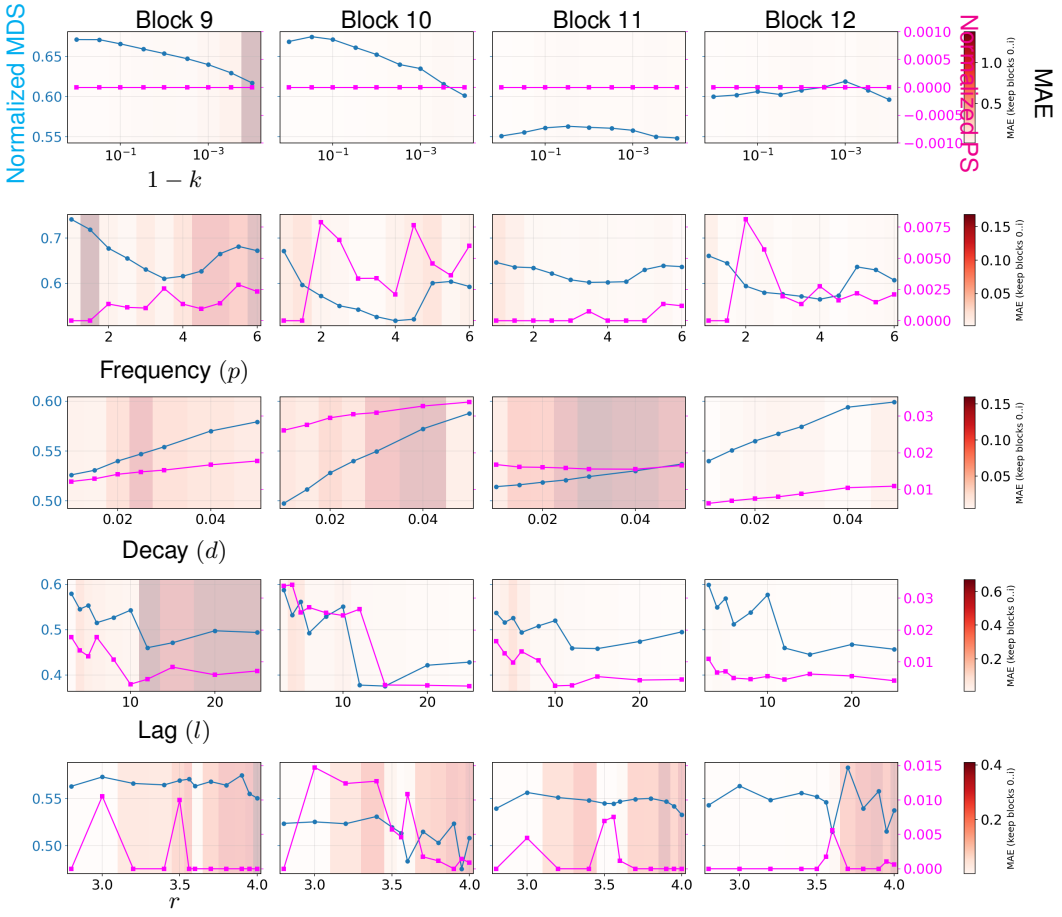

Figure 32: We plot diagonal strength (MDS) and periodicity strength (PS) at block 9-12 against the ablation MAE when each block is added to the encoder.

Table 6: Spearman ρ (p-value) between the attention metric read at Chronos-base encoder block i and the forecast error of the model with the encoder truncated after block i, for every depth. Blocks 6-12 shown here. **Bold** marks $p < 0.05$.

Series family	MAE vs. MDS	MAE vs. Periodicity	Skill vs. MDS	Skill vs. Periodicity
Block 6				
Elliptic (k)	**$+0.87$ (0.0025)**	–	**$+0.68$ (0.042)**	–
Elliptic (period)	-0.27 (0.42)	$+0.36$ (0.27)	$+0.22$ (0.52)	-0.07 (0.85)
Iterative (decay)	-0.46 (0.29)	-0.54 (0.22)	$+0.71$ (0.071)	$+0.64$ (0.12)
Iterative (lag)	**-0.88 (0.00081)**	**$+0.72$ (0.02)**	-0.15 (0.68)	$+0.43$ (0.22)
Logistic (r)	-0.45 (0.14)	–	-0.35 (0.3)	–
Block 7				
Elliptic (k)	$+0.60$ (0.088)	–	$+0.47$ (0.21)	–
Elliptic (period)	-0.45 (0.16)	–	-0.05 (0.87)	–
Iterative (decay)	-0.46 (0.29)	-0.64 (0.12)	**$+0.96$ (0.00045)**	**$+0.93$ (0.0025)**
Iterative (lag)	**-0.77 (0.0092)**	-0.52 (0.13)	-0.43 (0.21)	-0.30 (0.4)
Logistic (r)	$+0.06$ (0.85)	-0.48 (0.11)	-0.31 (0.36)	-0.22 (0.51)
Block 8				
Elliptic (k)	$+0.32$ (0.41)	–	$+0.23$ (0.55)	–
Elliptic (period)	-0.01 (0.98)	–	$+0.55$ (0.083)	–
Iterative (decay)	**$+0.82$ (0.023)**	**$+0.82$ (0.023)**	**$+0.96$ (0.00045)**	**$+0.96$ (0.00045)**
Iterative (lag)	**-0.84 (0.0022)**	**$+0.71$ (0.023)**	-0.30 (0.4)	$+0.58$ (0.08)
Logistic (r)	-0.56 (0.059)	–	-0.16 (0.63)	–
Block 9				
Elliptic (k)	**-0.95 ($<$1e-04)**	–	**-0.93 (0.00024)**	–
Elliptic (period)	$+0.17$ (0.61)	$+0.05$ (0.88)	$+0.49$ (0.13)	-0.07 (0.84)
Iterative (decay)	$+0.36$ (0.43)	$+0.36$ (0.43)	**$+0.79$ (0.036)**	**$+0.79$ (0.036)**
Iterative (lag)	**-0.77 (0.0092)**	-0.44 (0.2)	**-0.75 (0.013)**	-0.27 (0.45)
Logistic (r)	-0.20 (0.53)	-0.34 (0.27)	-0.26 (0.43)	-0.19 (0.58)
Block 10				
Elliptic (k)	**-0.93 (0.00024)**	–	**-0.90 (0.00094)**	–
Elliptic (period)	$+0.35$ (0.28)	-0.13 (0.7)	$+0.53$ (0.096)	-0.19 (0.58)
Iterative (decay)	$+0.18$ (0.7)	$+0.18$ (0.7)	**$+0.82$ (0.023)**	**$+0.82$ (0.023)**
Iterative (lag)	**$+0.78$ (0.0075)**	**$+0.71$ (0.022)**	**$+0.85$ (0.0016)**	**$+0.82$ (0.0038)**
Logistic (r)	-0.27 (0.4)	-0.34 (0.27)	$+0.14$ (0.69)	-0.17 (0.61)
Block 11				
Elliptic (k)	$+0.15$ (0.7)	–	$+0.40$ (0.29)	–
Elliptic (period)	$+0.39$ (0.23)	$+0.01$ (0.97)	$+0.45$ (0.16)	-0.12 (0.73)
Iterative (decay)	**$+0.82$ (0.023)**	-0.71 (0.071)	**$+1.00$ ($<$1e-04)**	-0.46 (0.29)
Iterative (lag)	$+0.61$ (0.06)	$+0.62$ (0.054)	**$+0.76$ (0.011)**	**$+0.64$ (0.048)**
Logistic (r)	$+0.09$ (0.78)	**-0.75 (0.0051)**	$+0.28$ (0.4)	**-0.84 (0.0011)**
Block 12				
Elliptic (k)	$+0.32$ (0.41)	–	$+0.25$ (0.52)	–
Elliptic (period)	$+0.09$ (0.79)	$+0.10$ (0.76)	$+0.54$ (0.089)	-0.04 (0.92)
Iterative (decay)	$+0.61$ (0.15)	$+0.61$ (0.15)	**$+0.79$ (0.036)**	**$+0.79$ (0.036)**
Iterative (lag)	**$+0.73$ (0.016)**	$+0.62$ (0.054)	**$+0.81$ (0.0049)**	$+0.62$ (0.054)
Logistic (r)	-0.20 (0.54)	$+0.11$ (0.74)	-0.03 (0.94)	-0.17 (0.62)

D MOIRAI-1 (PATCH EMBEDDING)

Compared to Chronos's token embedding, Moirai-1 uses patch embedding with patch sizes $(8, 16, 32, 64, 128)$. The number of tokens used to represent a time series is *(sequence length/patch size)*. We limit our time series data to the Jacobi elliptic functions and the cosine functions. We fix the number of tokens for each time series to 50 and generate for each patch size, a sequence with *context length = number of tokens * patch size*. Figures 33 and 34 show the forecasts made with each patch size candidate and the forecast MAE. We note that Moirai-1 allows longer prediction length.

We continue our analysis by selecting the model with the best patch size (best MAE) for each time series. Moirai-1 creates a mask over the tokens to be predicted, therefore we extract the submatrix corresponding to the context-to-context attention and normalize each row. Figures 35 and 36 illustrate the average attention matrices per encoder block with these models. We recognize that some canonical structures also emerge, and the diagonal bands are blurrier than in Chronos. Locality and periodicity profiles, layer pruning, and Spearman correlation to forecast error are shown in figures 37, 38, and table 7. Ablation results are similar to Chronos where adding the later blocks typically reduces forecast error. Compared to Chronos, Moirai-1 does not appear to leverage periodicity information for forecasts. One explanation is that patch embedding destroys frequency information.

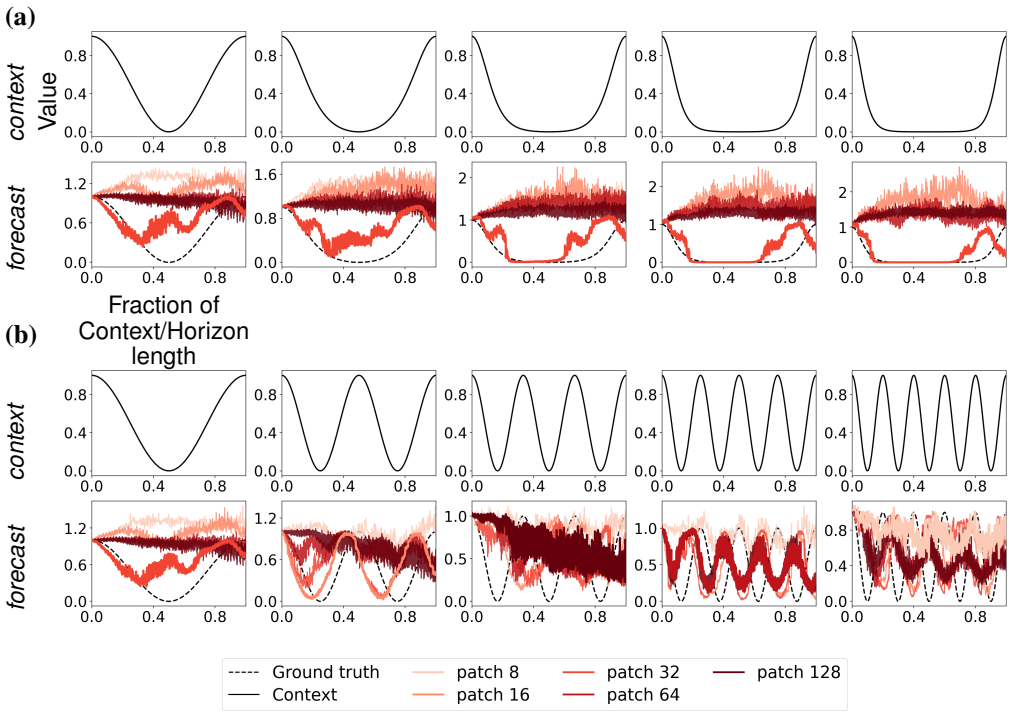

Figure 33: Moirai-1 forecasts of **(a)** Jacobi elliptic functions and **(b)** cosine functions when using different patch sizes. The top row shows the input given to the model while the bottom row shows the resulting forecast.

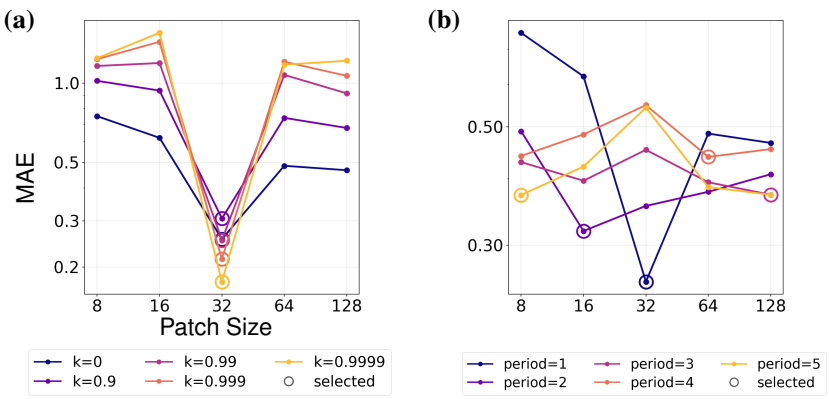

Figure 34: Mean absolute error of Moirai-1 forecasts made with each patch size.

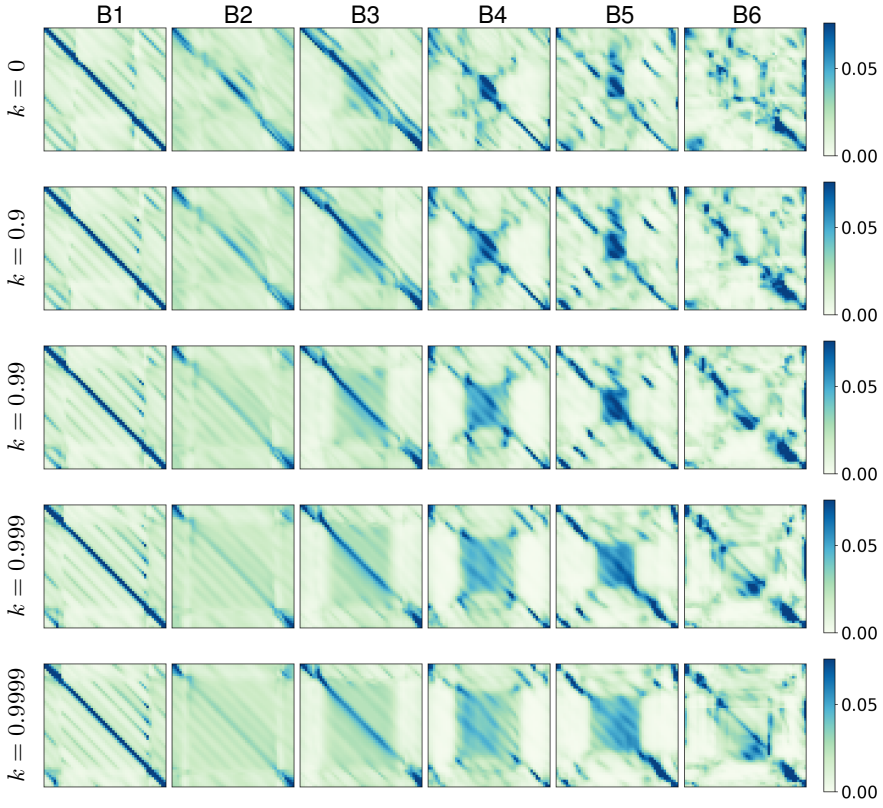

Figure 35: Average attention matrices across Moirai-1 encoder blocks for Jacobi elliptic functions with varying modulus k.

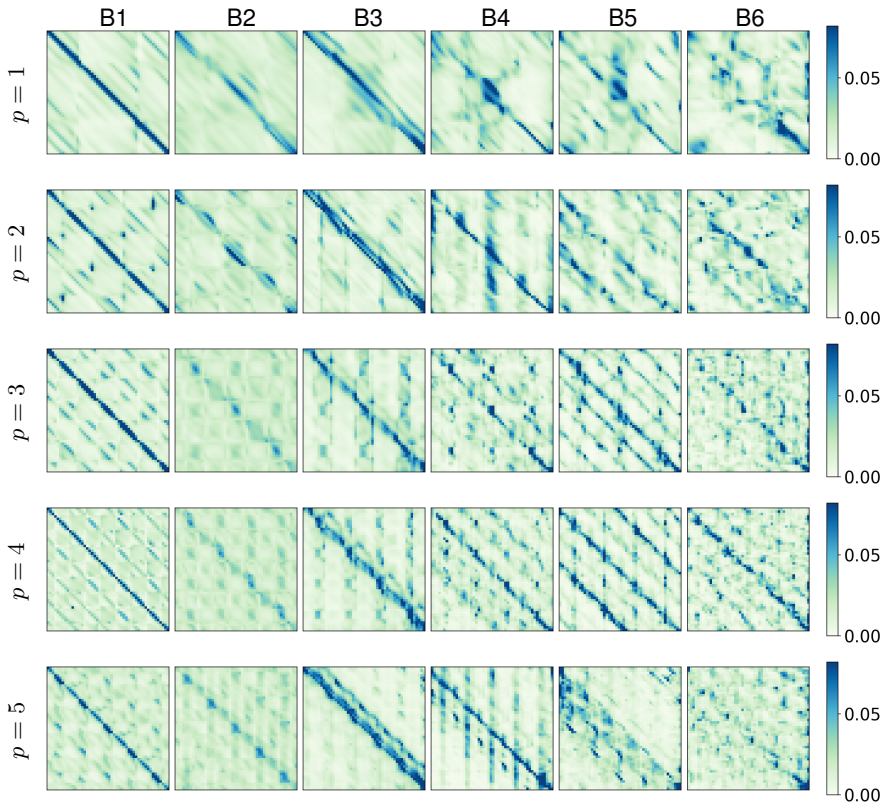

Figure 36: Average attention matrices across Moirai-1 encoder blocks for cosine functions with varying number of cycles p.

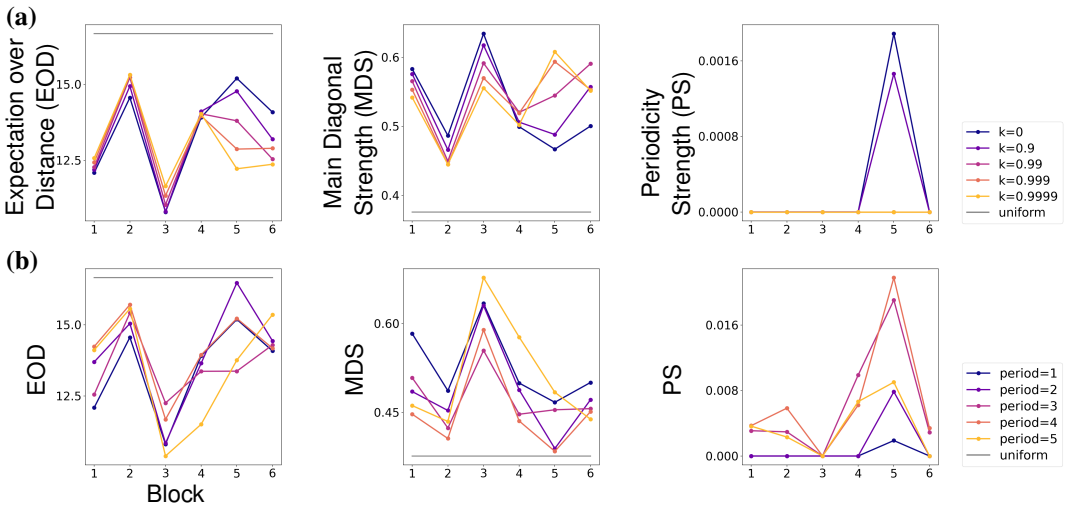

Figure 37: Locality and periodicity profiles on Moirai-1 for **(a)** Jacobi-elliptic functions with varying modulus k and **(b)** cosine functions with varying number of cycles p.

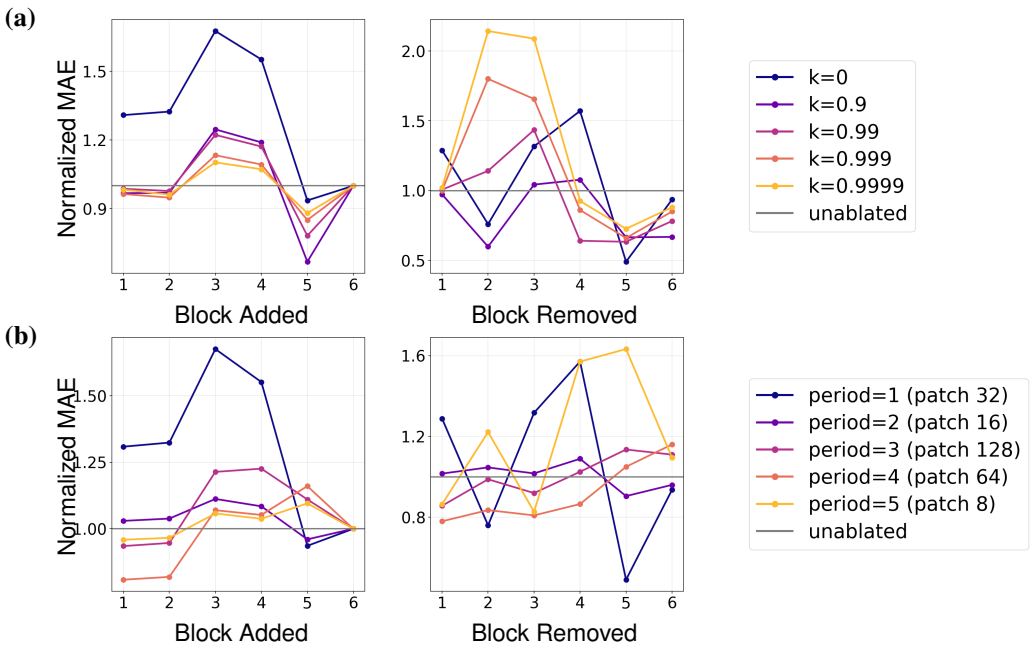

Figure 38: Mean absolute error (MAE) from two ablation studies on Moirai-1 for **(a)** Jacobi elliptic functions and **(b)** cosine functions. On the left, we measure MAE as blocks are sequentially added. On the right, we measure MAE after a block has been removed. Normalized MAE is computed by normalizing MAE with respect to the error when all blocks are intact.

Table 7: Spearman ρ (p-value) between the attention metric read at Moirai-1 encoder block i and the forecast error of the model with the encoder truncated after block i, for every depth. **Bold** marks $p < 0.05$.

Series family	MAE vs. MDS	MAE vs. Periodicity	Skill vs. MDS	Skill vs. Periodicity
Block 1				
Elliptic (k)	**+1.00 (<1e-04)**	–	**+1.00 (<1e-04)**	–
Elliptic (p)	−0.60 (0.28)	+0.87 (0.054)	−0.60 (0.28)	+0.87 (0.054)
Block 2				
Elliptic (k)	**+0.90 (0.037)**	–	**+0.90 (0.037)**	–
Elliptic (p)	−0.60 (0.28)	+0.67 (0.22)	−0.60 (0.28)	+0.67 (0.22)
Block 3				
Elliptic (k)	**+1.00 (<1e-04)**	–	**+1.00 (<1e-04)**	–
Elliptic (p)	−0.60 (0.28)	–	−0.60 (0.28)	–
Block 4				
Elliptic (k)	−0.30 (0.62)	–	−0.30 (0.62)	–
Elliptic (p)	−0.70 (0.19)	+0.41 (0.49)	−0.70 (0.19)	+0.41 (0.49)
Block 5				
Elliptic (k)	**−1.00 (<1e-04)**	**+0.89 (0.041)**	**−1.00 (<1e-04)**	**+0.89 (0.041)**
Elliptic (p)	−0.50 (0.39)	**+1.00 (<1e-04)**	−0.50 (0.39)	**+1.00 (<1e-04)**
Block 6				
Elliptic (k)	+0.20 (0.75)	–	−0.10 (0.87)	–
Elliptic (p)	**−0.90 (0.037)**	+0.67 (0.22)	**−0.90 (0.037)**	+0.67 (0.22)

E MOIRAI-2 (DECODER-ONLY)

Compared to Moirai-1, Moirai-2 uses a fixed patch size and a decoder-only architecture. For Moirai-2 small, which we examine here, the patch size is 16. We proceed with our analysis by fixing the number of tokens and adjusting the context length as *number of tokens * patch size*. Moirai-2 predicts 4 tokens at a time, rolling predicted tokens into the context. For this reason, we forecast 4 tokens or 64 time steps for our input sequences. As a decoder-only model, the attention matrices are causal (row i is distributed over columns 1 to i), therefore we have to adjust our metrics to a lower-triangular attention matrix. For expectation over distance and main diagonal strength, the computation is unchanged. However, we consider a different baseline since uniformity now differs across rows. As an alternative to the uniform attention matrix, we compare locality to a shuffled baseline where for each row, values are shuffled across supporting columns i.e. values at row i are shuffled across columns 1 to i. We then compute

$$\text{EOD Ratio}(X) = \frac{\text{EOD}(X)}{\text{EOD}(X_{\text{baseline}})} \tag{12}$$

$$\text{MDS Ratio}(X) = \frac{\text{MDS}(X)}{\text{MDS}(X_{\text{baseline}})} \tag{13}$$

We shuffle the attention matrix and recompute locality of the shuffled matrix a number of iterations before taking the mean. For periodicity, we only need to modify the band strength profile to track one-sided diagonal offsets, i.e. for each diagonal offset d, we compute the mean of $-d$ offset diagonal to obtain sequence β.

Again, we restrict the analysis to Jacobi-elliptics and cosine functions. Forecasts are shown in figure 39 and the average causal attention matrices are shown in figures 40 and 41. Similar to Moirai-1, periodicity in the decoder blocks, while present, tends to be blurry. Figure 42 compares locality and periodicity for different parameters while 43 demonstrates a steady decrease in MAE as blocks are added. We observe that there is strong Spearman correlation between locality and forecast error in blocks 5 and 6, and generally weak correlation with periodicity (table 8).

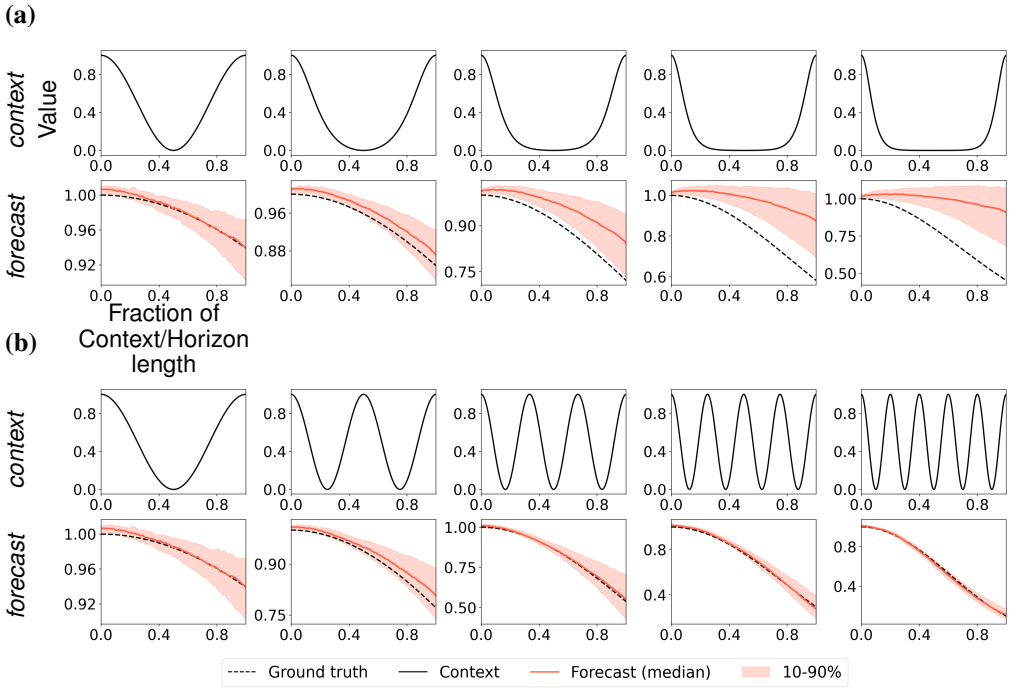

Figure 39: Moirai-2 forecasts of **(a)** Jacobi elliptic functions and **(b)** cosine functions. The top row shows the input given to the model while the bottom row shows the resulting forecast.

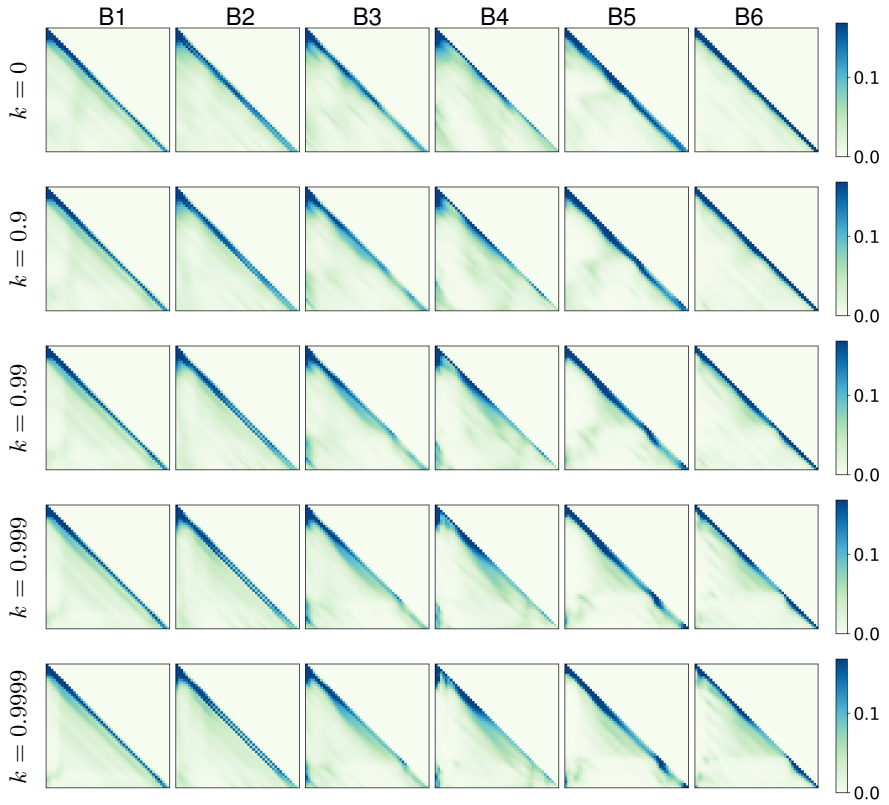

Figure 40: Average causal attention matrices across Moirai-2 decoder blocks for Jacobi elliptic functions with varying modulus k.

Table 8: Spearman ρ (p-value) between the attention metric read at Moirai-2 decoder block i and the forecast error of the model with truncation after block i, for every depth. **Bold** marks $p < 0.05$.

Series family	MAE vs. MDS obs/null	MAE vs. Periodicity	Skill vs. MDS obs/null	Skill vs. Periodicity
Block 1				
Elliptic (k)	−0.70 (0.19)	–	−0.70 (0.19)	–
Elliptic (p)	−0.70 (0.19)	−0.45 (0.45)	−0.70 (0.19)	−0.45 (0.45)
Block 2				
Elliptic (k)	−0.70 (0.19)	–	−0.70 (0.19)	–
Elliptic (p)	+0.60 (0.28)	+0.22 (0.72)	+0.60 (0.28)	+0.22 (0.72)
Block 3				
Elliptic (k)	+0.30 (0.62)	–	+0.30 (0.62)	–
Elliptic (p)	**+1.00 (<1e-04)**	+0.10 (0.87)	**+1.00 (<1e-04)**	+0.10 (0.87)
Block 4				
Elliptic (k)	**+1.00 (<1e-04)**	–	**+1.00 (<1e-04)**	–
Elliptic (p)	−0.60 (0.28)	−0.20 (0.75)	−0.60 (0.28)	−0.20 (0.75)
Block 5				
Elliptic (k)	**−0.90 (0.037)**	–	**−0.90 (0.037)**	–
Elliptic (p)	+0.10 (0.87)	+0.00 (1)	+0.10 (0.87)	+0.00 (1)
Block 6				
Elliptic (k)	**−1.00 (<1e-04)**	–	**−1.00 (<1e-04)**	–
Elliptic (p)	−0.30 (0.62)	+0.22 (0.72)	−0.30 (0.62)	+0.22 (0.72)

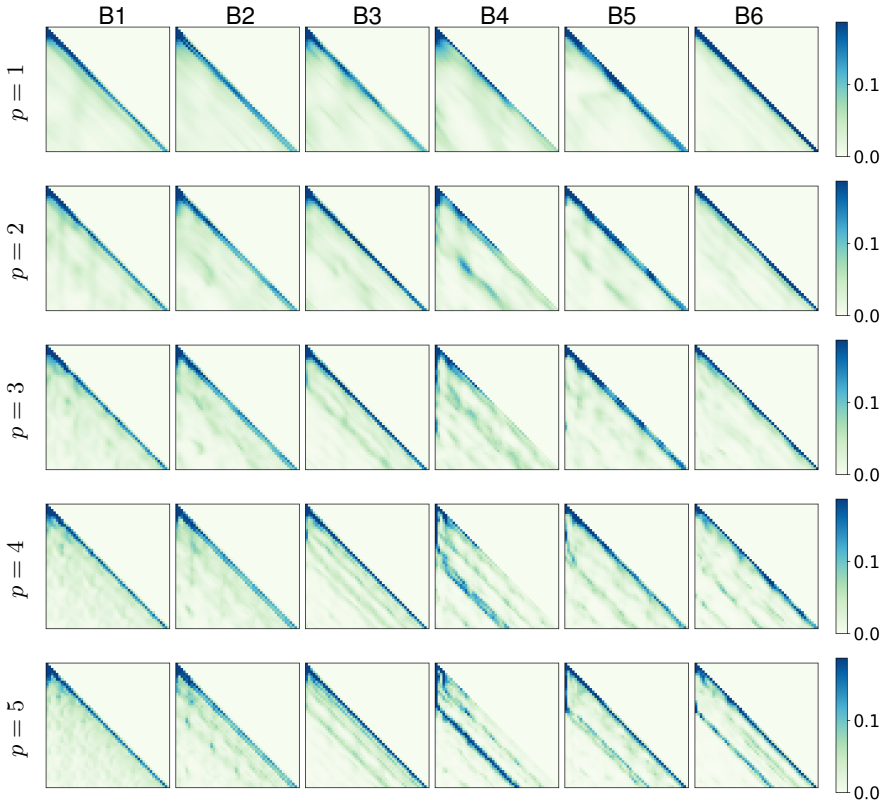

Figure 41: Average causal attention matrices across Moirai-2 decoder blocks for cosine functions with varying number of cycles p.

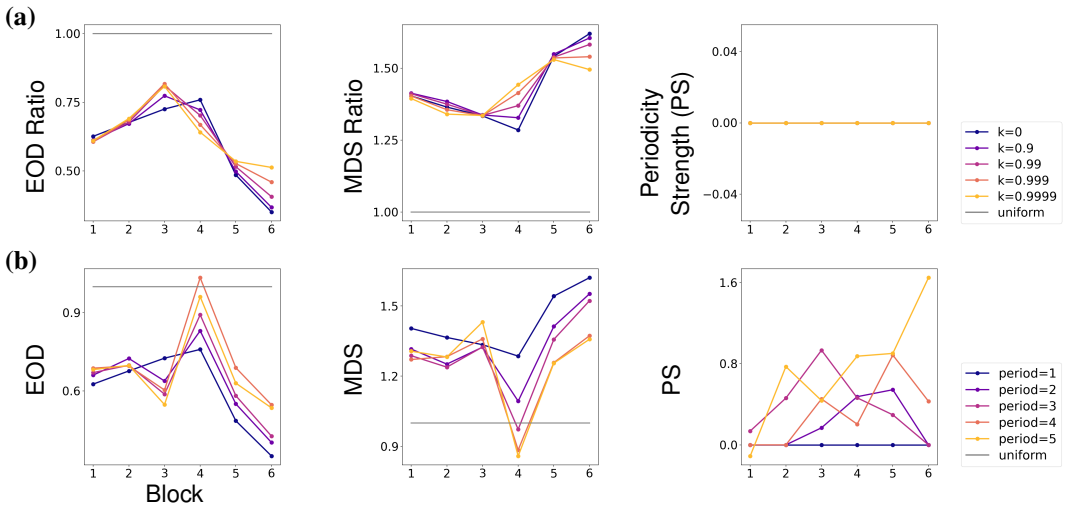

Figure 42: Locality and periodicity profiles on Moirai-2 for **(a)** Jacobi-elliptic functions with varying modulus k and **(b)** cosine functions with varying number of cycles p.

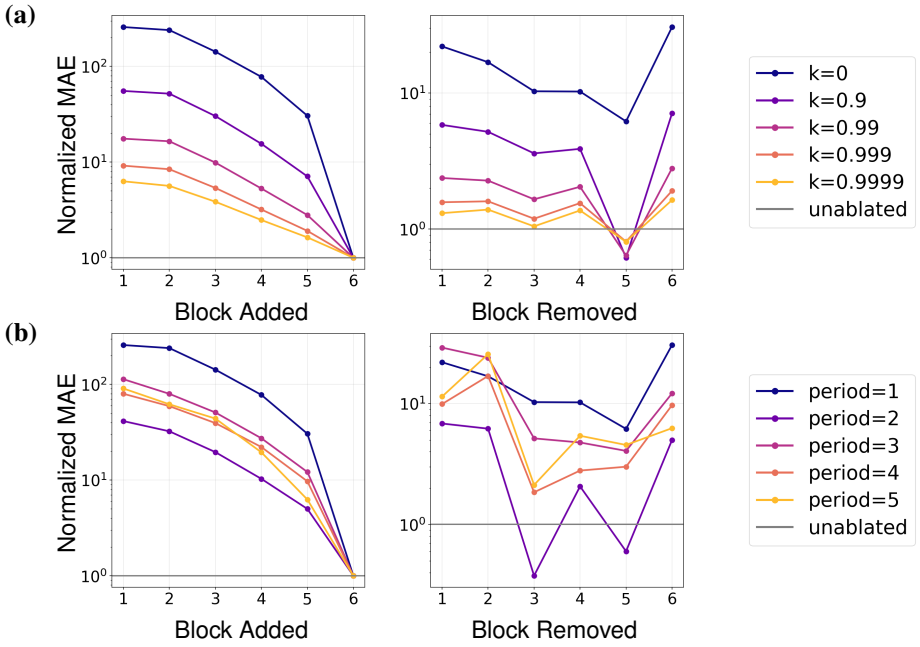

Figure 43: Mean absolute error (MAE) from two ablation studies on Moirai-2 for **(a)** Jacobi elliptic functions and **(b)** cosine functions. On the left, we measure MAE as blocks are sequentially added. On the right, we measure MAE after a block has been removed. Normalized MAE is computed by normalizing MAE with respect to the error when all blocks are intact.