Interpretability for Time Series Transformers using A Concept Bottleneck Framework

Angela van Sprang[1], Erman Acar[1], Willem Zuidema[1]
[1]University of Amsterdam
{a.v.vansprang,e.acar,w.h.zuidema@}@uva.nl

Abstract

Mechanistic interpretability focuses on *reverse engineering* the internal mechanisms learned by neural networks. We extend our focus and propose to mechanistically *forward engineer* using our framework based on Concept Bottleneck Models. In the context of long-term time series forecasting, we modify the training objective to encourage a model to develop representations which are similar to predefined, interpretable concepts using Centered Kernel Alignment. This steers the bottleneck components to learn the predefined concepts, while allowing other components to learn other, undefined concepts. We apply the framework to the Vanilla Transformer, Autoformer and FEDformer, and present an in-depth analysis on synthetic data and on a variety of benchmark datasets. We find that the model performance remains mostly unaffected, while the model shows much improved interpretability. Additionally, we verify the interpretation of the bottleneck components with an intervention experiment using activation patching.

1 Introduction

Transformers show great success for various types of sequential data, including language [Devlin, 2018, Brown, 2020], images [Dosovitskiy et al., 2021, Liu et al., 2021], and speech [Baevski et al., 2020]. Their ability to capture long-term dependencies has triggered substantial interest in applying them to time-series, which are naturally sequential, and in particular to the challenging task of long-term time series forecasting. Transformer-based architectures, indeed, often show superior performance on this task [Zhou et al., 2021, 2022, Wu et al., 2021, Ni et al., 2023, Chen et al., 2024], for an overview we refer to Wen et al. [2023].

However, due to their deep and complex architecture, transformers are difficult to interpret, which is especially important in high-stakes domains such as finance and energy demand prediction. There is a large body of work in the field of explainable AI to interpret neural networks [Bereska and Gavves, 2024], or increase their interpretability, including the approach of Concept Bottleneck Models (CBMs; Koh et al., 2020). This approach relies on the idea of constraining the model such that it first predicts human-interpretable concepts, and then uses only these concepts to make the final prediction. CBMs and their variants have become popular in various fields, especially in computer vision, but are so far unexplored in the context of time series forecasting.

In this paper, we propose a training framework to make any time series transformer into a Concept Bottleneck Model using time-series specific, yet domain-agnostic concepts, as shown in Figure 1. A key aspect of our training framework is to leave the model's architecture intact, while encouraging the learned representations to be similar - but not identical - to the interpretable concepts. We measure similarity with Centered Kernel Alignment (CKA; Kornblith et al., 2019) and include it in the loss function. The first concept is a simple, linear surrogate model and the second is time information (e.g. hour-of-day). Note that we propose a *global* interpretability method, which improves identifying

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

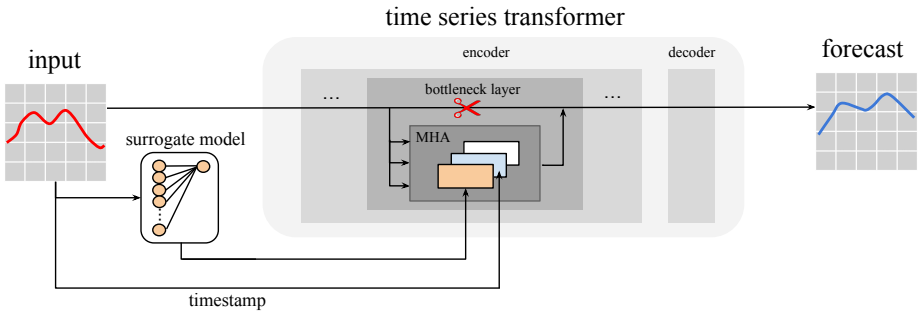


Figure 1: Overview of the concept bottleneck framework. The bottleneck is one encoder layer which is trained to be similar to pre-defined, interpretable concepts. The residual stream around the bottleneck is removed, such that all information passes through the bottleneck.

and localizing high-level concepts in the model's internal mechanisms, and is not comparable to local post-hoc interpretability methods such as SHAP, LIME, or attention-based visualizations which explain individual predictions.

We apply our concept bottleneck framework to three types of models: Vanilla Transformer [Vaswani et al., 2017], Autoformer [Wu et al., 2021] and FEDformer [Zhou et al., 2022]. Across extensive experiments on seven datasets, we show that our setup results in models that are more interpretable while the overall performance remains largely unaffected – in many cases surpassing results from the original Autoformer paper. Furthermore, we explicitly test the faithfulness of the obtained interpretations with an intervention study using activation patching.

Our contributions are summarized as follows:

1. We propose a novel training framework to increase the interpretability of transformers for time series.
2. We demonstrate the feasibility of applying this framework to time-series transformers by conducting extensive experiments on three types of transformers and seven datasets, and identify interpretable concepts in each of these transformers.
3. We assess the faithfulness of the interpretability analysis by performing an activation patching experiment, and obtain evidence that the identified components (in the concept bottleneck) indeed have the hypothesized unique and causal role in the predictions of the target model.

2 Background and Related Work

This paper combines and builds upon foundational works from different fields, including CBMs, knowledge transfer with CKA and time series transformers. CBMs have been applied to time series before [Ferfoglia et al., 2024], but not with the same interpretable concepts. Likewise, the similarity index CKA has been used before to transfer knowledge between models [Tian et al., 2023], yet, to the best of our knowledge, it has not been used to construct a CBM. This makes our work a unique contribution at the intersection of (mechanistic) interpretability, concept learning, and time series forecasting.

2.1 Concept Bottleneck Models

Concept Bottleneck Models (CBMs; Koh et al., 2020) have emerged as promising interpretable models [Poeta et al., 2023]. The concept bottlenecks constrain the model to first predict interpretable concepts, and then use only these concepts in the final downstream task. They are shown to be useful in multiple applications, such as model debugging and human intervention. The bottleneck allows for explaining which information the model is using and when it makes an error due to incorrect concept predictions.

One of the shortcomings of standard CBMs is that concept annotations are needed during training to learn the bottleneck, and concept labels do not necessarily contain all information needed to accurately perform the downstream task, and can therefore decrease the task accuracy [Mahinpei et al., 2021]. Therefore, Zarlenga et al. [2022] propose Concept Embedding Models, where concepts are represented as vectors, such that richer and more meaningful concept semantics can be captured.

CBMs and their variants are usually applied to the field of computer vision, and less frequently to natural language [Tan et al., 2024], graphs [Barbiero et al., 2023] or tabular data [Zarlenga et al., 2022]. In principle, the methodology can be applied to time series as well, but defining high-level, meaningful concepts is challenging. Ferfoglia et al. [2024] use Signal Temporal Logic (STL) formulas as concept embeddings for time series to convert them into natural language, and use these concepts as bottleneck for anomaly detection.

2.2 Knowledge Transfer with Centered Kernel Alignment

Inspired by neuroscience, CKA measures the similarity between different representations from neural networks [Kornblith et al., 2019]. By factoring out differences in scaling or orthogonal transformations, CKA captures intuitive notions of similarity between representations. To obtain the score, firstly, the similarity between every pair of examples in each representation separately is measured using a pre-defined kernel, and then the obtained similarity structures are compared. We use a linear kernel, which makes the CKA score defined as follows for representations X and Y:

$$\mathrm{CKA}(X, Y) = \frac{\left\|Y^\top X\right\|_F^2}{\left\|X^\top X\right\|_F \left\|Y^\top Y\right\|_F}. \tag{1}$$

The CKA score can be used to transfer knowledge between different models when included in the loss function [Tian et al., 2023]. In this work, the authors study knowledge distillation between a teacher and student model, and incorporate CKA into the loss function to transfer feature representation knowledge from the pretrained model to the incremental learning model [Parisi et al., 2019].

2.3 Time Series Transformers

Time series transformers for long-term time series forecasting, such as the Autoformer and FEDformer, obtain two types of input: (1) *data values* $\boldsymbol{X} \in \mathbb{R}^{I \times d}$, and (2) *timestamps* $\boldsymbol{T} \in \mathbb{R}^{I \times 4}$. More specifically, they can be regarded as a function $f : \mathbb{R}^{I \times d} \times \mathbb{R}^{I \times 4} \times \mathbb{R}^{O \times 4} \to \mathbb{R}^{O \times d}$, where I is the number of input time steps, O is the number of future time steps, and d is the number of variables in the time series. The additional four dimensions of timestamps $\boldsymbol{T}$ represent four time features, namely *hour-of-day, day-of-week, day-of-month*, and *day-of-year*. The future timestamps are also provided, for which the model should forecast the future data values. Note that we explicitly introduce a notation for the timestamps to later define the CKA scores and the intervention.

3 Method

We propose a training framework to make any transformer model interpretable by including a bottleneck based on knowledge transfer with CKA [Kornblith et al., 2019], as shown in Figure 2. The main idea is that we assign one of the encoder layers to be the *concept bottleneck*; representations in the bottleneck are subject to a soft constraint of being as similar as possible to predefined interpretable concepts. To this end, we calculate CKA scores with the interpretable concepts, and include these scores in the loss function.

3.1 Loss Function

The loss function should encourage the model to represent the interpretable concepts in the bottleneck layer. Therefore, we add a term $\mathcal{L}_{CKA}$ based on the CKA scores of the bottleneck and the interpretable concepts (Eq. 3). In particular, low similarity between the bottleneck and the interpretable concepts results in a higher value for $\mathcal{L}_{CKA}$. The total loss function $\mathcal{L}_{Total}$ (Eq. 2), then, is a weighted average

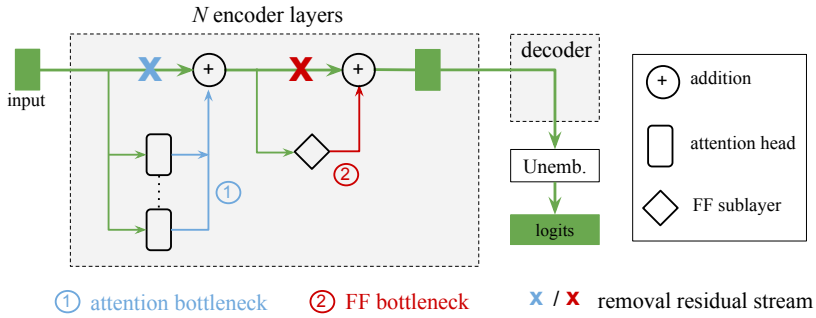


Figure 2: Architecture of a transformer with a concept bottleneck in the attention mechanism (blue) or the FF network (red). Note that the residual connection is removed at the location of the bottleneck (and the residual stream thus interrupted). Visualisation inspired by Rai et al., 2024.

of the Mean Squared Error (MSE) loss $\mathcal{L}_{MSE}$ and the CKA loss $\mathcal{L}_{CKA}$:

$$\mathcal{L}_{Total} = (1-\alpha)\,\mathcal{L}_{MSE} + \alpha\,\mathcal{L}_{CKA}, \tag{2}$$

$$\mathcal{L}_{CKA} = 1 - \frac{1}{c}\sum_{i=1}^{c} CKA_i, \tag{3}$$

where α is a hyperparameter, c is the number of concepts, and $CKA_i \in [0,1]$ is the CKA score (using a linear kernel) between the model's representation and concept i (see Section 3.2).

3.2 Interpretable Concepts in the Bottleneck

In this section, we describe how to calculate the CKA score to measure the presence of a concept. We refer to Appendix B for a more detailed description of the concept bottleneck framework.

Location bottleneck. We assign one encoder layer to be the bottleneck layer, because the encoder focuses on modelling seasonal information. Within the bottleneck layer, the latent representations can be taken from two different types of blocks: the attention block ($\tau = Att$) and the feed-forward block ($\tau = FF$). These two options are illustrated in Figure 2. We assign c interpretable concepts over the latent representations, with the goal of teaching the corresponding model component to represent the pre-defined interpretable concepts.

Since the attention block is multi-headed, different heads naturally form the components of the attention bottleneck. Moreover, the components need to be divided between the heads, which would be convenient when the number of heads is a multiple of the total number of concepts to maintain a uniform concept per head ratio. For the feed-forward bottleneck, we define the components to be slices from its output, such that stacking the components results in the original output.

Interpretable concepts. For the real-world time series, we use two domain-agnostic interpretable concepts which can be used for forecasting, namely: (1) a simple, human-interpretable surrogate forecasting model, (2) the input timestamps recorded with the time series. Note that each model token (time step) should map to each concept to calculate the CKA score.

1. We use a simple autoregressive model (AR) as a surrogate model, which predicts the next future value as a linear combination of its past values. This model is transparent, and the attribution of each input feature to the output can be simply interpreted by its weight. This concept can also be regarded as a baseline for the forecasting performance. The model is fit to the same training data as the transformer (with its order being the length of the input). We use its activations to calculate the CKA score.
2. We use the hour-of-day feature from the timestamps $\boldsymbol{T}$ as interpretable time concept, denoted by $\boldsymbol{T}_{hourofday}$. This provides the bottleneck with a simplified notion of time.

Removal of residual connection. Any transformer layer contains residual connections around the attention and feed-forward blocks. To ensure that all information passes through the bottleneck, we

remove the residual connection around the bottleneck, potentially at the cost of a loss in performance. Otherwise, any concept, including the interpretable concepts, can be passed through the residual connection and compromise the bottleneck.

In the scenario that the number of components is equal to the number of interpretable concepts ($c = 2$), the construction of the bottleneck limits learning domain-specific features from the data, other than the interpretable concepts. Therefore, we perform experiments where we allow an extra component in the bottleneck to not learn any pre-defined concept ($c = 3$). In other words, the extra component serves as a *side-channel* or *free component*, on which no CKA loss is calculated. The free component may partly restore the information lost by removing the residual connection, but with the advantage that we can monitor which information goes through it, and even visualize it (as in Section 4.3.2).

3.3 Implementation details.

In our experiments, we use transformer models with three encoder layers, of which the bottleneck layer is the second layer. Similar to the original Autoformer paper, we use one decoder layer, employ the Adam optimizer [Kingma and Ba, 2017] with an initial learning rate of 10^{-4}, and use a batch size of 32. The training process is early stopped within 25 epochs. All experiments are repeated five times on different seeds, using hyperparameter $\alpha = 0.3$. Each model is trained on 1 Nvidia GeForce GTX 1080 Ti with 30 GB for approximately 30 minutes.

4 Experiments

We evaluate our framework on three models and seven datasets, including synthetic and real-world data. The six real-world benchmarks consider the domains of energy, traffic, economics, weather, and disease, similar to Wu et al. [2021]. These datasets are multivariate, and the task is to predict the future values of all variates. For example, the electricity dataset consists of hourly measurements of the electricity consumption of 321 customers from 2012 to 2014. For more information on the datasets, we refer the reader to Appendix A. We apply the experiments to the Vanilla Transformer, Autoformer and FEDformer. First, we train a simple AR model on the same data, so that its outputs can be used to align the representations of the bottleneck. Then, we train the transformers with and without bottleneck, using different configurations for the bottleneck.

4.1 Synthetic Data

To show the general applicability of the bottleneck framework, we first train an Autoformer on a synthetic time series. In particular, we generate the dataset as the sum of different sines using the function f_{Total} with time t as follows:

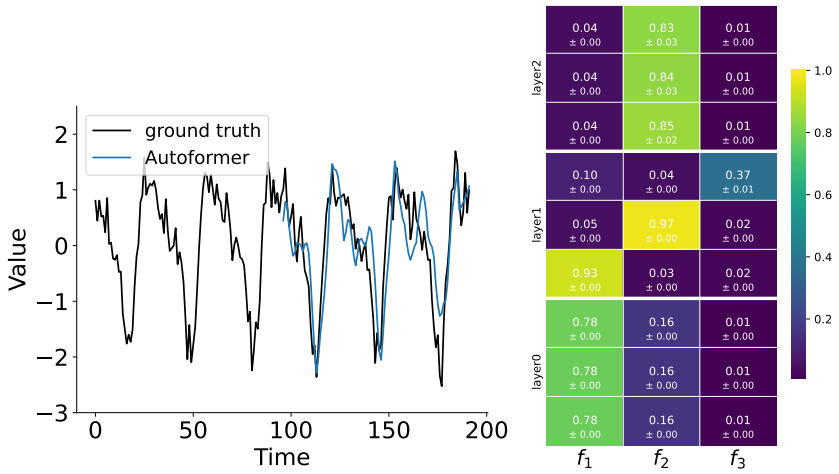
$$f_{Total}(t) = f_1(t) + f_2(t) + f_3(t),$$

where:

$$\begin{aligned} f_1(t) &= \sin(2\pi t), \\ f_2(t) &= \frac{1}{2}\sin(4\pi t + \frac{\pi}{4}), \\ f_3(t) &= \frac{1}{4}\sin(6\pi t + \frac{\pi}{2}) + \epsilon_t. \end{aligned}$$

Note that all functions f_1, f_2 and f_3 follow a periodic structure, and f_3 contains random noise ϵ from a normal distribution with standard deviation of 0.2.

Each concept in the bottleneck is defined as one of the underlying functions (i.e., f_1, f_2 or f_3), for which the ground-truth is known by construction. For hyperparameter $\alpha = 0.8$ (see Section 3.1), we find that the model is able to forecast well, while achieving very high similarity scores. That is, the model obtains a Mean Squared Error (MSE) of 0.36 ± 0.17 and Mean Absolute Error (MAE) of 0.46 ± 0.12 on 5 different seeds. See Figure 3 for a sample forecast on the test data and the CKA scores of the model's representations with the concept representations. The heads in the bottleneck `layer1` show high similarity for their respective concepts, e.g. a score of 0.93 for the head trained on f_1 (recall that CKA scores range from 0 for totally dissimilar to 1.0 for identical, although potentially scaled and rotated). We refer to Appendix H for more results on the synthetic dataset.



(a) Forecast (b) CKA scores

Figure 3: Forecast and CKA scores of the attention bottleneck Autoformer on synthetic data, where the three heads of each layer (vertically) are compared with the three concept vectors (horizontally).

4.2 Real-world data

Table 1 shows the performance of the Autoformer with our bottleneck on the benchmark datasets, compared to the AR surrogate model (i.e. the first interpretable concept) and Wu et al. [2021] (i.e. the original Autoformer model). Note that the bottleneck models are trained with a free component, i.e., $c = 3$, and the original Autoformer is of a different size (two encoder layers with eight heads per layer). Visualizations of the forecasts from these models are shown in Appendix C.

Table 1: Error scores of different Autoformer models. For both metrics, it holds that a lower score indicates a better performance, where the best results are **bold**, and the second-best are underlined.

	Att bottleneck		**FF bottleneck**		**No bottleneck**		**AR**		**Wu et al.**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	0.231	0.338	0.207	0.320	0.280	0.368	0.497	0.522	**0.201**	**0.317**
Traffic	0.642	0.393	**0.393**	**0.377**	0.619	0.387	0.420	0.494	0.613	0.388
Weather	0.290	0.354	0.271	0.341	0.269	0.344	**0.006**	**0.062**	0.266	0.336
Illness	3.586	1.313	3.661	1.322	3.405	1.295	**1.027**	**0.820**	3.483	1.287
Exchange rate	0.195	0.323	0.155	0.290	0.152	0.283	**0.082**	**0.230**	0.197	0.323
ETT	0.177	0.282	0.174	0.280	0.155	0.265	**0.034**	**0.117**	0.255	0.339

We find that including a bottleneck (either **Att bottleneck** or **FF bottleneck**) outperforms **Wu et al.** for three datasets (traffic, exchange rate and ETT), and stays within 5% of the MSE and MAE for the other three datasets. Surprisingly, the surrogate AR model outperforms the other models for most datasets w.r.t. both MSE and MAE, even though this model is very simple.[1] More detailed results are presented in Appendix D and E, where the first includes the results for bottlenecks without free component (including the standard deviation for different seeds), and the latter includes a sensitivity analysis to hyperparameter α.

Similar to the Autoformer, the Vanilla Transformer and FEDformer with a bottleneck outperform models without bottleneck for some datasets, see Appendix F and G for a full analysis, respectively.

[1]Note that the phenomenon that simple models sometimes beat time series transformers [Zeng et al., 2022] has been observed before. There has been a vivid discussion about the relevance of these results, for instance here. These discussions are beyond the scope of our paper, which rather targets interpretability of time series transformers. For more information on the effect of AR as surrogate model, see Appendix I.

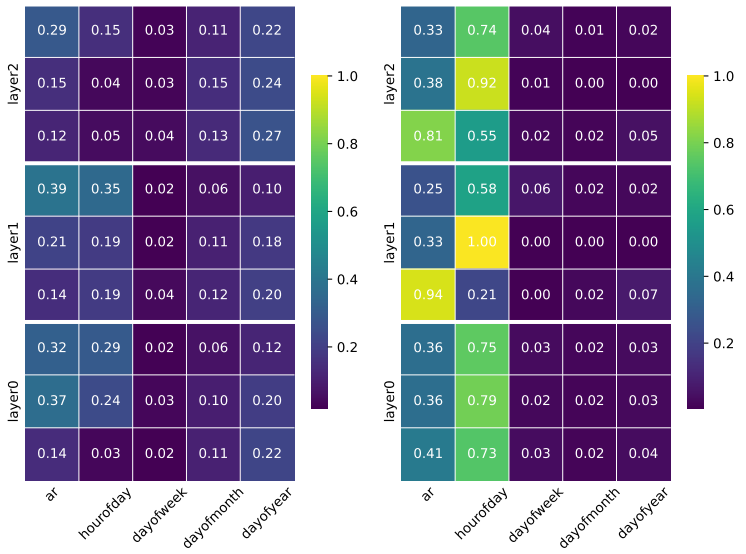
4.3 Interpretability Analysis

To demonstrate the impact of the bottleneck on model interpretability, we first conduct a CKA analysis on the bottleneck layer with the corresponding interpretable concepts, and then visually demonstrate how each component contributes to the final forecast.

4.3.1 CKA Analysis

To test the extent to which the bottleneck represents the interpretable concepts, we calculate the CKA scores of the model's representations with the concept representations. The scores of the feed-forward bottleneck on the electricity dataset are shown in Figure 4 (see Appendix E for more scores on the Autoformer). Note that the bottom, middle and upper layer of `layer1` correspond to the AR, hour-of-day, and free component of the bottleneck, respectively.

The scores show that the representations in the bottleneck layer are much more similar to the intended concepts than the representations from the model without bottleneck: 0.94 for the AR model, and 1.00 for the hour-of-day feature, whereas the model without bottleneck does not show high similarity to the interpretable concepts. This indicates that the training framework can encourage the components to form representations that are perfectly similar to the interpretable concepts. Additionally, note that the CKA scores of other layers than the bottleneck layer are also higher in Fig. 4b, which indicates that these other model components also learn to represent the interpretable concepts. This does not affect the interpretability of the bottleneck layer itself (Section 4.4).



(a) No bottleneck (b) Ours (FF bottleneck)

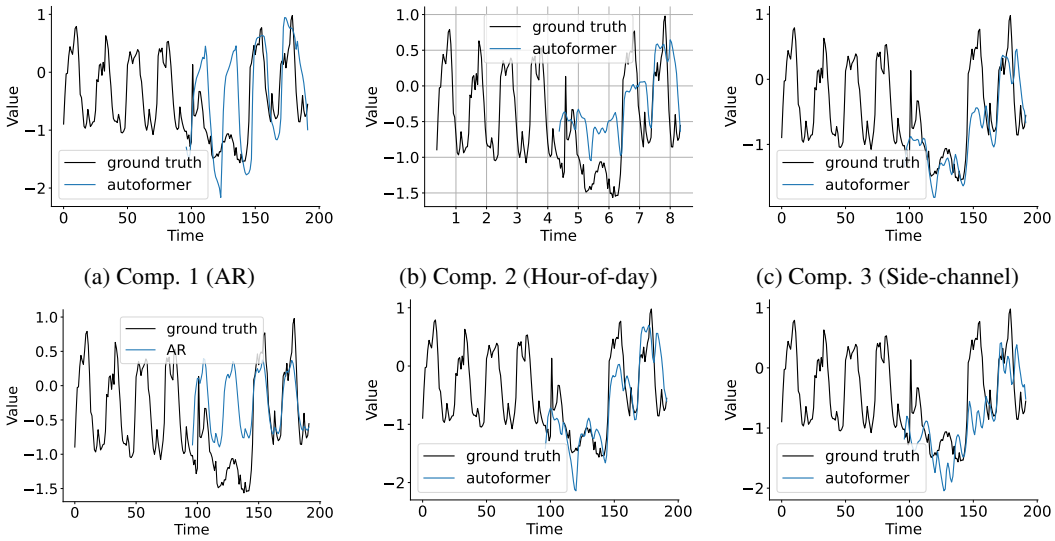
Figure 4: CKA scores on different concepts for the encoder of the Vanilla Transformer without bottleneck and with FF bottleneck. Both models contain three heads per layer. The first component of `layer1` (lower row) of the attention bottleneck is trained to be similar to AR, and the second component (middle row) to the hour-of-day concept. The scores are calculated on three batches of size 32 from the electricity test data. Recall that CKA is defined on a scale from 0 to 1, where 1 denotes perfect similarity.

4.3.2 Component Visualizations

Because the model components all read and write from the residual stream [Elhage et al., 2021], we can visualize the contributions to the final prediction of each component separately by applying the entire decoder to the component representations (Decoder Lens method, Langedijk et al. [2023]). This way, we obtain visualizations of the contributions of each component in the bottleneck, see Figure 5. We obtain the output from the full bottleneck by applying the decoder to the output of the

bottleneck (after performing layer normalization). The output from each component individually is obtained by masking the other components with zero (close to the mean).

From Figure 5a and 5b we see that the different bottleneck components are similar to the concepts they were trained on. In particular, the first component shows a forecast with correct periodicity and few irregularities, similar to the actual forecast from the AR model. Likewise, the second component shows a periodicity to the actual hour-of-day feature. The third component is not trained to be similar to an interpretable concept, and seems to pick up on the high-frequency patterns in the data, e.g., the low, second peak in the forecast. This observation is further strengthened by Figure 5f, which shows that the final forecast consists of many high-frequency patterns when using only the third component from the bottleneck. We find similar component visualizations on the Vanilla Transformer, see Appendix F.3.



(a) Comp. 1 (AR) (b) Comp. 2 (Hour-of-day) (c) Comp. 3 (Side-channel)

(d) AR (e) Full bottleneck (f) Final, using only comp. 3

Figure 5: Forecasts from individual bottleneck components by masking the other components with zero in 5a, 5b and 5c (FF bottleneck Autoformer on electricity data). The first half of the ground truth forms the input to the model. Note that the horizontal axes are the same across all figures, but Figure 5b contains a grid of days instead of numbered hours. Figure 5d shows the forecast made by the surrogate model AR; Figure 5e shows the forecast of the entire layer (i.e., all components together), and 5f shows the forecast of the final layer when only the third component is used in the bottleneck layer. Note the difference between Figures 5c and 5f, where we decode from the bottleneck and the final layer, respectively.

4.4 Intervention

The main benefit of interpreting trained models is gaining a deeper understanding and, possibly, more control of the model's behavior. This can be useful in the scenario of out-of-distribution data at inference time. If the data changes in features that can be interpreted in the model, it is feasible to intervene locally in these concepts to exclusively employ the model with data from its training distribution. Additionally, an intervention can be regarded as a causal interpretability test, where a successful intervention indicates a successful representation of the concept of interest.

To show such benefit of our framework, we perform activation patching (or causal tracing, Meng et al., 2023), where causal effects of hidden state activations are researched by evaluating the model on clean and corrupted inputs. We evaluate the trained model on data with shifted timestamps and compare it with performing an intervention on the shifted concept.

More specifically, we delay the input timestamps $\boldsymbol{T} \in \mathbb{R}^{I \times 4}$ with a fixed number of hours to obtain the shifted timestamps $\widetilde{\boldsymbol{T}}$, so that the learned patterns associated to the hour-of-day feature are misleading. We run the model on both types of timestamps, and perform an intervention in the

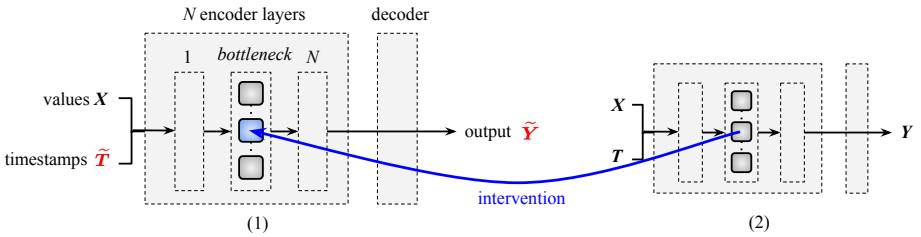
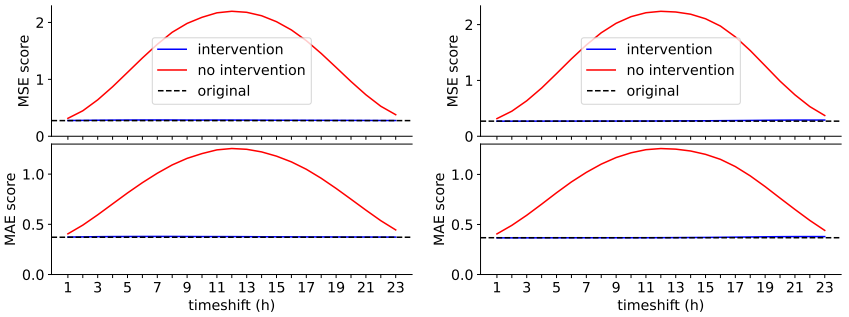


Figure 6: Intervention experiment, where we run the model on time-shifted input (1), but replace the activations of the hour-of-day component with those from a run (2) on unshifted timestamps $\boldsymbol{T}$.

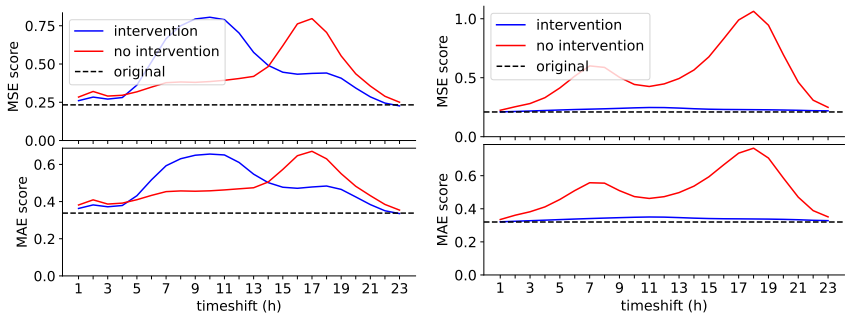
bottleneck by substituting the activations based on the shifted time with the activations based on the original, see Figure 6 for an overview.

We perform the intervention experiment with the electricity dataset, and perform shifts of up to and including 23 hours. We compare the performance of the intervention with out-of-the-box performance of the same model on the shifted dataset. The results of the Vanilla Transformer shown in Figure 7.



(a) V. Transformer - Att

(b) V. Transformer - FF



(c) Autoformer - Att

(d) Autoformer - FF

Figure 7: MSE of the attention (Att) and feed-forward (FF) bottleneck models on electricity data with shifted timestamps. The dashed line represents the performance on the data without timeshift.

Remarkably, the intervention on the Vanilla Transformer achieves the original performance for all timeshifts. This indicates that the bottleneck models effectively learn to represent the hour-of-day concept in the dedicated bottleneck component. Most interestingly, the models only utilise this interpretable concept in the bottleneck layer, but not in other encoder layers (because the experiment only intervenes in the bottleneck).

5 Discussion and Conclusions

In this work, we propose a training framework based on Concept Bottleneck Models to enforce interpretability of time series transformers. We introduce a new loss function based on the similarity score CKA of the model's representations and interpretable concepts. We apply our framework to the Vanilla Transformer, Autoformer and FEDformer using synthetic data and six benchmark datasets. Our results indicate that the overall performance remains unaffected, while the model's components become more interpretable. Additionally, it becomes possible to perform a local intervention when employing the model after a temporal data shift.

The main limitation of our concept bottleneck framework is that interpretable concepts have to be decided on before training, which might require domain knowledge. Representations for these concepts have to be available during training. However, domain-agnostic concepts such as the AR surrogate model and hour-of-day information are sufficient. Additionally, our framework increases computational complexity. This might be problematic if the size of the architecture increases.

An interesting direction for future research would be to optimize the number and type of interpretable concepts in the bottleneck, and extend the framework to other modalities. We trained mostly using two domain-agnostic concepts (AR and hour-of-day), but including more concepts, possibly domain-specific, would be very interesting. For example, one could consider choosing speech and music concepts for audio time series. Additionally, the framework should also work for transformers in other modalities, e.g., language and vision, although these models are usually of larger size. We hope our work contributes to a deeper understanding of (time series) transformers and their behavior in different fields. In particular, recent progress in the field of mechanistic interpretability is based on the observation that the residual stream of the transformer encourages modular solutions, which enables localized concepts or specialized circuitry to perform a specific task. Instead of relying on post-hoc localization of these concepts, our paper presents a demonstration that we can encourage locality of concepts, without a significant loss in performance.

Regarding societal impact, this work enables transparent time series forecasting models, which enable explainable forecasts. However, in the case of malicious use, biases could be included in the models. Harm could be prevented by developing mechanistic interpretability techniques for bias detection in time series models.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable Neural-Symbolic Concept Reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1801–1825. PMLR, July 2023. URL https://proceedings.mlr.press/v202/barbiero23a.html. ISSN: 2640-3498.

Leonard Bereska and Efstratios Gavves. Mechanistic Interpretability for AI Safety – A Review, April 2024. URL http://arxiv.org/abs/2404.14082. arXiv:2404.14082 [cs].

Tom B. Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. URL https://splab.sdu.edu.cn/GPT3.pdf.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale Transformers with Adaptive Pathways for Time Series Forecasting, September 2024. URL http://arxiv.org/abs/2402.05956. arXiv:2402.05956 [cs].

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. URL https://bibbase.org/service/mendeley/bfbbf840-4c42-3914-a463-19024f50b30c/file/6375d223-e085-74b3-392f-f3fed829cd72/Devlin_et_al___2019___BERT_Pre_training_of_Deep_Bidirectional_Transform.pdf.pdf.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and Tom Conerly. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Irene Ferfoglia, Gaia Saveri, Laura Nenzi, and Luca Bortolussi. ECATS: Explainable-by-design concept-based anomaly detection for time series, July 2024. URL http://arxiv.org/abs/2405.10608. arXiv:2405.10608 [cs].

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348. PMLR, November 2020. URL https://proceedings.mlr.press/v119/koh20a.html. ISSN: 2640-3498.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited, July 2019. URL http://arxiv.org/abs/1905.00414. arXiv:1905.00414 [cs, q-bio, stat].

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. DecoderLens: Layerwise Interpretation of Encoder-Decoder Transformers, October 2023. URL http://arxiv.org/abs/2310.03686. arXiv:2310.03686 [cs].

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. pages 10012–10022, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and Pitfalls of Black-Box Concept Learning Models, June 2021. URL http://arxiv.org/abs/2106.13314. arXiv:2106.13314 [cs].

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2023. URL http://arxiv.org/abs/2202.05262. arXiv:2202.05262 [cs].

Zelin Ni, Hang Yu, Shizhan Liu, Jianguo Li, and Weiyao Lin. BasisFormer: Attention-based Time Series Forecasting with Learnable and Interpretable Basis. *Advances in Neural Information Processing Systems*, 36:71222–71241, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e150e6d0a1e5214740c39c6e4503ba7a-Abstract-Conference.html.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, May 2019. ISSN 0893-6080. doi: 10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based Explainable Artificial Intelligence: A Survey, December 2023. URL http://arxiv.org/abs/2312.12936. arXiv:2312.12936 [cs].

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models, July 2024. URL http://arxiv.org/abs/2407.02646. arXiv:2407.02646 [cs].

Zhen Tan, Lu Cheng, Song Wang, Bo Yuan, Jundong Li, and Huan Liu. Interpreting Pretrained Language Models via Concept Bottlenecks. In De-Nian Yang, Xing Xie, Vincent S. Tseng, Jian Pei, Jen-Wei Huang, and Jerry Chun-Wei Lin, editors, *Advances in Knowledge Discovery and Data Mining*, pages 56–74, Singapore, 2024. Springer Nature. ISBN 978-981-9722-59-4. doi: 10.1007/978-981-97-2259-4_5.

Songsong Tian, Weijun Li, Xin Ning, Hang Ran, Hong Qin, and Prayag Tiwari. Continuous transfer of neural network representational similarity for incremental learning. *Neurocomputing*, 545: 126300, August 2023. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.126300. URL https://www.sciencedirect.com/science/article/pii/S092523122300423X.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in Time Series: A Survey, May 2023. URL http://arxiv.org/abs/2202.07125. arXiv:2202.07125 [cs, eess, stat].

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off, December 2022. URL http://arxiv.org/abs/2209.09056. arXiv:2209.09056 [cs].

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers Effective for Time Series Forecasting?, August 2022. URL http://arxiv.org/abs/2205.13504. arXiv:2205.13504 [cs].

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, March 2021. URL http://arxiv.org/abs/2012.07436. arXiv:2012.07436 [cs].

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting, June 2022. URL http://arxiv.org/abs/2201.12740. arXiv:2201.12740 [cs, stat].

A Datasets

We evaluate the Autoformer model on six real-world benchmarks, covering the five domains of energy, traffic, economics, weather, and disease. We use the same datasets as Wu et al. [2021], and provide additional information in Table 2, as given in the original Autoformer paper.

Table 2: Descriptions of the datasets, as given by Wu et al. [2021] and shared online. 'Pred len' denotes the prediction length used in our experiments.

Dataset	Pred len	Description
Electricity	96	Hourly electricity consumption of 321 customers from 2012 to 2014.
Traffic	96	Hourly data from California Department of Transportation, which describes the road occupancy rates measured by different sensors on San Francisco Bay area freeways.
Weather	96	Recorded every 10 minutes for 2020 whole year, which contains 21 meteorological indicators, such as air temperature, humidity, etc.
Illness	24	Includes the weekly recorded influenza-like illness (ILI) patients data from Centers for Disease Control and Prevention of the United States between 2002 and 2021, which describes the ratio of patients seen with ILI and the total number of the patients.
Exchange rate	96	Daily exchange rates of eight different countries ranging from 1990 to 2016.
ETT	96	Data collected from electricity transformers, including load and oil temperature that are recorded every 15 minutes between July 2016 and July 2018.

B Formalization of Concept Bottleneck Framework

Any time series Transformer obtains two types of input: (1) *data values* $\boldsymbol{X} \in \mathbb{R}^{I \times d}$, and (2) *timestamps* $\boldsymbol{T} \in \mathbb{R}^{I \times 4}$. The transformer consists of an encoder and a decoder, which are both constructed from one or multiple layers. Any encoder layer contains two sub-layers: a multi-head attention mechanism (Att) and a fully connected neural network (FF). Every sub-layer contains a residual connection around it. More specifically, the output $\boldsymbol{X}^\ell$ of any encoder layer ℓ is:

$$\begin{aligned}\boldsymbol{X}^\ell &= \text{Encoder}(\boldsymbol{X}^{\ell-1}) \\ &= \text{LayerNorm}(\text{FF}(\boldsymbol{S}^\ell) + \boldsymbol{S}^\ell), \\ \boldsymbol{S}^\ell &= \text{LayerNorm}(\text{Att}(\boldsymbol{X}^{\ell-1}) + \boldsymbol{X}^{\ell-1}),\end{aligned}$$

where

$$\begin{aligned}\text{FF}(\mathbf{x}) &= \max(0,\, \mathbf{x}\boldsymbol{W}_1 + \mathbf{b}_1)\, \boldsymbol{W}_2 + \mathbf{b}_2, \\ \text{Att}(\mathbf{x}) &= \boldsymbol{W}_0 \cdot \text{Concat}\left(\text{h}_1(\mathbf{x}),\, \ldots,\, \text{h}_h(\mathbf{x})\right).\end{aligned}$$

For future reference, we denote the output of the feed-forward module as follows: $\text{FF}(\boldsymbol{S}^\ell) = \boldsymbol{Z}^\ell \in \mathbb{R}^{d_1 \times d_2}$. We omit the definition of the decoder, because our bottleneck framework does not include it. Note that the exact implementation of each (sub-)layer depends on the type of Transformer.

B.1 Bottleneck Layer

We assign one encoder layer to be the bottleneck and construct it such that it contains c latent representations or *components*, i.e., $(\boldsymbol{H}_i)_{i=1}^{c}$. Depending on the bottleneck type τ, these latent representations are either taken from the attention mechanism or the feed-forward module. More specifically:

$$\boldsymbol{H}_i = \begin{cases} \mathrm{h}_i(\mathbf{x}) & \text{if bottleneck type } \tau = \text{Att}, \\ \boldsymbol{Z}_i & \text{if bottleneck type } \tau = \text{FF}. \end{cases}$$

Since the attention block is multi-headed, different heads naturally form the components of the attention bottleneck. For the feed-forward bottleneck, we define the components to be slices (in d_1) from its output $\boldsymbol{Z}$, such that stacking the components results in the original output.

Note that the residual connection around the corresponding bottleneck component is removed, and that each component $\boldsymbol{H}_i$ should represent a pre-defined interpretable concept.

B.2 Intervention

In the intervention experiment, we shift the time stamps $\boldsymbol{T}$ to obtain $\widetilde{\boldsymbol{T}}$. The key aspect of the experiment is to run the Transformer on the shifted time stamps $\widetilde{\boldsymbol{T}}$, and replace the input representations $\widetilde{\boldsymbol{X}}^{b-1}$ of the bottleneck layer b with $\boldsymbol{X}^{b-1}$ (based on $\boldsymbol{T}$), but only in the component that represents the time concept.

More specifically, if type $\tau = \text{Att}$, we intervene on the attention block in the bottleneck as follows:

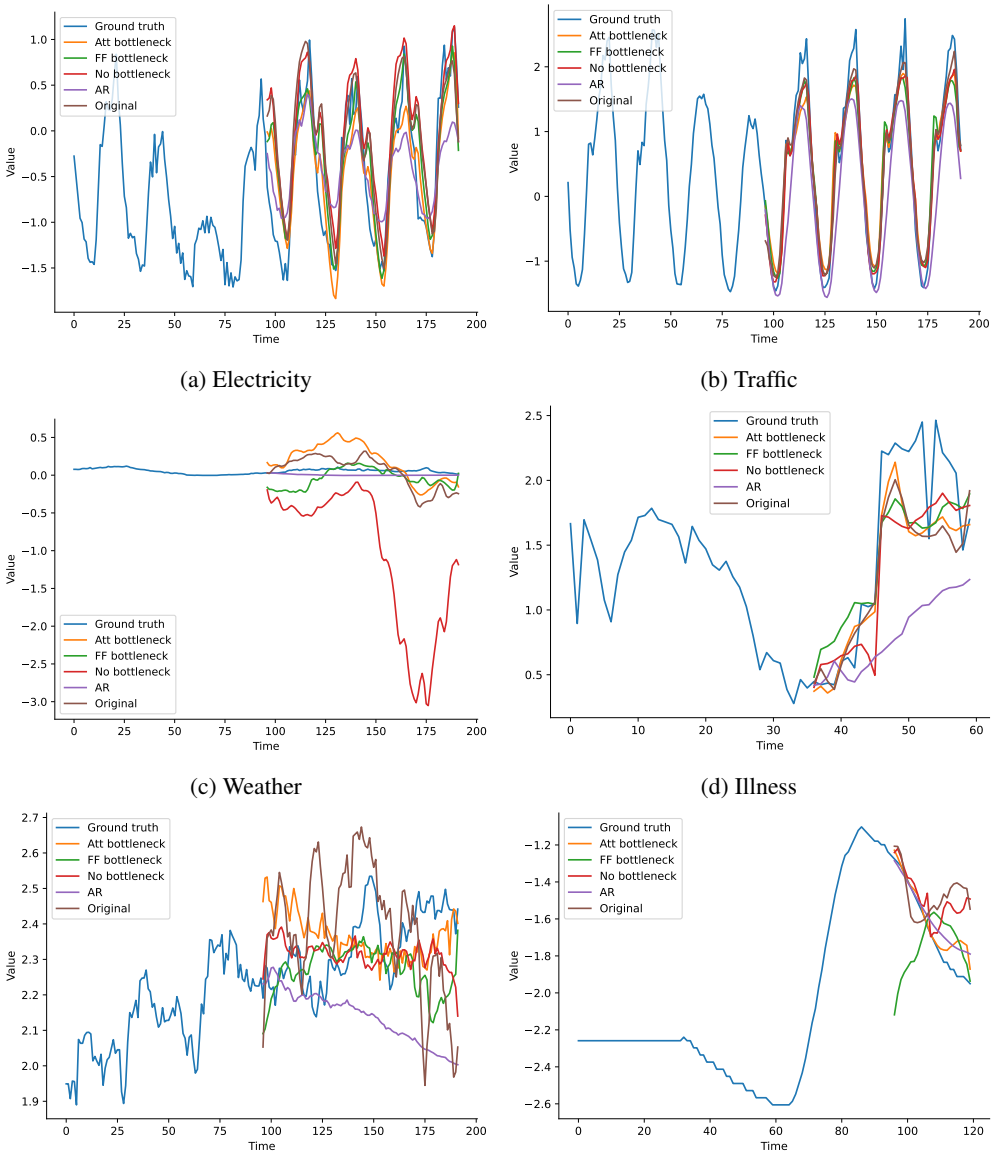
$$\text{Att}(\mathbf{x}, \widetilde{\mathbf{x}}) = \boldsymbol{W}_0 \cdot \text{Concat}\left(\mathrm{h}_1(\widetilde{\mathbf{x}}),\, \mathrm{h}_2(\mathbf{x}),\, \mathrm{h}_3(\widetilde{\mathbf{x}})\right),$$

and, if type $\tau = \text{FF}$, as follows:

$$\text{FF}(\mathbf{x}, \widetilde{\mathbf{x}}) = \text{Stack}(\widetilde{\boldsymbol{Z}}_1, \boldsymbol{Z}_2, \widetilde{\boldsymbol{Z}}_3).$$

In both functions we make use of the fact that the time concept is represented in the second component, and there are three components in total. This intervention can be done in the bottleneck only, because, by construction, its location of the concept representations is known.

C Qualitative Results



(a) Electricity

(b) Traffic

(c) Weather

(d) Illness

(e) Exchange rate

(f) ETT

Figure 8: Forecasts on different datasets. The first part of the ground truth (shown in blue) is the input for the models, and the test set is used for each dataset.

D Detailed results

Table 3: Performance of different models in Mean Squared Error (MSE) and Mean Absolute Error (MAE). The bottlenecks **do** contain a free component ($c = 3$), and use AR as surrogate model. The model with no bottleneck is an original Autoformer of similar size. For all datasets, the shortest prediction lengths from Wu et al. [2021] are used, see Table 2. The standard deviation is determined using five different seeds.

Free component	**Att bottleneck**		**FF bottleneck**		**No bottleneck**		**AR**		**Wu et al.**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	0.231 ± 0.009	0.338 ± 0.005	0.207 ± 0.005	0.320 ± 0.005	0.280 ± 0.165	0.368 ± 0.111	0.497	0.522	**0.201** ± 0.003	**0.317** ± 0.004
Traffic	0.642 ± 0.022	0.393 ± 0.013	**0.393** ± 0.013	**0.377** ± 0.006	0.619 ± 0.015	0.387 ± 0.005	0.420	0.494	0.613 ± 0.028	0.388 ± 0.012
Weather	0.290 ± 0.027	0.354 ± 0.020	0.271 ± 0.016	0.341 ± 0.011	0.269 ± 0.000	0.344 ± 0.000	**0.006**	**0.062**	0.266 ± 0.007	0.336 ± 0.006
Illness	3.586 ± 0.241	1.313 ± 0.040	3.661 ± 0.237	1.322 ± 0.050	3.405 ± 0.208	1.295 ± 0.044	**1.027**	**0.820**	3.483 ± 0.107	1.287 ± 0.018
Exchange rate	0.195 ± 0.029	0.323 ± 0.025	0.155 ± 0.010	0.290 ± 0.013	0.152 ± 0.003	0.283 ± 0.003	**0.082**	**0.230**	0.197 ± 0.019	0.323 ± 0.012
ETT	0.177 ± 0.003	0.282 ± 0.004	0.174 ± 0.006	0.280 ± 0.005	0.155 ± 0.004	0.265 ± 0.002	**0.034**	**0.117**	0.255 ± 0.020	0.339 ± 0.020

Table 4: Performance on different datasets, where the bottlenecks **do not** contain a free component ($c = 2$). AR is used as surrogate model in the bottlenecks. The model with no bottleneck is an original Autoformer of similar size. For all datasets, the shortest prediction lengths from Wu et al. [2021] are used, see Table 2. The standard deviation is determined using five different seeds.

No free component	**Att bottleneck**		**FF bottleneck**		**No bottleneck**		**AR**		**Wu et al.**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	0.224 ± 0.006	0.332 ± 0.003	0.206 ± 0.009	0.321 ± 0.009	0.202 ± 0.006	0.318 ± 0.007	0.497	0.522	**0.201** ± 0.003	**0.317** ± 0.004
Traffic	0.629 ± 0.023	0.394 ± 0.015	0.627 ± 0.031	0.392 ± 0.025	0.613 ± 0.018	**0.378** ± 0.007	**0.420**	0.494	0.613 ± 0.028	0.388 ± 0.012
Weather	0.281 ± 0.025	0.348 ± 0.018	0.260 ± 0.015	0.333 ± 0.013	0.257 ± 0.004	0.332 ± 0.005	**0.006**	**0.062**	0.266 ± 0.007	0.336 ± 0.006
Illness	3.966 ± 0.296	1.401 ± 0.073	3.721 ± 0.268	1.351 ± 0.053	3.585 ± 0.331	1.333 ± 0.070	**1.027**	**0.820**	3.483 ± 0.107	1.287 ± 0.018
Exchange rate	0.208 ± 0.026	0.333 ± 0.022	0.158 ± 0.009	0.293 ± 0.009	0.152 ± 0.006	0.284 ± 0.007	**0.082**	**0.230**	0.197 ± 0.019	0.323 ± 0.012
ETT	0.178 ± 0.011	0.283 ± 0.007	0.174 ± 0.01	0.283 ± 0.009	0.165 ± 0.004	0.274 ± 0.004	**0.034**	**0.117**	0.255 ± 0.020	0.339 ± 0.020

E Hyper-Parameter Sensitivity

To verify the sensitivity to hyperparameter α in the loss function, we train the Autoformer with a feed-forward bottleneck on different values for α, where the bottleneck contains a free component ($c = 3$) and the model is trained on the electricity dataset. The results are given in Figure 9. Interestingly, the error scores for all $\alpha < 1$ are close in value, which verifies that additionally training for interpretability does not hurt the performance, at least not in this set-up. Note that a low forecasting error cannot be expected for $\alpha = 1$, because in this edge case the loss function does not contain any term that represents the forecasting performance.

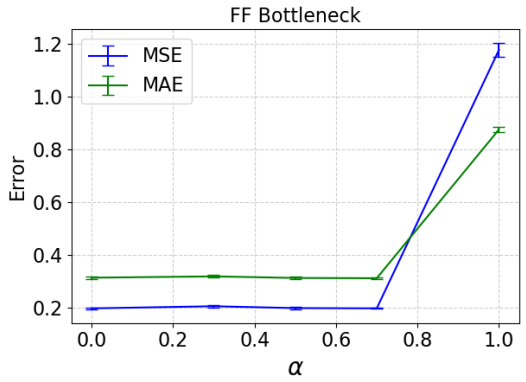
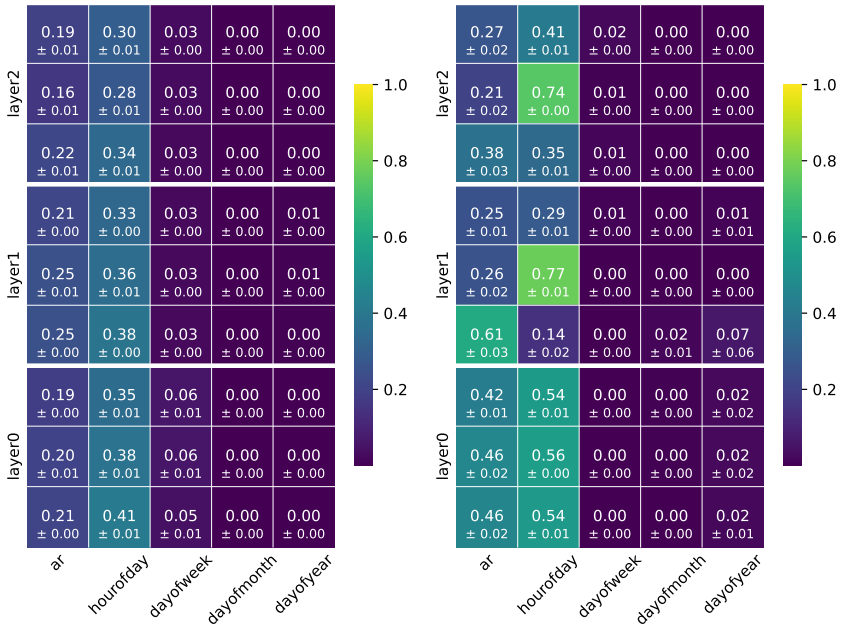


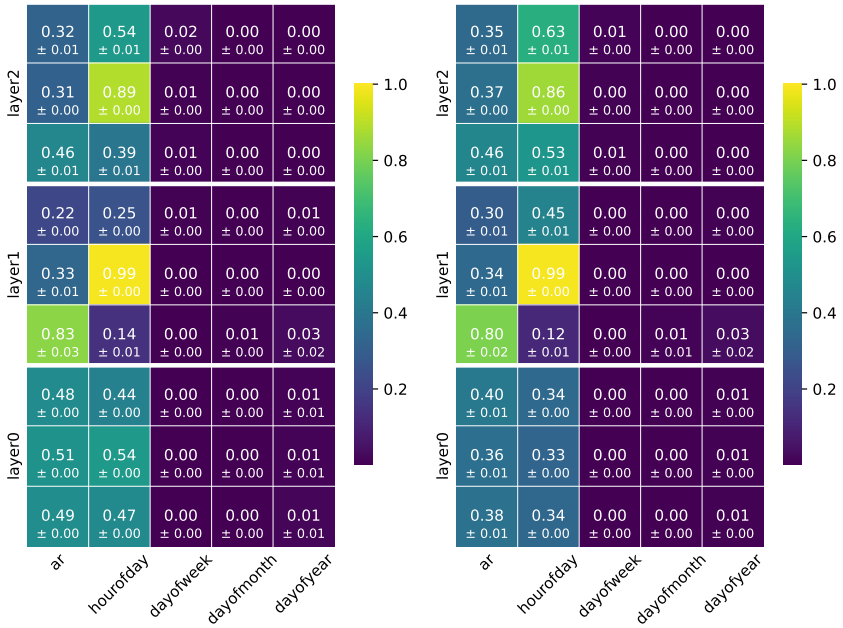
Figure 9: Performance of the Autoformer for different values of α in MSE and MAE.

Additionally, the CKA scores of the different models with the interpretable concepts (and other time features) are given in Figures 10, 11, and 12. Naturally, the CKA scores are the lowest in the setting $\alpha = 0$, and the scores from the bottleneck (`layer1`) increase over α. Interestingly, the CKA scores from the bottleneck do not increase for higher values than $\alpha = 0.5$, although the scores of some other components do increase. This indicates that perfect similarity (i.e. CKA score of 1) to some interpretable concepts may not be reached.



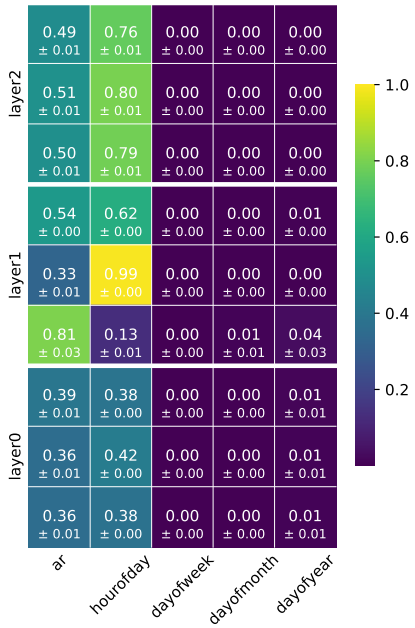
(a) $\alpha = 0$ (b) $\alpha = 0.3$

Figure 10: CKA scores of the feed-forward bottleneck Autoformer on electricity data for different values of hyperparameter α. The scores are calculated using three batches of size 32 of the test data set.



(a) $\alpha = 0.5$ (b) $\alpha = 0.7$

Figure 11: CKA scores of the feed-forward bottleneck Autoformer on electricity data for different values of hyperparameter α. The scores are calculated using three batches of size 32 of the test data set.



(a) $\alpha = 1$

Figure 12: CKA scores of the feed-forward bottleneck Autoformer on electricity data for hyperparameter $\alpha = 1$. The scores are calculated using three batches of size 32 of the test data set.

F Application of Framework to Vanilla Transformer

To demonstrate the generality of the concept bottleneck framework, we apply it to an additional Transformer architecture, namely the *vanilla Transformer* (the original architecture from which all Transformer models, including all time series Transformers, are derived). We train it using the same six benchmark datasets and perform a similar, but less extensive, analysis as done for the Autoformer model. Note that the architecture of the Transformer is *not* modified, and the timestamps are included as an embedding (in addition to the positional embedding).

F.1 Performance Analysis

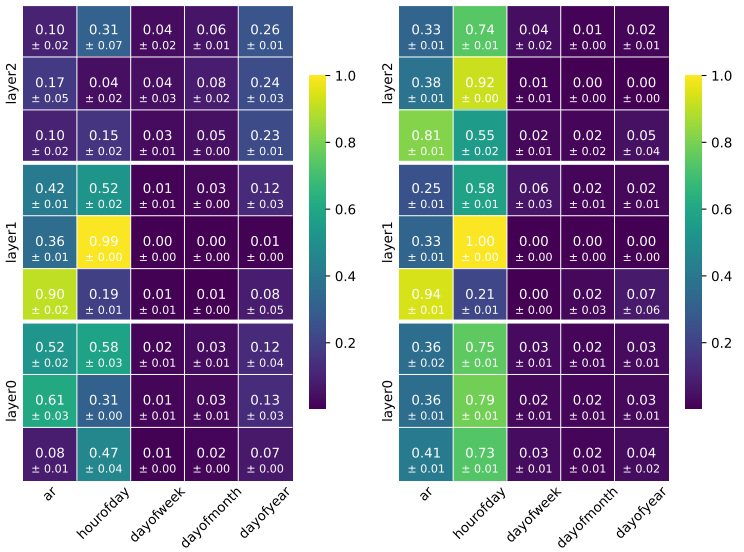
The performance of the vanilla Transformer model with and without bottleneck is given in Table 5. We train the bottleneck with a 'free' component (the side channel), i.e., with $c = 3$. Note that Wu et al. [2021] do not provide scores for these benchmark forecasting datasets, therefore we cannot include them in the table. The results show that the vanilla Transformer performs, unsurprisingly, worse than the Autoformer, and for most datasets also worse than the linear AR model. However, most relevant, for our purposes, is that across the datasets using a concept bottleneck does not hurt the overall performance of the vanilla Transformer.

F.2 CKA Analysis

After training the vanilla Transformer with the bottleneck framework, we evaluate the similarity of its hidden representations to the interpretable concepts using CKA, see Figure 13. Recall that CKA scores are defined in the range from 0 to 1, where 1 indicates perfect similarity. Both components in the two types of bottleneck show very high similarity to their target concept. Interestingly, the first component in the bottleneck (the AR concept) shows a higher similarity to the AR representations than the Autoformer (see Figure 4), presumably because the decomposition structure of the Autoformer hinders learning a linear function.

Table 5: Performance of different vanilla Transformer models. For both metrics, it holds that a lower score indicates a better performance, where the best results are **bold**, and the second-best are underlined.

	Att bottleneck		FF bottleneck		No bottleneck		AR	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	0.275	0.371	**0.268**	**0.362**	0.275	0.371	0.497	0.522
Traffic	0.708	0.394	0.703	0.397	0.684	**0.376**	**0.420**	0.494
Weather	0.400	0.450	0.381	0.410	0.362	0.415	**0.006**	**0.062**
Illness	3.380	1.280	3.323	1.252	3.321	1.273	**1.027**	**0.820**
Exchange rate	0.675	0.642	0.677	0.633	0.694	0.662	**0.082**	**0.230**
ETT	0.230	0.328	0.185	0.299	0.166	0.294	**0.034**	**0.117**

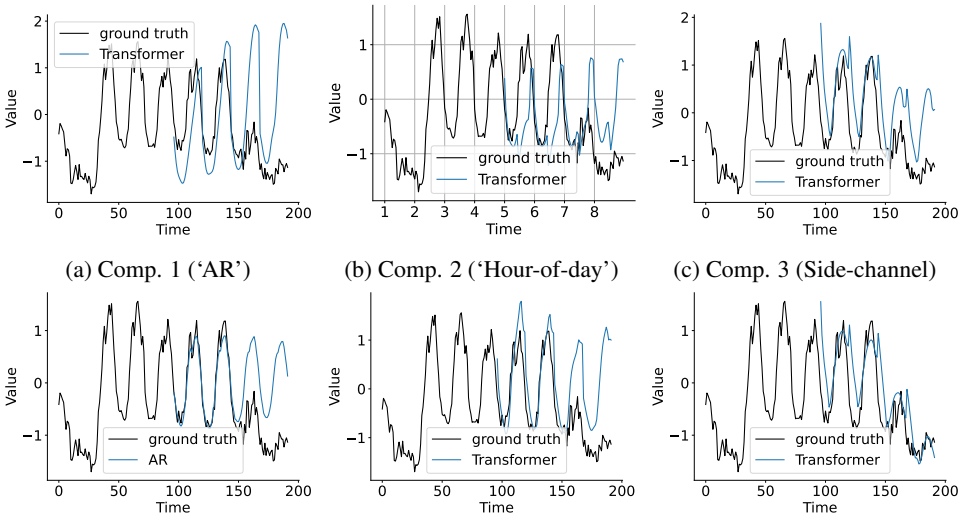


(a) Att bottleneck (b) FF bottleneck

Figure 13: CKA scores of the vanilla Transformer's encoder (containing three heads per layer) from the attention and feed-forward bottleneck on the electricity dataset, where each score denotes the similarity of an individual component. The first component of `layer1` is trained to be similar to AR, and the second component to the hour-of-day concept (lower and middle row in the figure, respectively). The scores are calculated using three batches of size 32 from the test data set.

F.3 Component Visualizations

We visualize the contributions of each component in the bottleneck using the Decoder Lens method [Langedijk et al., 2023], see Figure 14. We obtain the output from each component individually by masking the other components with zero (close to the mean). Each component seems to provide similar contributions to the forecast as their respective counterpart in the Autoformer model. In particular, the first component (see Figure 14a) produces forecasts of correct seasonality and few irregularities, similar to the AR model. The second component (see Figure 14b) follows the hour-of-day feature, and the free head (see Figure 14c) picks up on high-frequency data patterns.



(a) Comp. 1 ('AR') (b) Comp. 2 ('Hour-of-day') (c) Comp. 3 (Side-channel)

(d) AR (e) Full bottleneck (f) Final, using only comp. 3

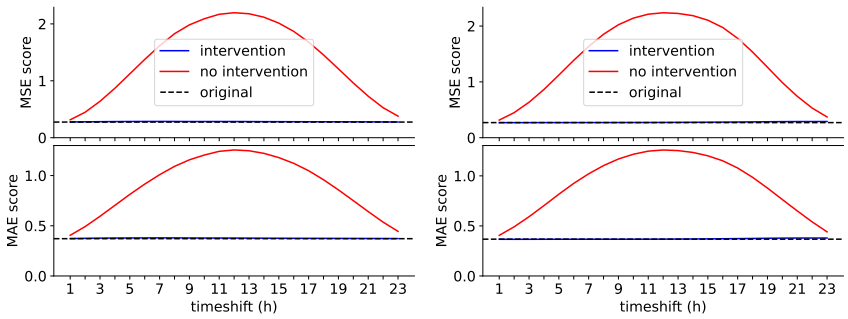
Figure 14: Vanilla Transformer forecasts from the components in the bottleneck layer (FF bottleneck on electricity data) in 14a, 14b and 14c. They are obtained by masking the other components with zero (the mean). The first half of the ground truth forms the input to the model. Note that the horizontal axes are the same across all figures, but Figure 14b contains a grid of days instead of numbered hours. Figure 14d shows the forecast made by the surrogate model AR; Figure 14e shows the forecast of the entire layer (i.e., all components together), and 14f shows the forecast of the final layer when only the third component is used in the bottleneck layer. Note the difference between Figures 14c and 14f, where we decode from the bottleneck and the final layer, respectively.

F.4 Intervention

We perform the intervention experiment in the same set-up as for the Autoformer model. That is, we delay the input timestamps with a fixed number of hours to obtain shifted timestamps, and perform an intervention in the bottleneck by substituting the activations based on the shifted time with the activations from the original time. We use a vanilla Transformer trained on the electricity dataset, and perform shifts of up to and including 23 hours. We compare the performance of the intervention with out-of-the-box performance of the same model on the shifted dataset. The results are shown in Figure 15. For both types of bottlenecks, the intervention performs best for all timeshifts, by keeping the error scores marginally close to the original performance (with no timeshift). This indicates that the model effectively learns to represent the hour-of-day concept in the dedicated head, which is able to provide control over the model's behavior.

F.5 Conclusion

By repeating the set of experiments for the vanilla Transformer model, we provided further evidence for the generality of the concept bottleneck framework. In particular, we showed that the framework can be applied to the vanilla Transformer model, without having any significant impact on the overall model performance, while providing improved interpretability.



(a) Attention bottleneck (b) Feed-forward bottleneck

Figure 15: Performance of the bottleneck vanilla Transformer on electricity data with shifted timestamps. The dashed line represents the performance of the same model on the original data, i.e., with no timeshift.

G Application of Framework to FEDformer

To demonstrate the generality of our concept bottleneck framework, we apply it to *FEDformer* [Zhou et al., 2022]. This is a Transformer architecture containing *Fourier enhanced blocks* and *wavelet enhanced blocks* to represent time series in the frequency domain. For more details, we refer to the original authors Zhou et al. [2022]. We train the model on the same six datasets and perform an interpretability analysis.

G.1 Performance Analysis

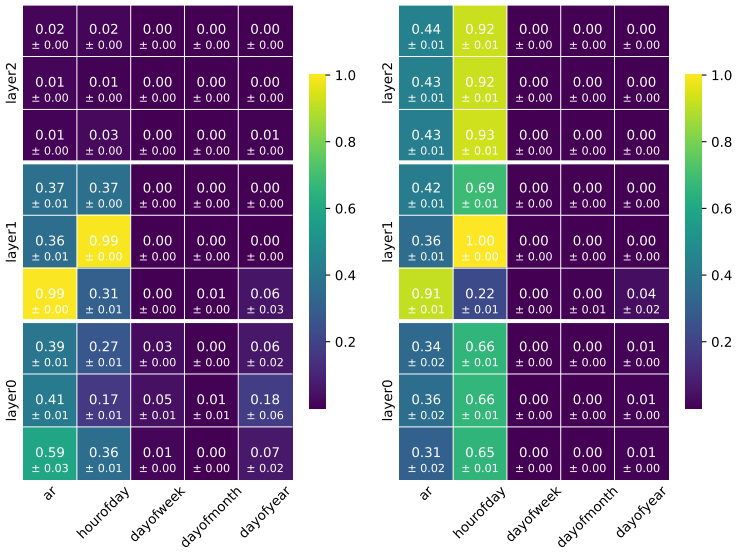
The performance of the FEDformer with and without bottleneck is given in Table 6. We train the bottleneck with a 'free' component (the side channel), i.e., with $c = 3$. Note that the model by Zhou et al. [2022] is of a different size (two encoder layers with eight heads per layer). Interestingly, we find for some datasets (e.g. electricity and illness) that including a bottleneck increases the performance, while it has little effect on the performance for the other datasets. We can conclude for all datasets that including a bottleneck does not hurt performance.

Table 6: Performance of FEDformer. For both metrics, it holds that a lower score indicates a better performance, where the best results are **bold**, and the second-best are underlined.

	Att bottleneck		**FF bottleneck**		**No bottleneck**		**AR**		**Zhou et al.**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	**0.185**	**0.302**	0.186	0.303	0.189	0.304	0.497	0.522	0.193	0.308
Traffic	0.585	0.364	0.585	0.364	0.573	**0.358**	**0.420**	0.494	0.587	0.366
Weather	0.221	0.299	0.219	0.296	0.334	0.397	**0.006**	**0.062**	0.217	0.296
Illness	3.070	1.217	3.076	1.219	3.111	1.232	**1.027**	**0.820**	3.228	1.260
Exchange rate	0.147	0.277	0.145	0.275	0.146	0.276	**0.082**	**0.230**	0.148	0.278
ETT	0.079	0.193	0.079	0.192	0.077	0.190	**0.034**	**0.117**	0.203	0.287

G.2 CKA Analysis

After training the FEDformer with our concept bottleneck framework, we evaluate the similarity of the hidden representations to the interpretable concepts using CKA, see Figure 16. Recall that CKA scores are defined in the range from 0 to 1, where 1 indicates perfect similarity. Both components in the two types of bottleneck show a very high similarity to their target concept, indicating a successful training on interpretability.

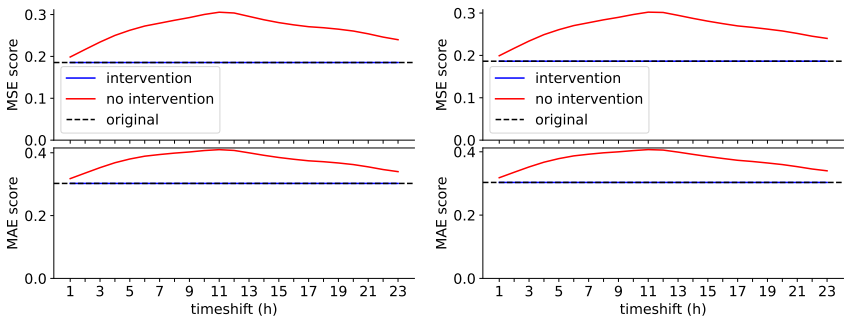


(a) Att bottleneck (b) FF bottleneck

Figure 16: CKA scores of the FEDformer's encoder (containing three heads per layer) from the attention and feed-forward bottleneck on the electricity dataset, where each score denotes the similarity of an individual component. The first component of `layer1` is trained to be similar to AR, and the second component to the hour-of-day concept (lower and middle row in the figure, respectively). The scores are calculated using three batches of size 32 from the test data set.

G.3 Intervention

Additionally, we perform the intervention experiment in the same set-up as for the other Transformer models. That is, we delay the input timestamps with a fixed number of hours and perform an intervention in the bottleneck by substituting the activations with those based on the original time. We compare the performance of the intervention with out-of-the-box performance of the same model on the shifted dataset. The results are shown in Figure 17. For both types of bottlenecks, the intervention performs best for all timeshifts, by keeping the error marginally close to the original performance (without timeshift). This indicates that the model effectively learns to represent the hour-of-day concept in the dedicated head, which is able to provide control over the model's behavior.



(a) Attention bottleneck (b) Feed-forward bottleneck

Figure 17: Performance of the bottleneck FEDformer on electricity data with shifted timestamps. The dashed line represents the performance of the same model on the original data, i.e., with no timeshift.

G.4 Conclusion

By repeating the set of experiments for the FEDformer model, we provided further evidence for the generality of the concept bottleneck framework. In particular, we showed that the framework can be applied to the FEDformer model, without having any significant impact on the overall model performance, while providing improved interpretability.

H Synthetic Data

To increase the understanding of how the concepts in the bottleneck can be leveraged, we train the model on a synthetic dataset.

H.1 Dataset

We generate a synthetic time series as the sum of different functions. In particular, the dataset is generated using the function f_{Total} with time t as follows:

$$f_{Total}(t) = f_1(t) + f_2(t) + f_3(t),$$

where:

$$\begin{aligned} f_1(t) &= \sin(2\pi t), \\ f_2(t) &= \frac{1}{2}\sin(4\pi t + \frac{\pi}{4}), \\ f_3(t) &= \frac{1}{4}\sin(6\pi t + \frac{\pi}{2}) + \epsilon_t. \end{aligned}$$

Note that all functions f_1, f_2 and f_3 follow a periodic structure, and f_3 contains random noise ϵ from a normal distribution with standard deviation of 0.2. See Figure 18 for a visualization of the functions.

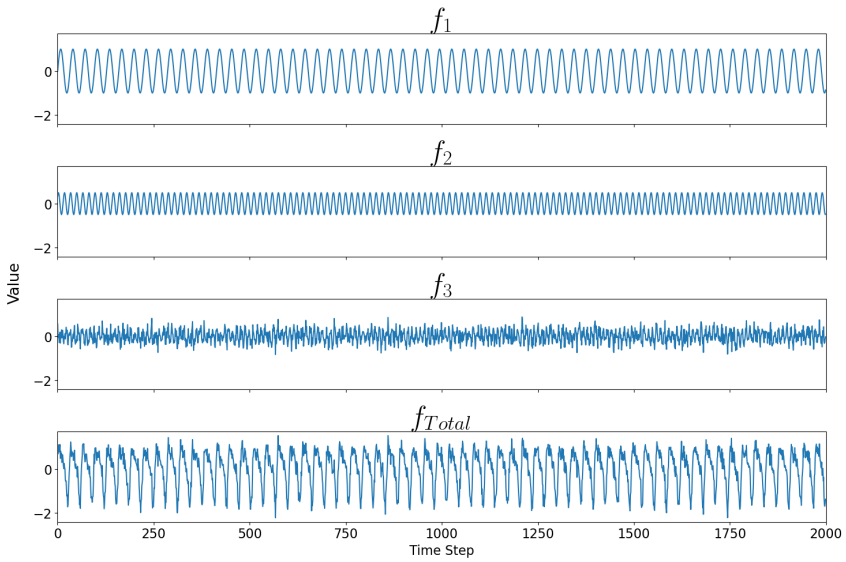


Figure 18: The synthetic time series dataset.

H.2 Experiment and Results

We train an Autoformer model on the synthetic dataset using the concept bottleneck framework. Each concept in the bottleneck is defined as one of the underlying functions (i.e., f_1, f_2 or f_3), for which the ground-truth is known by construction. The model contains three encoder layers, with three attention heads per layer. We apply the bottleneck to the attention heads of the second encoder layer. Additionally, we train the bottleneck using different values for hyperparameter α, which controls the weight of the CKA loss in the total loss function (see Section 3.1).

As expected, we find for all values $\alpha < 1$ that the model is able to forecast the dataset well, see Figure 19. Note that a low forecasting error cannot be expected for $\alpha = 1$, because in this edge case the loss function does not contain any term that represents the forecasting error. Remarkably, for all other cases, the performance of the Autoformer seems to improve as α increases. This suggests that properly chosen concepts improve the performance of the model, at least when the ground-truth

underlying functions are known. It should be noted that the standard deviation is higher for all $\alpha > 0$, which indicates that initialization of the parameters is important when learning the bottleneck. Additionally, visualizations of the predictions are given in Figure 20.

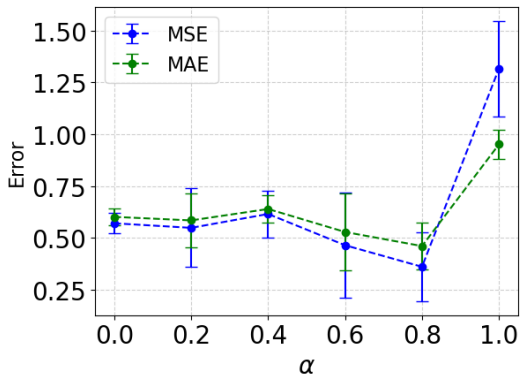
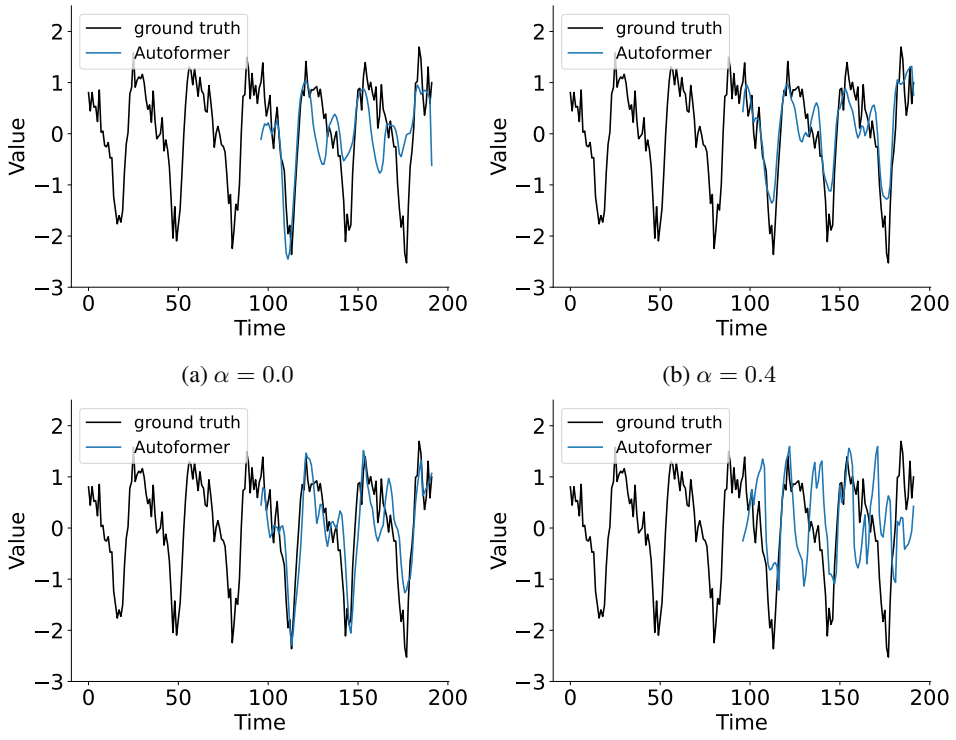


Figure 19: Performance on the synthetic dataset for different values of α, using an Autoformer with attention bottleneck. For both metrics, it holds that a lower score indicates a better performance. The standard deviation is provided over 5 different seeds.



(a) $\alpha = 0.0$ (b) $\alpha = 0.4$

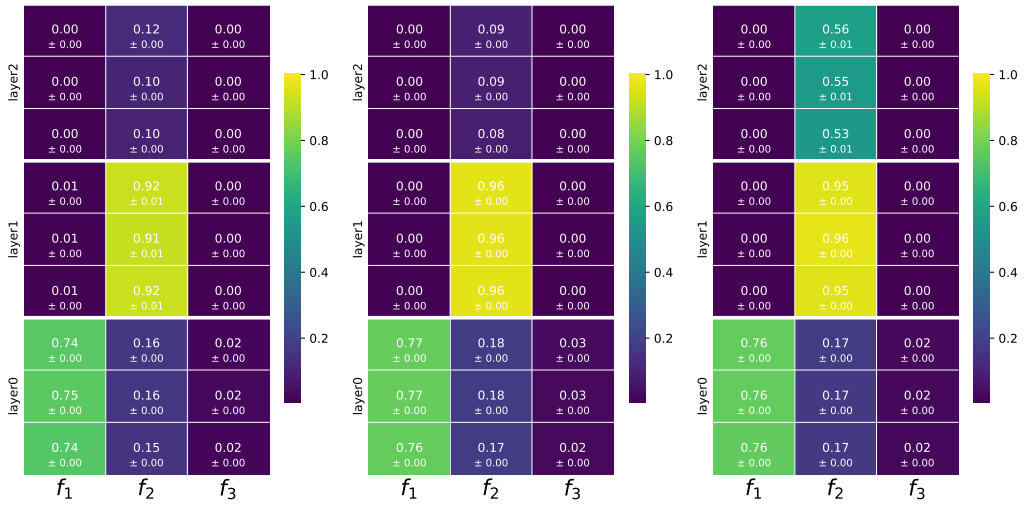
(c) $\alpha = 0.8$ (d) $\alpha = 1.0$

Figure 20: Predictions of the Autoformer model on a sample from the test dataset. The Autoformer is trained with an attention bottleneck using different values of hyperparameter α and the same seed.

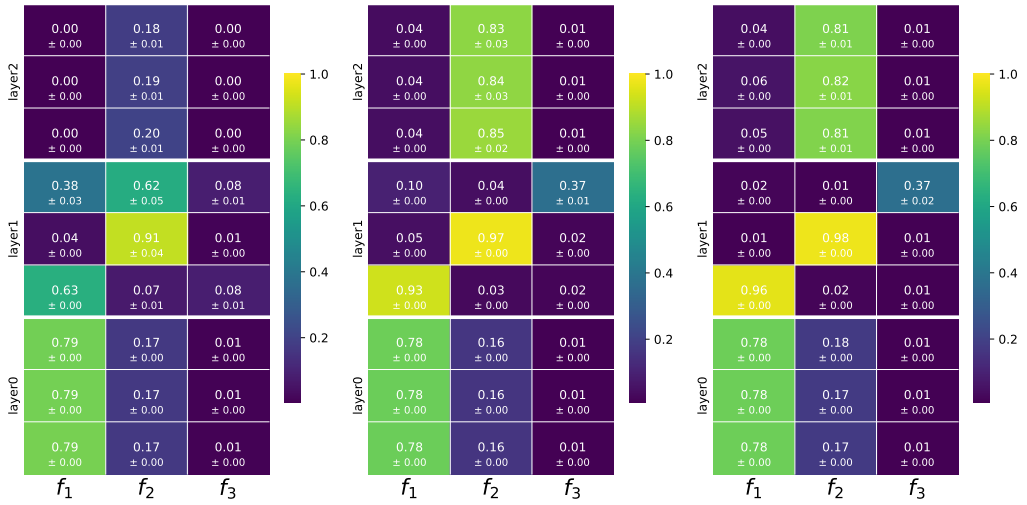
Additionally, the different values of hyperparameter α show clearly how the different concepts are leveraged by the model, see Figure 21. The figure shows the similarity scores between the attention heads and the different underlying functions of the dataset. Without the CKA loss, at $\alpha = 0$, the different heads in `layer1` of the model do not show high similarity to their respective concepts, i.e., functions. Instead, all heads have a high similarity to concept f_2. This is different for higher values

of α, where the different heads show higher similarity to their respective concepts. Note that the third concept f_3 cannot be perfectly learned by the model because of the random noise component.

All in all, these results show that a higher value for α, which is equivalent to a higher weight of the CKA loss in the total loss function, results in more similarity of the bottleneck components to their respective concepts, as expected.



(a) $\alpha = 0$ (b) $\alpha = 0.2$ (c) $\alpha = 0.4$



(d) $\alpha = 0.6$ (e) $\alpha = 0.8$ (f) $\alpha = 1$

Figure 21: CKA scores of the attention bottleneck Autoformer on synthetic data for different values of hyperparameter α. The scores are calculated using three batches of size 32 of the test data set.

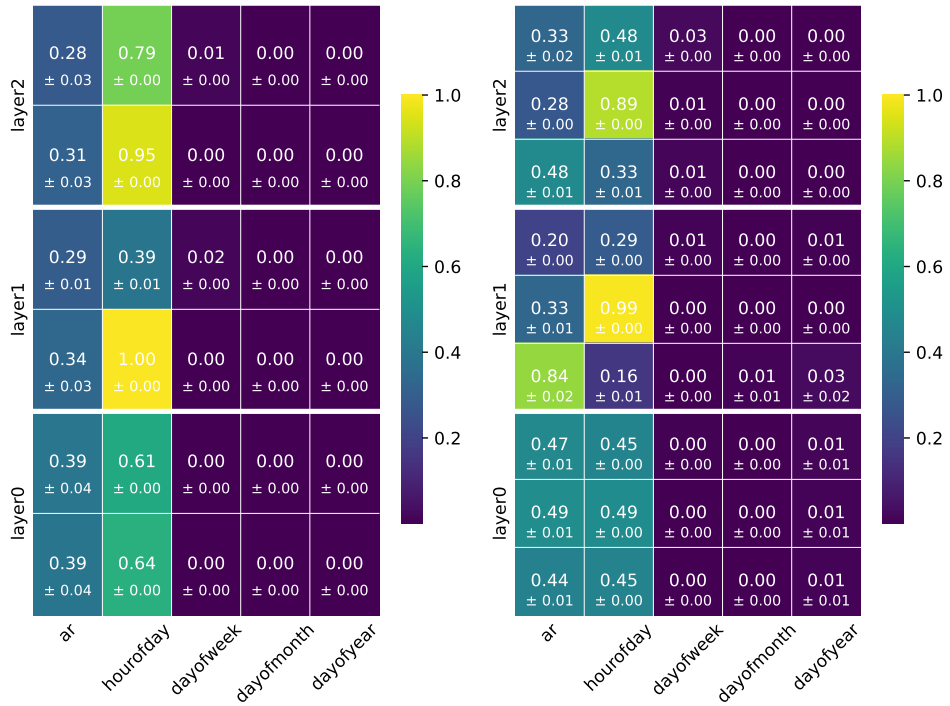
I Effect of AR as Surrogate Model

Interestingly, the AR model outperforms the Autoformer for some datasets (see Table 1). This raises the question whether the AR surrogate model makes up for any loss in performance introduced by the concept bottleneck.

To test this, we train an Autoformer without the AR concept. Specifically, we include the time concept and a free component in the feed-forward bottleneck. Here, the free component refers to a component in the bottleneck that is not included in the CKA loss (see Section 3.2).

The performance on the electricity data for this model is (MSE: 0.206, MAE: 0.321), which is seemingly identical to the original performance of (MSE: 0.207, MAE: 0.320). This suggests that it is not the AR head that makes up for the loss in performance. The CKA plots, see Figure 22, verify that there is no component in the minimal set-up (without AR) that is very similar to the AR model, unlike in the original set-up. So, these results show that the AR model does not add performance to the bottleneck model, merely interpretability.

Additionally, we refer the reader to Appendix H, where we perform more experiments on training the bottleneck without the AR surrogate model.



(a) Without AR (MSE: 0.206, MAE: 0.321)

(b) With AR (MSE: 0.207, MAE: 0.320)

Figure 22: CKA plots of two Autoformer models with feed-forward bottlenecks. The model in 22a is trained without AR in the bottleneck, while the model in 22b is trained with AR. Note that the upper component in `layer1` is the free component in both plots.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses the limitations of the work, including the assumptions and computational efficiency of the method.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 discloses all information needed to reproduce the main experimental results. Additionally, Appendix B provides a formalization of the proposed method.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Upon acceptance of the paper in an archive, we will release the code. The data is open-source, and already available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3.3 specificies the implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix D and E report error margins and details on hyperparameter sensitivity.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.3 provides the information on computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: There are no violations with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Potential social impacts are discussed in Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators of the assets used in the paper are cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.