# Frequency-Masked Embedding Inference: A Non-Contrastive Approach for Time Series Representation Learning

**En Fu**[1], **Yanyan Hu**[1,2,*],

[1]School of Intelligence Science and Technology, University of Science and Technology Beijing
[2]Institute of Artificial Intelligence, University of Science and Technology Beijing
fuen@xs.ustb.edu.cn, huyanyan@ustb.edu.cn

## Abstract

Contrastive learning underpins most current self-supervised time series representation methods. The strategy for constructing positive and negative sample pairs significantly affects the final representation quality. However, due to the continuous nature of time series semantics, the modeling approach of contrastive learning struggles to accommodate the characteristics of time series data. This results in issues such as difficulties in constructing hard negative samples and the potential introduction of inappropriate biases during positive sample construction. Although some recent works have developed several scientific strategies for constructing positive and negative sample pairs with improved effectiveness, they remain constrained by the contrastive learning framework. To fundamentally overcome the limitations of contrastive learning, this paper introduces Frequency-masked Embedding Inference (FEI), a novel non-contrastive method that completely eliminates the need for positive and negative samples. The proposed FEI constructs 2 inference branches based on a prompting strategy: 1) Using frequency masking as prompts to infer the embedding representation of the target series with missing frequency bands in the embedding space, and 2) Using the target series as prompts to infer its frequency masking embedding. In this way, FEI enables continuous semantic relationship modeling for time series. Experiments on 8 widely used time series datasets for classification and regression tasks, using linear evaluation and end-to-end fine-tuning, show that FEI significantly outperforms existing contrastive-based methods in terms of generalization. This study provides new insights into self-supervised representation learning for time series.

**Code/Appendix** —
https://github.com/USTBInnovationPark/Frequency-masked-Embedding-Inference

## Introduction

Time series data is crucial in various industries' production processes and economic activities, serving as a foundational data format(Trirat et al. 2024; Ozyurt, Feuerriegel, and Zhang 2023). Effective representation methods are essential for building highly transferable and generalizable pattern recognition modules, significantly reducing the optimization difficulty of deep models on small sample datasets(Tonekaboni, Eytan, and Goldenberg 2021; Zerveas et al. 2021). This has become a consensus in the deep learning community, particularly in the fields of computer vision (CV)(Han et al. 2022; Li et al. 2021; LeCun 2022) and natural language processing (NLP)(Devlin et al. 2019; Fang et al. 2020; Yan et al. 2021).

However, in the field of time series analysis, self-supervised representation learning has not yet become to a community standard due to the insufficient generalization performance. Contrastive learning strategies have shown certain advantages and have become the foundational framework for much research in recent years(Yue et al. 2022; Zhang et al. 2022; Luo et al. 2023). Contrastive learning optimizes models by constructing positive and negative samples for anchor sample. Positive samples are created using data augmentation techniques to provide different views under similar semantics to the anchor sample, while negative samples are selected or constructed to have opposing semantics to the anchor sample. The main issue with this framework is the difficulty in defining appropriate positive and negative samples for time series data.

Due to the inherent continuity of time series, key characteristics such as trends and frequencies change continuously and cannot be fully enumerated. **This makes the boundaries between semantics in time series less clear compared to other data formats.** For example, it is easy to define a picture of a cat and a picture of a dog as opposites, but it is challenging to define whether a series with a 7-day cycle and that with a 6.5-day cycle are similar or opposite. They have differences but are not fundamentally opposed. Contrastive learning uses a discrete modeling approach to define absolute opposite semantics for all samples, which contradicts the inherent continuity of time series.

Some methods(Luo et al. 2023; Liu and Chen 2024) have developed more reasonable and flexible strategies for constructing positive and negative samples to mitigate the issues mentioned above. However, they still operate within the discrete modeling framework of contrastive learning, failing to address the root cause of the problem. Therefore, this study aims to define the semantic differences in time series in a completely new way.

In this paper, a novel non-contrastive time series representation learning framework, Frequency-masked Embed-

*Corresponding Author

ding Inference(FEI), is proposed. Inspired by the successful application of Joint-Embedding Predictive Architecture (JEPA)(LeCun 2022; Assran et al. 2023) in CV, FEI establishes continuous relationships between different semantics in time series through the concept of embedding inference, constructing a continuous embedding space sensitive to frequency variations. FEI infers specific frequency sample directly in the embedding space using frequency masking prompts. Unlike contrastive learning, which models distinctions between semantics, FEI focuses on modeling the relationships between different semantics, achieving continuous modeling through a prompting strategy.

Overall, the contributions of this study include:

- This paper proposes FEI, a novel self-supervised representation learning framework for time series that eliminates the need for positive and negative sample pairs. Using frequency masking prompts, FEI performs embedding inference of different frequency bands in the embedding space, enabling continuous semantic modeling.
- We validate the quality of the representation through linear evaluation and end-to-end fine-tuning experiments. The proposed FEI achieves new state-of-the-art performance across 8 benchmark datasets for classification and regression tasks requiring high-level representations.
- This study demonstrates the feasibility and effectiveness of non-contrastive learning for time series self-supervised representation learning, providing new insights for further research in this area.

## Related Works

**Time series representation learning.** In recent years, contrastive learning has become a fundamental paradigm for time series high-level representation learning. For example, in addition to learning series reconstruction representations in the embedding space, SimMTM(Dong et al. 2023) also uses contrastive loss to constrain the representation distance between positive and negative sample pairs. TimeDRL(Chang et al. 2024) uses CLS tokens to obtain high-level representations and generates positive and negative sample pairs at different stages of subspace mapping through gradient truncation to guide the high-level representation optimization of CLS tokens. COMET(Wang et al. 2023) constructs multi-level contrast constraints to achieve a stable high-level representation learning process. Due to the discrete modeling nature of the contrastive learning framework, the construction of positive and negative sample pairs can significantly impact the optimization results. Inappropriate pairing frequently occur in some self-supervised representation learning approaches based on contrastive learning. For this issue, TS2Vec(Yue et al. 2022) proposing an enhanced contextual construction strategy to eliminate improper priors in positive and negative sample pairs. TimesURL(Liu and Chen 2024) and TF-C(Zhang et al. 2022) both construct positive and negative samples from a frequency domain perspective; the former mixes the spectra of multiple samples to generate new augmented samples, while the latter constructs augmented samples through spectral masking and enhancement. In addition, both Literature (Lan et al. 2024) and (Xu et al. 2024) have also investigated the issues related to contrastive learning in time series modeling. However, these methods still operate within the contrastive learning framework and cannot fundamentally address the issues inherent in the discrete modeling approach of contrastive learning.

**Joint-Embedding Predictive Architecture.** JEPA (LeCun 2022) is a self-supervised learning concept that introduces additional covariates in the latent space to guide the model in learning the relationships between multiple semantics. I-JEPA(Assran et al. 2023), a successful application of this concept, demonstrates that human priors are unnecessary in the self-supervised process. This provides a potential solution to fundamentally address the issue of inappropriate biases introduced by sample augmentation in self-supervised representation learning. JEPA has garnered widespread attention in CV(Mo and Yun 2024; Saito and Poovvancheri 2024) and time series forecasting(Verdenius, Zerio, and Wang 2024). However, JEPA has not yet been applied to time series representation learning.

## Frequency-masked Embedding Inference

The main architecture of the proposed FEI is illustrated in Figure 1. The core architecture of FEI consists of 2 sets of inference branches: **target embedding inference** based on mask prompts and **mask inference** based on target embedding prompts. The overall pre-training objective is to obtain a universal time series encoder $f_\theta$.

### Original Encoder

Given the original time series $x \in \mathbf{R}^l$ with length $l$, the encoder $f_\theta$ generates an embedding vector $e \in \mathbf{R}^d$ for $x$, where $d$ represents the embedding dimension of $f_\theta$. This process does not impose any restrictions on the specific structure of encoder $f_\theta$. After obtaining the embedding vector, a projector $s_\phi$ serves as a relaxation factor to reduce the training complexity of the subsequent inference process, is used to obtain the subspace embedding $u \in \mathbf{R}^h$, where $h = d/2$. In our implementation, $s_\phi$ is implemented by a single linear layer.

### Momentum Encoder

As FEI lacks explicit training constraints like those in contrastive learning architectures, directly performing embedding inference based on the single encoder may lead to representation collapse. A momentum encoder is an effective means to prevent this issue(Assran et al. 2023). Therefore, FEI employs smoothly updated copies, $f_{\theta'}$ and $s_{\phi'}$, of the original encoder $f_\theta$ and projector $s_\phi$ as the target series encoder. These copies are updated using an exponential moving average and do not directly participate in gradient calculation. The update process is as follows:

$$\theta'_t = \alpha * \theta'_{t-1} + (1 - \alpha) * \theta_t \tag{1}$$

where $\theta_t$ represents the weights of $f_\theta$ after the $t$-th gradient descent, $\theta'_t$ represents the weights of momentum encoder $f_{\theta'}$, and $\theta'_0 = \theta_0$. The same applies to subspace projector $s_{\phi'}$ and $s_\phi$. Similarly, the target series embedding $u'$ can be generated by $f_{\theta'}$ and $s_{\phi'}$.

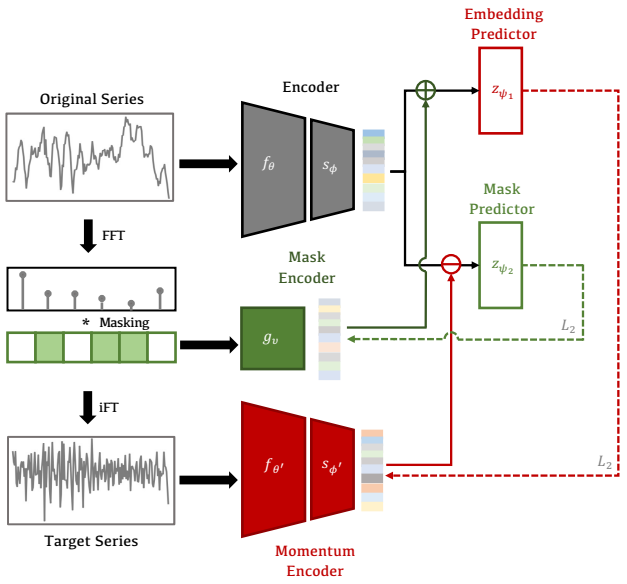


Figure 1: The overall structure of the proposed **FEI**. The original time series is fed into the encoder $f_\theta$ and a subspace projector $s_\phi$ to generate the original embedding. The target time series is constructed by applying random frequency masking, which is then fed into the momentum encoder—a smoothed copy of the original encoder updated via exponential moving average—to produce the target embedding. The goal of FEI is to enable the encoder to generate high-quality representation embeddings that can accurately infer the target embedding despite the presence of randomly masked frequency components(red dashed branch). Additionally, the representation embedding should also be capable of inferring mask embeddings by leveraging the differences between the target series and the original series (green dashed branch).

## Frequency Masking

To obtain the target series $x'$ corresponding to the original series $x$ while preserving its continuity, random frequency masking is employed. The original series $x$ undergoes a Fast Fourier Transform (FFT) to obtain its amplitude spectrum. A random mask $M \in \{0,1\}^n$ with $k$ mask positions is then generated, where $n = \lfloor \frac{l}{2} \rfloor + 1$. The masked frequency component positions are marked as 1, and the unmasked positions are marked as 0. This mask is applied to the amplitude spectrum of the original series, covering some of the frequency components, and the Inverse Fourier Transform (IFT) is then used to obtain the target time series $x'$ with masked frequencies.

It is important to note that the generation process of the frequency mask involves 2 steps. First, the number $k$ of masked frequency components is a random variable, with the masking ratio following a uniform distribution $U(\beta_1, \beta_2), 0 \leq \beta_1 < \beta_2 < 1$. This means that each masking operation covers $\beta_1$ to $\beta_2$ of the total frequency components of the original series, providing sufficiently rich variation semantics during training. Then, $k$ frequency components are randomly selected as the final mask. This design avoids the issue of semantic uniformity caused by using the consistent masking ratio, while avoiding the introduction of inappropriate biases from any fixed masking pattern into the model training process.

## Mask Encoder

To convert the frequency mask $M \in \{0,1\}^n$ into a embedding vector that can be used as an embedding space prompt, we design a dedicated mask encoder to generate the mask embedding $m$ that adheres to the following principles: when no frequency components are masked, i.e., $M = \{0\}^n$, the mask embedding $m$ should also be 0. Additionally, the mean and variance of $m$ should remain relatively stable as $k$ varies. The mask encoder proposed is as follows:

$$m = g_\upsilon(M) = \frac{MW_{emb}}{\sqrt{k}} \tag{2}$$

where $W_{emb} = \{w_1, ..., w_n\}^T \in \mathbf{R}^{n \times h}$ constructs an embedding matrix, with each $w_i$ representing an embedding vector for a frequency component. $W_{emb}$ is randomly initialized from a normal distribution $N(0,1)$ and updated during training.

## Embedding Inference

The embedding inference part consists of 2 branches: the target embedding inference and the mask inference. Both branches use 1-layer MLP-based predictors, but differ in their prompt embeddings and merging methods. When inferring the target embedding, the mask embedding $m$ used to generate the target series is employed. Conversely, when inferring the mask embedding, the target embedding $u'$ is utilized. The computational process for this section is as follows:

$$\hat{u}' = z_{\psi_1}(u + D(m)) \tag{3}$$

$$\hat{m} = z_{\psi_2}(D(u) - u') \tag{4}$$

where $D(\cdot)$ represents the gradient detaching, $\hat{u}'$ is the inferred target embedding based on mask prompting, and $\hat{m}$ indicates the inferred mask embedding based on the target embedding. Gradient detaching is used to force the model to focus on different optimization aspects during the two inference processes. In the process of inferring the target embedding, the gradient calculation of the mask embedding $m$ is detached, and thus the mask encoder $g$ is not optimized. The encoder $f_\theta$ is driven to obtain better original embedding $u$ to complete the inference. Conversely, in the process of inferring the mask embedding, the gradient calculation of the original embedding $u$ is detached, and only the mask encoder is optimized to obtain a better mask embedding. This design clarifies the optimization objectives of the 2 branches and avoiding gradient conflicts during the training process.

The design of the embedding inference is the core module for the effectiveness of FEI and embodies the original intention behind FEI's design. Firstly, we aim for the model to infer any changes in the series's frequency band within the embedding space using appropriate prompts. Secondly,

| Task | Dataset | Freq.(Hz) |
|---|---|---|
| Pre-training | SleepEEG | 100 |
| Classification | Gesture | 100 |
| | FD-B | 64k |
| | EMG | 4k |
| | EPI | 174 |
| | HAR | 50 |
| | 128 UCR | - |
| Regression | C-MAPSS | 1 |
| | Bearing | 25.6k |

Table 1: The datasets used in experiments.

we aim for the model to infer the frequency differences between the original embedding and the frequency-masked target embedding. They jointly guide the model in understanding frequency variations in time series.

### Loss Function

Based on the embedding inference results, the training loss is simply computed using the $L_2$ distance between the inferred target embedding $\hat{u}'$ and true target embedding $u'$, as well as between the inferred mask embedding $\hat{m}$ and the true mask embedding $m$, as follows:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|u'_i - \hat{u}'_i\|_2^2 + \|m_i - \hat{m}_i\|_2^2 \tag{5}$$

## Experiments

This section reports the transfer performance of FEI on classification and regression downstream tasks, along with an ablation study analyzing the effectiveness of the FEI design. The full experiment results and more details can be found in Appendix.

### Dataset

The dataset used in the experimental section is shown in Table 1. For the pre-training phase, we use the commonly utilized SLeepEEG dataset, which provides ample samples and is widely used for pre-training in various transfer learning methods (Zhang et al. 2022; Dong et al. 2023). For the downstream tasks, 8 publicly available datasets are utilized: 1) Gesture (Liu et al. 2009), 2) FD-B (Lessmeier et al. 2016), 3) EMG (Goldberger et al. 2000), 4) EPI(Andrzejak et al. 2001), 5) HAR (Anguita et al. 2013), 6) 128 UCR (Dau et al. 2018), 7) C-MAPSS (Saxena et al. 2008), and 8) Bearing (Wang et al. 2018). The first 6 datasets are commonly used for classification tasks, while the last 2 are typically used for regression tasks(Gao, Wen, and Wu 2021; Wang, Zhu, and Zhao 2024).

### Baselines

To fully verify the representation advantages of FEI, we selected the state-of-the-art methods from recent years as baselines for comparison. These methods are: 1) TimesURL (Liu and Chen 2024), 2) SimMTM (Dong et al. 2023), 3) InfoTS (Luo et al. 2023), 4) TimeDRL (Chang et al. 2024), 5) TF-C (Zhang et al. 2022), and 6) TS2Vec (Yue et al. 2022). All these methods are representative of the current time series representation learning field and are based on contrastive learning frameworks.

We reproduce these baselines using the official open-source code. To ensure a fair comparison of the generalization performance of pre-training methods, we use 1-D ResNet as the encoder network for all methods except TimesURL, with the embedding dimension set to 1024. The TimesURL method encounter an Out-of-Memory issue, preventing us from replacing it with a larger encoder network. However, we adjust the embedding dimension of TimesURL to be consistent with all other methods.

### Pre-training Setup

The primary significance of representation learning is to enable the pre-trained model to be applicable to as many unknown downstream tasks as possible. Therefore, our experimental process fully adheres to this principle: **after completing pre-training, the encoder is directly transferred to different downstream datasets without any additional pre-training or hyperparameter adjustments for each dataset**. A single linear layer is used as the task-specific output layer. This approach fully verifies the differences in the representation quality of each method.

The basic setup of proposed FEI for the pre-training process follows SimMTM and TF-C. Additional configurations can be found in the Appendix.

### Task 1: Classification

**Setup.** Time series classification is a key benchmark for evaluating representation learning algorithms. We perform linear evaluation on 6 commonly used datasets and conduct end-to-end fine-tuning on 5 datasets with limited training samples to validate the generalization capabilities of each method.

In the linear evaluation process, we freeze the encoder, and optimize only a linear classifier, with a maximum training iteration of 300 and an initial learning rate of 1e-4. In the end-to-end fine-tuning process, both the encoder and the classifier are optimized, with a maximum iteration of 100 and a smaller learning rate of 1e-5 to prevent overfitting. For both linear evaluation and end-to-end fine-tuning, all baseline encoders use the same hyperparameters. **The model with the lowest validation loss is used as the final test model** by early stopping for all datasets except for the 128 UCR dataset, which does not have a validation set division, so all models are tested directly after training ends.

We evaluate the accuracy, precision, recall, and F1 score of each method on classification tasks, with the accuracy results of linear evaluation shown in Table 2 and the fine-tuning results shown in Table 3.

In these tables, "Rand. Init." represents the performance of a randomly initialized encoder used directly for downstream tasks without pre-training.

| Datasets | Rand. Init. | TS2Vec | TimeDRL | TF-C | TimesURL | SimMTM | InfoTS | **FEI(Ours)** |
|---|---|---|---|---|---|---|---|---|
| Gesture | 12.50 | 63.33 | 50.00 | 57.50 | 69.72 | 74.17 | 64.17 | **75.00** |
| FD-B | 11.39 | 43.59 | 40.63 | 45.53 | 54.44 | 60.74 | 60.71 | **67.25** |
| EMG | 46.34 | **92.68** | 63.41 | 78.05 | **92.68** | 85.37 | 87.80 | 87.80 |
| EPI | 19.79 | 96.41 | 77.85 | 85.75 | 95.42 | 96.42 | 96.27 | **96.84** |
| HAR | 36.51 | 78.91 | 70.31 | 67.56 | 79.10 | 77.13 | 78.35 | **79.54** |
| 128 UCR | 39.03 | 72.50 | 61.61 | 61.88 | 69.53 | 75.34 | 73.13 | **78.17** |
| Avg. | 27.59 | 72.66 | 60.64 | 64.38 | 76.82 | 78.19 | 76.74 | **80.77** |

Table 2: The **linear evaluation** accuracy(%) of all methods on classification task.

| Datasets | Rand. Init. | TS2Vec | TimeDRL | TF-C | TimesURL | SimMTM | InfoTS | **FEI(Ours)** |
|---|---|---|---|---|---|---|---|---|
| Gesture | 68.33 | 72.50 | 73.33 | 70.00 | 73.33 | 76.67 | 71.67 | **77.50** |
| FD-B | 69.61 | 48.31 | 47.97 | 65.48 | 54.41 | 63.49 | 62.99 | **70.99** |
| EMG | 95.12 | 78.05 | 78.05 | 92.68 | 73.17 | 87.80 | **97.56** | **97.56** |
| EPI | 80.21 | 95.60 | 94.05 | 95.28 | 96.67 | 96.22 | 97.07 | **97.24** |
| 128 UCR | 72.44 | 67.44 | 63.36 | 78.50 | 79.40 | 80.42 | 81.77 | **82.65** |
| Avg. | 77.14 | 73.62 | 71.69 | 80.86 | 74.40 | 81.05 | 82.32 | **85.82** |

Table 3: The **end-to-end fine-tuning** accuracy(%) of all methods on small-sample classification datasets.

**Analysis.** From the results, it is evident that the proposed FEI achieves the best transfer performance in almost all cases. In linear evaluation scenarios, FEI achieves an average accuracy improvement of 2.15% over existing methods on the compared datasets. In end-to-end fine-tuning, most models show varying degrees of improvement compared to linear evaluation results. The encoder trained by the proposed FEI still achieves the best average performance, with an average accuracy improvement of approximately 3.50% on the 5 small-sample datasets compared to the best existing methods. The proposed FEI adopts an inference-based modeling approach for different frequency bands, fully leveraging the semantics of each frequency band in the pretraining dataset. This results in a broader learning corpus compared to contrastive-based methods, significantly enhancing the robustness of the encoder to variations in data frequency.

## Task 2: Regression

**Setup.** Regression tasks are a classic type of time series task. In this paper, we use 2 equipment health status analysis datasets, C-MAPSS and Bearing, to analyze the transfer performance of FEI in regression tasks. These datasets consist of 4 and 3 sub-datasets, respectively, each using input signal to regress the Remaining Useful Life (RUL) ratio of the equipment. These datasets are commonly used for time series regression tasks in equipment health status analysis.

Moreover, the sampling frequency of these 2 datasets significantly differs from the pre-training dataset SleepEEG. The C-MAPSS dataset has a very low sampling frequency (1Hz), while the Bearing dataset has a very high sampling frequency (25.6Hz). This difference allows for a better evaluation of the model's generalization ability across different data characteristics.

The experimental settings for linear evaluation and end-to-end fine-tuning are the same as for classification tasks. Due to the smaller number of training samples in the Bearing dataset compared to the C-MAPSS dataset, end-to-end fine-tuning experiments are conducted on the Bearing dataset. Mean Squared Error (MSE) and Mean Absolute Error (MAE) are used as performance metrics, with results shown in Tables 4 and 5, respectively. In the tables, all performance results are averaged across all sub-datasets in each dataset.

| Datasets | Methods | MSE | MAE |
|---|---|---|---|
| C-MAPSS | Rand. Init. | 0.0714 | 0.2439 |
| | TS2Vec | 0.1172 | 0.2679 |
| | TimeDRL | 0.0644 | 0.2185 |
| | TF-C | 0.4499 | 0.5612 |
| | TimesURL | 0.1044 | 0.2566 |
| | SimMTM | 0.0599 | 0.2056 |
| | InfoTS | 0.0623 | 0.2145 |
| | **FEI(Ours)** | **0.0584** | **0.1992** |
| Bearing | Rand. Init. | 0.5274 | 0.6705 |
| | TS2Vec | 0.1268 | 0.1972 |
| | TimeDRL | 0.0919 | 0.1875 |
| | TF-C | 0.6022 | 0.6447 |
| | TimesURL | 0.0782 | 0.1813 |
| | SimMTM | 0.0959 | 0.1983 |
| | InfoTS | 0.0829 | 0.1905 |
| | **FEI(Ours)** | **0.0744** | **0.1747** |

Table 4: The average **linear evaluation** results of all methods on regression task.

| Datasets | Methods | MSE | MAE |
|---|---|---|---|
| Bearing | Rand. Init. | 0.0728 | 0.1844 |
| | TS2Vec | 0.1837 | 0.2649 |
| | TimeDRL | 0.1467 | 0.2885 |
| | TF-C | 0.6040 | 0.6446 |
| | TimesURL | 0.1641 | 0.2561 |
| | SimMTM | 0.0539 | 0.1397 |
| | InfoTS | 0.0732 | 0.1776 |
| | **FEI(Ours)** | **0.0331** | **0.1150** |

Table 5: The **end-to-end fine-tuning** results of all methods on small-sample regression dataset.

**Analysis.** As shown in the table, the proposed FEI method also achieves the best average performance in regression tasks. On the C-MAPSS and Bearing datasets, which exhibit significant semantic differences from the pre-training data, the TimesURL method improves the strategy for constructing positive and negative samples. Compared with TS2Vec, TimeDRL, and TF-C, its performance improves significantly. However, its MSE and MAE are still 9.31% and 8.83% higher than those of FEI on the C-MAPSS dataset, respectively. The FEI method completely abandons explicit negative sample constraints and instead establishes flexible correlations of frequency band features between samples.

## Model Analysis

**Ablation.** To further explore the role of each module in the proposed FEI, an ablation analysis is conducted. We target the core modules of FEI for ablation, constructing 6 ablation models:

- $w/o$ **emb. infer.**, which removes the target embedding inference branch, retaining only the mask inference.
- $w/o$ **mask prompt**, which removes the mask prompting for target series embedding inference.
- $w/o$ **momentum**, which removes the momentum encoder and directly uses the original encoder for encoding.
- $w/o$ **subspace**, which removes the subspace projector $s_\phi, s'_\phi$ and directly learns in the original embedding space.
- $w/o$ **mask infer.**, which removes the mask inference branch, retaining only the target embedding inference.
- $w/o$ **detach**, which removes the gradient detachment $D(\cdot)$ in Equations (3) and (4) training.

The linear evaluation precision of each ablation method on the EMG dataset is shown in Table 6.

Due to the design of the mask inference branch, removing the momentum encoder from FEI does not lead to significant representation collapse. This is because trivial solutions, where the encoder encodes all samples into the same embedding vector, cannot minimize the loss of the mask inference branch.

Additionally, the mask inference stabilizes the FEI training process, as shown by the loss value comparison over the first 20 epochs in Figure 2. The colored lines represent the loss of target embedding inference, while the gray lines represent the overall loss (target embedding inference + mask inference) with $w/$ mask infer. The mask inference provides a clearer optimization goal for the mask encoder $g_v$, and the smooth loss descent process indicates a flatter loss surface and better generalization(Petzka et al. 2021).

| Model | Prec.(%) |
|---|---|
| **FEI** | **90.20** |
| $w/o$ emb. infer. | 57.69 ↓32.51 |
| $w/o$ mask prompt | 62.75 ↓27.45 |
| $w/o$ momentum | 62.75 ↓27.50 |
| $w/o$ subspace | 51.60 ↓38.60 |
| $w/o$ mask infer. | 79.24 ↓10.96 |
| $w/o$ detach | 85.70 ↓4.50 |

Table 6: The ablation results of FEI on EMG dataset.

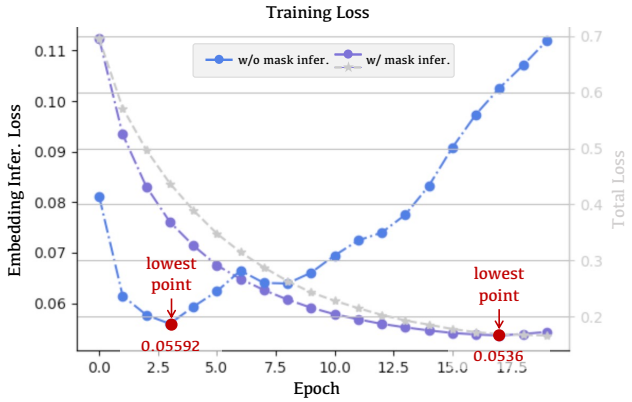


Figure 2: Comparison of training loss curves for the first 20 epochs $w/$ and $w/o$ mask inference.

**Visualization of embedding inference.** Embedding inference is a core module of FEI. In this section, we select a original series sample from Gesture dataset and construct 5 representative target series using random masks. **It should be noted that FEI has never been trained on this dataset.** Using t-SNE(Van der Maaten and Hinton 2008), we visualize the original embedding and target embeddings after dimensionality reduction, as shown in Figure 3. The left side displays the relationship between the embeddings of the original series and the target series, with inverted triangles representing the target embeddings obtained directly using the original encoder $f_\theta$ and subspace projector $s_\phi$, and stars representing the inferred embeddings obtained by the embedding predictor $z_{\psi_1}$. The right side shows the 5 masks used to construct the target series, where the dark color represents the masked frequency component. From the figure, it can be observed that the green target series has the fewest masks, and its embedding result and inference are close to the original embedding. The red and yellow target series have masks mainly in the mid and high-frequency bands, with only a small amount in the low-frequency band, resulting in their embeddings being further from the orig-

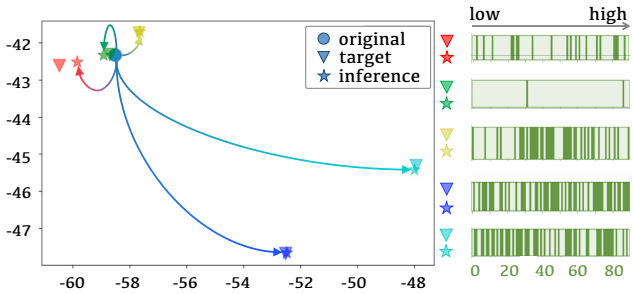


Figure 3: Visualization of embedding inference results on Gesture dataset. The left side shows the inference results, and the right side shows the masks used to construct the target series.

inal embedding. The cyan and blue target series have the highest masking ratio and are significantly farther from the original embedding. FEI accurately complete the embedding inference for previously unseen series. The modeling approach of FEI encourages the encoder to establish a **continual frequency-sensitive embedding space**, significantly enhancing its general representation ability. Additional visualizations and details are available in the Appendix.

**Masking Strategies.** To construct target sequences, we experiment with various masking strategies in the design of FEI, specifically:

- Discrete Frequency Masking (DFM): Completely random frequency band masking strategy (used in FEI).
- Continuous Frequency Masking (CFM): Continuous frequency masking with random start and end points.
- Time-domain Masking (TDM): Random masking directly in the series's time domain.

We compare the performance differences of these masking strategies, and their accuracy results on the FD-B dataset are shown in Table 7.

The key difference between frequency-domain masking(DFM, CFM) and time-domain masking(TDM) lies in the nature of the information loss they produce. Frequency-domain masking creates target series with soft information loss, where the semantic differences between the target and original series change continuously with the position and amount of frequency masking. In contrast, time-domain masking generates target series with hard information loss, causing semantic discontinuities in the time domain. We found that TDM leads to better transfer performance of FEI on datasets with the similar frequency (e.g., Gesture), but its performance significantly declines on datasets with different frequencies, particularly on the FD-B dataset with large frequency differences, which severely reduces FEI's generalization ability.

Similarly, the SimMTM method, also based on time-domain masking, exhibits similar characteristics. As shown in Tables 2 and 4, SimMTM demonstrates excellent transfer performance on datasets with the same frequency as the pre-training dataset (e.g., Gesture). However, its transfer performance significantly deteriorates on datasets with

| Masking Strategy | FD-B | Gesture |
|---|---|---|
| DFM | **67.25** | 75.00 |
| CFM | 60.65 ↓ 6.60 | 71.67 ↓ 3.33 |
| TDM | 54.02 ↓ 13.23 | **79.17** ↑ 4.17 |

Table 7: The main results of different masking strategies on FD-B and Gesture datasets.

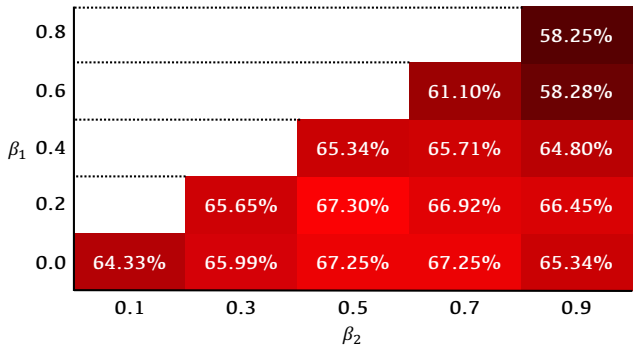


Figure 4: Accuracy results of FEI on the FD-B dataset under different masking ratios. The random masking ratio generated during each FEI training session lies between $\beta_1$ and $\beta_2$, as described in **Section Frequency Masking**.

frequencies that differ greatly from the pre-training dataset (e.g., EMG, FD-B, Bearing). **This seems to suggest that time-domain masking, as a method for non-continuous semantic modeling, is more easily transferable to data with the same frequency but struggles to train models with strong generalization ability.** This phenomenon provides new insights for selecting sample augmentation strategies (time-domain processing, frequency-domain processing) in time series modeling. This further emphasizes the importance of continuous semantic modeling methods for achieving robust time series representations. Additional results can be found in the Appendix.

**Sensitivity** We further analyze the impact of different masking ratios on the performance of FEI. Figure 4 shows the linear evaluation accuracy of FEI under different masking ratios on the FD-B dataset. Ultimately, we use $\beta_1 = 0.0$ and $\beta_2 = 0.7$. More sensitivity analysis can be found in the Appendix.

## Conclusion

This paper introduces FEI, a method for modeling the continuous semantic relationships of time series without the need for complex positive and negative sample pair construction. By inferring between different frequency-masked samples within the embedding space, FEI overcomes the limitations of contrastive learning. Experimental results demonstrate that the encoder trained by FEI achieves superior generalization performance compared to existing contrastive learning-based methods.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 62273038 and U21A20483, and in part by the Innovative Talent Training Foundation for University of Science and Technology Beijing.

## References

Andrzejak, R. G.; Lehnertz, K.; Mormann, F.; Rieke, C.; David, P.; and Elger, C. E. 2001. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6): 061907.

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J. L.; et al. 2013. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, 3.

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15619–15629.

Chang, C.; Chan, C.-T.; Wang, W.-Y.; Peng, W.-C.; and Chen, T.-F. 2024. TimeDRL: Disentangled Representation Learning for Multivariate Time-Series. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 625–638. IEEE.

Dau, H. A.; Keogh, E.; Kamgar, K.; Yeh, C.-C. M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C. A.; Yanping; Hu, B.; Begum, N.; Bagnall, A.; Mueen, A.; Batista, G.; and Hexagon-ML. 2018. The UCR Time Series Classification Archive. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/. Accessed: 2024-08-01.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dong, J.; Wu, H.; Zhang, H.; Zhang, L.; Wang, J.; and Long, M. 2023. SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 29996–30025. Curran Associates, Inc.

Fang, H.; Wang, S.; Zhou, M.; Ding, J.; and Xie, P. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Gao, Y.; Wen, Y.; and Wu, J. 2021. A Neural Network-Based Joint Prognostic Model for Data Fusion and Remaining Useful Life Prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1): 117–127.

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

Lan, X.; Yan, H.; Hong, S.; and Feng, M. 2024. Towards Enhancing Time Series Contrastive Learning: A Dynamic Bad Pair Mining Approach. In *The Twelfth International Conference on Learning Representations*.

LeCun, Y. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *OpenReview*.

Lessmeier, C.; Kimotho, J. K.; Zimmer, D.; and Sextro, W. 2016. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *PHM Society European Conference*, volume 3.

Li, Z.; Chen, Z.; Yang, F.; Li, W.; Zhu, Y.; Zhao, C.; Deng, R.; Wu, L.; Zhao, R.; Tang, M.; et al. 2021. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34: 13165–13176.

Liu, J.; and Chen, S. 2024. TimesURL: Self-Supervised Contrastive Learning for Universal Time Series Representation Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12): 13918–13926.

Liu, J.; Zhong, L.; Wickramasuriya, J.; and Vasudevan, V. 2009. uWave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6): 657–675.

Luo, D.; Cheng, W.; Wang, Y.; Xu, D.; Ni, J.; Yu, W.; Zhang, X.; Liu, Y.; Chen, Y.; Chen, H.; and Zhang, X. 2023. Time Series Contrastive Learning with Information-Aware Augmentations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4): 4534–4542.

Mo, S.; and Yun, S. 2024. DMT-JEPA: Discriminative Masked Targets for Joint-Embedding Predictive Architecture. *arXiv preprint arXiv:2405.17995*.

Ozyurt, Y.; Feuerriegel, S.; and Zhang, C. 2023. Contrastive Learning for Unsupervised Domain Adaptation of Time Series. In *The Eleventh International Conference on Learning Representations*.

Petzka, H.; Kamp, M.; Adilova, L.; Sminchisescu, C.; and Boley, M. 2021. Relative flatness and generalization. *Advances in neural information processing systems*, 34: 18420–18432.

Saito, A.; and Poovvancheri, J. 2024. Point-JEPA: A Joint Embedding Predictive Architecture for Self-Supervised Learning on Point Cloud. *arXiv preprint arXiv:2404.16432*.

Saxena, A.; Goebel, K.; Simon, D.; and Eklund, N. 2008. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management*, 1–9. IEEE.

Tonekaboni, S.; Eytan, D.; and Goldenberg, A. 2021. Unsupervised Representation Learning for Time Series with Temporal Neighborhood Coding. In *International Conference on Learning Representations*.

Trirat, P.; Shin, Y.; Kang, J.; Nam, Y.; Na, J.; Bae, M.; Kim, J.; Kim, B.; and Lee, J.-G. 2024. Universal Time-Series Representation Learning: A Survey. *arXiv preprint arXiv:2401.03717*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Verdenius, S.; Zerio, A.; and Wang, R. L. 2024. LaT-PFN: A Joint Embedding Predictive Architecture for In-context Time-series Forecasting. *arXiv preprint arXiv:2405.10093*.

Wang, B.; Lei, Y.; Li, N.; and Li, N. 2018. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*, 69(1): 401–412.

Wang, L.; Zhu, Z.; and Zhao, X. 2024. Dynamic predictive maintenance strategy for system remaining useful life prediction via deep learning ensemble method. *Reliability Engineering & System Safety*, 245: 110012.

Wang, Y.; Han, Y.; Wang, H.; and Zhang, X. 2023. Contrast Everything: A Hierarchical Contrastive Framework for Medical Time-Series. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 55694–55717. Curran Associates, Inc.

Xu, M.; Moreno, A.; Wei, H.; Marlin, B.; and Rehg, J. M. 2024. REBAR: Retrieval-Based Reconstruction for Time-series Contrastive Learning. In *The Twelfth International Conference on Learning Representations*.

Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; and Xu, W. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5065–5075.

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2022. TS2Vec: Towards Universal Representation of Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8): 8980–8987.

Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; and Eickhoff, C. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2114–2124.

Zhang, X.; Zhao, Z.; Tsiligkaridis, T.; and Zitnik, M. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35: 3988–4003.