# Masking Strategy Variations in Self-Distilled and Contrastive Time-Series Model Convergence and Accuracy

Assignee Research

June 11, 2026

**Abstract**

In recent years, the introduction of self-supervised contrastive learning (SSCL) has demonstrated remarkable improvements in representation learning across various domains, including natural language processing and computer vision. By leveraging the inherent benefits of self-supervision, SSCL enables the pre-training of representation models using vast amounts of unlabeled data. Despite these advances, there remains a significant gap in understanding the impact of different SSCL strategies on time series forecasting performance, as well as the specific benefits that SSCL can bring. This paper

## 1 Introduction

This paper examines: What Constitutes Good Contrastive Learning in Time-Series Forecasting?. Research question: What is the impact of masking strategy variations on the convergence speed and final accuracy of self-distilled versus contrastive time-series models?.

## 2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

## 3 Results

13 papers retrieved. 17 claims extracted; 13 independently verified. Quality review score: 7.4/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

| Claim | Verified | Confidence |
|---|---|---|
| The ETT datasets consist of data from six power load features and measurements of oil temperature. | ✓ | 0.23 |
| The ECL dataset contains the electricity consumption of 312 clients. | ✓ | 0.16 |
| The ECL dataset was converted into hourly-level measurements, following previous work (Yue et al., 2022). | ✓ | 0.28 |
| The LSTM-based encoder includes five unidirectional LSTM layers with a hidden dimension of 320 in each layer. | ✓ | 0.31 |
| The LSTM-based encoder architecture has a total of 660K learnable parameters. | ✓ | 0.26 |
| The TCN-based encoder consists of 10 dilated convolution layers with a hidden dimension of 320 and a kernel size of 3 pe | ✓ | 0.33 |
| The TCN-based encoder architecture has a total of 637K learnable parameters. | ✓ | 0.26 |
| The Transformer-based encoder contains 5 Informer layers with a hidden size of 128 and 8 attention heads. | ✓ | 0.27 |
| The Transformer-based encoder architecture has a total of 655K learnable parameters. | ✓ | 0.26 |
| All three models utilize a linear layer as the first embedding layer to map the input features into hidden dimensions. | ✓ | 0.26 |
| All three models are trained with a peak learning rate of 0.001. | × | 0.13 |
| A cosine learning rate scheduler is used for MoCo2-framework models. | × | 0.15 |
| Each model trains for 30 epochs with early stopping based on the Dev performance to prevent overfitting. | ✓ | 0.25 |
| For two-step learning, an encoder is trained on the training set with SSCL for 600 training iterations. | ✓ | 0.20 |
| When fine-tuning a pre-trained encoder, the same training hyper-parameters as in end-to-end training are used. | ✓ | 0.19 |
| The MSE is used as the main evaluation metric, while MAE results are also reported. | × | 0.12 |
| The overall performance of each model is represented by averaging the MSE across all prediction lengths and the three da | × | 0.14 |

## References


- http://arxiv.org/abs/2405.08815v1
- http://arxiv.org/abs/2306.10125v4
- http://arxiv.org/abs/2306.12086v2