# Rotary Masked Autoencoders are Versatile Learners

**Uros Zivanovic**[1], **Serafina Di Gioia**[2, 3], **Andre Scaffidi**[3], **Martín de los Rios**[3], **Gabriella Contardo**[7, 3], and **Roberto Trotta**[3, 4, 5, 6]

[1]University of Trieste, Italy
[2]Abdus Salam International Centre for Theoretical Physics (ICTP), Italy
[3]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Italy
[4]INFN – National Institute for Nuclear Physics, Italy
[5]ICSC - Centro Nazionale di Ricerca in High Performance Computing, Italy
[6]Imperial College London, United Kingdom
[7]University of Nova Gorica, Slovenia

## Abstract

Applying Transformers to irregular time-series typically requires specializations to their baseline architecture, which can result in additional computational overhead and increased method complexity. We present the Rotary Masked Autoencoder (RoMAE), which utilizes the popular Rotary Positional Embedding (RoPE) method for continuous positions. RoMAE is an extension to the Masked Autoencoder (MAE) that enables interpolation and representation learning with multidimensional continuous positional information while avoiding any time-series-specific architectural specializations. We showcase RoMAE's performance on a variety of modalities including irregular and multivariate time-series, images, and audio, demonstrating that RoMAE surpasses specialized time-series architectures on difficult datasets such as the DESC ELAsTiCC Challenge while maintaining MAE's usual performance across other modalities. In addition, we investigate RoMAE's ability to reconstruct the embedded continuous positions, demonstrating that including learned embeddings in the input sequence breaks RoPE's relative position property.

## 1 Introduction

The framework offered by Foundation models (FM) has shifted the machine learning landscape by establishing new benchmarks on a variety of modalities and tasks. Specifically, Transformers [58] have achieved state-of-the-art performance across many domains, from vision [15] to natural language processing [61]. Given the ability of Transformers to handle sequential data such as natural language, they naturally became appealing for time-series, which arise in a large variety of domains, including health, finance and astrophysics. Such data can often be irregularly sampled in the temporal dimension. Being originally designed for sequences of text, the base Transformer architecture is not able to deal with such irregularly sampled data, by default only supporting quantized positional information as is found in natural language. This lack of support for continuous positional information becomes a limitation when extending Transformers to other modalities, degrading performance on tasks requiring precise temporal modelling and hindering the model's ability to capture complex patterns in non-uniformly sampled time-series.

Various specializations to the Transformer have been proposed to overcome this limitation. These can be divided into two main types: modifications of the internal architecture of the Transformer (e.g. modifying the feedforward layer in the Transformer Encoder Block [19, 46]) and novel positional embeddings (e.g. using a neural ODE [12]). Alternatively, modern State Space Models like

Preprint. Under review.

Mamba [22] or S5 [51] are natively able to model various modalities such as text and images in addition to irregular time-series. Extending the capability of Transformers to irregular time-series while staying within existing frameworks developed for fields such as Natural Language Processing (NLP) and computer vision would allow one to easily benefit from ongoing developments within the Transformer "ecosystem".

To this end, we propose a new representation learning method utilizing Rotary Positional Embeddings (RoPE) [54] for continuous position in combination with Masked Autoencoder (MAE) [23] pre-training: the Rotary Masked Autoencoder (RoMAE). We investigate the performance of this framework on a variety of tasks with different modalities. RoMAE obtains highly competitive results when compared to specialized, state-of-the-art approaches for individual tasks while maintaining MAE's excellent performance in computer vision and audio. RoMAE is therefore extremely versatile while being built up from only standard Transformer methods.
Our contributions are threefold:

1. **Continuous Positional Embedding with RoPE:** We investigate how RoPE can be used to embed continuous positions, expanding on its original concept introduced in the *RoFormer* architecture [54], which did not address non-uniform or real-valued timestamps. We show that learned input embeddings can invalidate RoPE's relative distance guarantees, and we provide new empirical insights into RoPE's ability to embed varying positional scales.
2. **RoMAE**: An expansion of MAE that works natively with irregular multivariate time-series without sacrificing any performance on standard modalities such as images and audio. Utilizing standard off-the-shelf methods developed for Transformers in NLP and Computer Vision, RoMAE shows that a specialized architecture is not required to achieve strong performance on irregular time-series.
3. **Experimental Analysis**: We compare RoMAE with state-of-the-art deep learning (DL) models, conducting experiments on the following tasks and modalities: (i) irregularly sampled multi-variate time-series classification, (ii) image classification, (iii) irregularly sampled time-series interpolation and (iv) audio classification.

This work is structured as follows: Section 2 covers related works on MAE, RoPE, and irregular time-series. Section 3 provides the necessary background material. Section 4 details the workings of RoMAE and establishes the theoretical framework we use to tackle a variety of modalities. Section 5 presents the results of our experiments. Section 6 discusses our results. Section 7 concludes the paper.

## 2 Related Work

**Rotary Positional Embeddings**: RoPE was initially proposed in RoFormer [54] as a simple and effective Relative Position Embedding (RPE) method that is independent from the Multi-Head Attention (MHA) implementation being used. Despite RoPE encoding relative position, models incorporating RoPE have been shown not to generalize to sequences longer than the ones shown during training. To improve sequence length extrapolation, various works have proposed increasing RoPE's base wavelength from 10 000 up to 500 000 [63, 43, 62]. Alternatively, YaRN [38] proposes to interpolate RoPE's frequencies $\theta_i$ during inference to avoid out-of-distribution angles. There has also been recent discussion on the usefulness of the long-term decay property of RoPE [4]. In this work, we provide additional experimental evidence supporting the argument against the long-term decay of RoPE.

**RoPE in Vision Transformers**: 2D Hand-crafted and learned Absolute Positional Embedding (APE) methods have both been shown to give similar performance on Computer Vision benchmarks when used with Vision Transformers (ViT) [15]. Later works [25, 34] have shown the impact of RoPE on multi-resolution inference, finding that RoPE improves ViT's extrapolation performance. To extend RoPE to 2D, Axial RoPE [17] applies RoPE twice, once for each dimension. Additional learnable parameters can be added to Axial RoPE to encode diagonal positional information as well [25]. Taking advantage of RoPE's independence from the specific Multi-Head Attention (MHA) implementation, Vision X-former [29] (ViX) is a variant of ViT that utilizes RoPE with linear MHA implementations, making it more computationally efficient. Overall, adding RoPE to ViT has been shown to be a beneficial change, improving model performance, and enabling extrapolation to higher resolutions during inference.

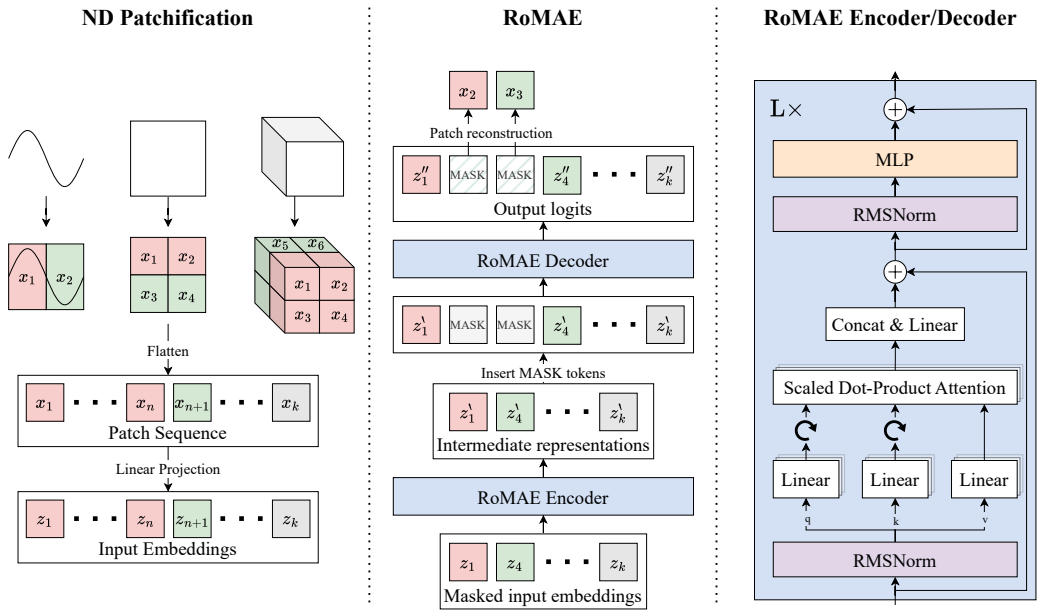


Figure 1: Overview of the RoMAE pipeline. **Left**: Visualisation of data embedding via multi-dimensional (ND) patchification for illustrative data realisations in 1, 2 and 3D. **Centre**: Full depiction of RoMAE architecture. The optional `[CLS]` token is omitted from the input sequence for simplicity. **Right**: The RoMAE encoder/decoder with ROPE operations denoted by rotational arrows.

**Masked Autoencoders**: Architectures such as BERT [14] and GPT [40, 41, 6] have shown that self-supervised pre-training tasks greatly boost the downstream fine-tuning performance of Transformers in NLP. MAE [23] is an approach that adapts BERT's masked modelling pre-training task to images. Although MAE is not the first work to conduct Masked Image Modelling, it is one of the most widely used methods. MAE has been shown to be both data-efficient [57] and scalable [23, 59]. It has also been adapted successfully to a variety of modalities other than images such as video [57, 18], audio [28], and point clouds [55]. MAE has also been combined with RoPE in MAETok [8], which uses the trained model as a tokenizer for diffusion models. Although MAE has been used in a variety of contexts, we highlight that many of these contexts have required task-specific specializations to the backbone Transformer architecture, unlike the approach we present here.

**Transformers for Irregular Time-series**: Adapting the Transformer architecture for irregular time-series is a long-standing research topic with many methods having been proposed [60]. When tokenizing the input, a popular approach is to insert the time for each point through positional embeddings – an approach recently used by models such as ContiFormer [12], Timer [33], and the concurrent work TrajGPT [52], which, similar to RoMAE, uses RoPE for the task. Alternatively, one can also convert the time-series data into 2D images and process them using an off-the-shelf ViT [32]. As a pre-training task, methods focus either on autoregressive trajectory prediction [52, 33], or BERT-style interpolation through masked modelling [37]. MAE has not been used for irregular time-series pre-training yet.

## 3 Background

**Attention:** We follow the formulation for Attention proposed in the original Transformer [58]. Let $\mathbf{z}$ be a sequence of $k$ embeddings $z_i \in \mathbb{R}^{d_{\text{model}}}$ for $i \in [1, 2, ..., k]$ and $d_{\text{model}}$ be the dimensionality of the model. Three linear layers are applied to transform $\mathbf{z}$ into sequences containing queries ($q_i$), keys ($k_i$), and values ($v_i$). E.g., $q_i = W_q z_i$, $k_i = W_k z_i$, $v_i = W_v z_i$. The matrices $Q, K, V$ containing $q$,$k$,$v$ are then passed through Scaled Dot-Product Attention (SDPA) as defined in Equation (1):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{\text{model}}}}\right)V \tag{1}$$

The key operation in SDPA is the dot product between $q_i$ and $k_i$, which determines how the values $V$ will be mixed with one-another. Because Attention is permutation-invariant, positional information must be encoded within $q_i$ and $k_i$ to allow SDPA to reason about position.

**Regular and Irregular Dimensions:** RoMAE is designed to work with both regular and irregular dimensional data. Specifically, we consider inputs of the form: $\mathbf{x} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_D}$ where $D$ is the number of dimensions in $\mathbf{x}$ and $d_i$ is the size of dimension $i$ for $i \in [1, \cdots, D]$.

**Definition 3.1** (Regular and irregular dimensions)**.** A *regular* dimension is one where all points are equally spaced, while an *irregular* dimension is one where the distance between points varies.

Some dimensions in the input $\mathbf{x}$ may be irregular while others could be regular. For example, a set of images sampled at irregular times has height and width as two regular dimensions and time as one irregular dimension.

### 3.1 Rotary Positional Embeddings

Given input $x_m \in \mathbb{R}^{d_x}$ with position $m$ and even dimensionality $d_x$, RoPE partitions $x_m$ into disjoint 2D subspaces $x_m^{(i)}$ with $i \in [1, 2, ..., d_x/2]$, rotating each subspace as:

$$\begin{pmatrix} \cos\ m\theta_i & -\sin\ m\theta_i \\ \sin\ m\theta_i & \cos\ m\theta_i \end{pmatrix} x_m^{(i)} \tag{2}$$

The $\theta_i$'s are generated in the same way as for sinusoidal positional embeddings [58]: $\theta_i = 10000^{-2(i-1)/d_x}$. Therefore, each subspace is rotated by a different amount depending on the individual $\theta_i$. RoPE is applied directly to the queries and keys before they enter SDPA. Because the dot product relies only on the angle between two vectors and their magnitude, RoPE is a Relative Positional Embedding (RPE) method.

**$p$-RoPE:** In this work we make use of $p$-RoPE [4], a truncated version of RoPE where only the $p$ percent of smallest $\theta_i$ values are kept. This cuts out a fraction $(1-p)$ of the frequencies in RoPE and thus leaves a portion of the input embedding space unchanged. The unchanged region in the embedding space provides the model with a data channel it can use to pass information into SDPA without any modifications by RoPE, making the model more robust to varying sequence length. We use a value $p = 0.75$, which has been shown by Barbero et. al. [4] to work well.

**Axial RoPE:** To encode multi-dimensional position, we utilize Axial RoPE [17]. In Axial RoPE, the input is split into $D$ subspaces of size $d_{\text{model}}/D$, where $D$ is the number of positional dimensions. Then we apply $p$-RoPE to each subspace, encoding the positional value for that dimension. We note that since RoPE requires that embeddings be even and Axial RoPE requires that embeddings be divisible by $D$, this puts constraints on the possible values that $d_{\text{model}}$ can take.

## 4 Method

An overview of RoMAE is shown in Figure 1.

**N-Dimensional (ND) Patchification:** Given inputs $\mathbf{x}$ as described in Section 3, we define a patch size $(p_1, \cdots, p_D)$ and divide each dimension into $N_i = d_i/p_i$ non-overlapping segments, where $d_i$ is the size of dimension $i$. These are flattened, creating a sequence of patches with length $k = \prod_{i=1}^{D} N_i$ and number of elements per patch $n_p = \prod_{i=1}^{D} p_i$. Finally, a linear layer $W^{n_p \times d_{\text{model}}}$ is applied to each patch to project it into the embedding dimension $d_{\text{model}}$. This process is illustrated in Figure 1 and is the same as employed by ViT [15] for images and ViViT [1] for video. It can also be understood as using non-overlapping $N$-dimensional convolutions with the number of channels equal to $d_{\text{model}}$. After ND-patchification, we have a sequence of embeddings $\mathbf{z}$, which can be passed into the RoMAE Encoder.

**Proposition 4.1.** For any *irregular* dimension $d_i$ in $\mathbf{x}$, the corresponding patch size for that dimension $p_i$ must be equal to 1.

**Discussion:** This limitation emerges from the requirement that each patch has the same number of data points $n_p$ inside it. Note that mixing irregular and regular dimensions is not an issue. E.g., for an

irregularly-sampled time-series of images, one could choose a patch size of (1, 16, 16) for the time, height, and width respectively.

Because the ND-patchification process flattens all dimensions into a single sequence, RoMAE is able to jointly model and attend to all patches across all dimensions at once. The drawback of this is that the number of tokens grows exponentially with the number of dimensions. In this work we only utilize this process up to $D = 3$. For highly dimensional multivariate time-series, we utilize a different approach that is described in Section 4.2.

### 4.1 Overall Structure

RoMAE's structure follows MAE's, using an asymmetric encoder/decoder, with the encoder being much larger than the decoder. Although both the encoder and decoder in RoMAE are a Transformer Encoder similar to BERT [14], we bring over recent developments in NLP from models such as Llama [43]. Specifically, we utilize the popular Sigmoid Linear Unit [24, 16] for the non-linearity. We also use RMSNorm [66] instead of Layer Normalization [2], as it has been shown to be more computationally efficient while maintaining the same level of model convergence. For architectural details, including definitions of various model sizes, we refer to Appendix A.2.

**Pre-Training Task:** Given a sequence of input patches, we uniformly mask a percentage of them. After projecting the unmasked patches into embeddings, a learned `[CLS]` token is optionally appended to the start of the sequence. This token becomes useful during fine-tuning, when an MLP head can be placed on top of it to conduct classification. The embeddings are then passed through the RoMAE Encoder to create an intermediate representation $\mathbf{z}'$. A set of learnable `[MASK]` tokens is then appended to $\mathbf{z}'$, with each `[MASK]` token receiving the positional information corresponding to a patch that was masked out. This sequence is then passed through the RoMAE Decoder, after which the model head predicts $n_p$ values for each patch that was masked out. After pre-training, the decoder is removed and the intermediate representations $\mathbf{z}'$ are used for down-stream tasks.

### 4.2 Positional Information in RoMAE

**Continuous Axial RoPE:** Although RoPE is originally designed to be used with discrete positions such as those found in text, we observe that Equation (2) works with any $m \in \mathbb{R}$. We make use of this in RoMAE to encode continuous position. Specifically, alongside the input values $\mathbf{x}$, RoMAE also accepts a sequence $\mathbf{s} = [s_1, \cdots, s_k],\ s_i \in \mathbb{R}^D$, containing the positional information for each patch. This is then applied within RoMAE using Axial RoPE as described in Section 3.1.

**Dealing With Many Irregular Dimensions:** Although we are able to encode multi-dimensional positional information using Axial RoPE, this does not scale well to a large number of dimensions due to having to divide $d_{\text{model}}$ by $D$. To overcome this, we optionally reserve a dimension in Axial RoPE that is used to store the dimensional index $i$ for $i \in [1, 2, \cdots, D]$. E.g., if an embedding belongs to dimension 4, it will receive a positional encoding of 4. When using this approach we include the learned `[CLS]` token, which allows the model to recover the dimensional index despite RoPE being a RPE method (as a consequence of Proposition 4.2). We utilize this method with the ELAsTiCC dataset in Section 5.4 which allows us to reduce the number of positional dimensions from 6 to 2.

### 4.3 Effects of Relative Position in RoMAE

Here we present an analysis of the effects of switching from absolute to relative position in RoMAE.

**Proposition 4.2** (Reconstructing absolute position)**.** When a learned `[CLS]` token is included in the input sequence, the RoMAE Encoder is able to reconstruct the original set of positions $\mathbf{s}$ as described in Section 4.2.

**Intuition:** The `[CLS]` token provides the model with an "anchor". This allows the model to compare each input embedding to the `[CLS]` token, and reconstruct its absolute position. We provide a proof in Appendix C, as well as experimental evidence in Section 5.1.

**Corollary 4.1** (Translational invariance in the RoMAE Encoder)**.** When no learned `[CLS]` token is included in the input sequence, positional information in the RoMAE Encoder is relative. This makes the RoMAE Encoder invariant to translations of the input embedding positions.

**Discussion:** Translational invariance in the RoMAE Encoder makes the pre-training task more difficult because the model cannot make predictions based on the overall position of a token in the input. E.g., with absolute position, the model can learn that objects of interest may often appear near the centre of an image.

**Corollary 4.2** (Effect of distance on absolute position reconstruction)**.** RoMAE is able to recover the absolute position of any embedding $z_i$ with regards to Proposition 4.2, irrespective of the position $s_i$ of that embedding.

**Discussion:** A claimed property of RoPE is that it causes the dot product between queries and keys to decay as their positions grow further apart [54]. In Appendix B.1, we provide empirical evidence showing that RoMAE is able to reconstruct position almost perfectly over two different scales of distance. We also refer to the proof by Barbero et. al. [4], showing that it is possible to construct a key for any non-zero query and any distance such that the softmax value in Attention is maximized at that distance.

Overall, relative position is a key element that influences the training dynamics of RoMAE. This allows us to investigate RoPE from a new angle, drawing new conclusions on the effect of learned embeddings and supporting prior claims that the long-term decay property is not significant to the functioning of RoPE.

# 5 Experiments

Throughout the experiments we make use of different sizes of RoMAE: RoMAE-tiny, RoMAE-small, and RoMAE-base, as detailed in Appendix A.1.

**Compute Details:**

The experiment on the Tiny ImageNet data set (Section 5.2) was run on one node of a Slurm cluster, utilizing two NVIDIA Tesla V100 GPUs for 5 hours.
The experiment on the DESC ELAsTICC Challenge[1] (Section 5.4), was run on a Slurm cluster using 4 nodes for $\sim 4$ hours with each having 4 Nvidia A100 (with 64GB memory) GPUs.
Together, the experiments on the UEA Time Series Archive [3] (Section 5.4), Pendulum dataset [5] (Section 5.4), and absolute position experiments (Section 5.1) were run on a 1080ti GPU for a total of $\sim 1.5$ hours.
All interpolation experiments (Sec. 5.5) were run on a single NVIDIA A100-PCIE-40GB GPU (internal cluster), utilising $\leq$5 GB memory, $\sim$ 10min for the spirals dataset, $\sim$ 30 mins for the synthetic dataset, and $\sim$ 3 hours for PhysioNet.

## 5.1 Reconstructing Absolute Position

To verify the model's ability to reconstruct absolute positional information according to Proposition 4.2, we give the model a sequence of 10 identical values as input. Each embedding is then given a 1D position sampled uniformly between 0 and 50. We then use the same linear head to predict the position for all tokens. Because the model dimension $d_{\text{model}}$ has an effect on the number of $\theta_i$'s RoPE uses, we also test a variety of model sizes. We run each test 5 times and report the mean and standard deviation. Our generated training set has 20 000 samples while our generated test set has 4000. Reported Mean Squared Error (MSE) is the average MSE over the test set. Results are shown in Table 1.

[1] https://portal.nersc.gov/cfs/lsst/DESC_TD_PUBLIC/ELASTICC/

Table 1: Position reconstruction MSE (mean $\pm$ std) for various sizes of RoMAE.

| Model size | With [CLS] | No [CLS] |
|---|---|---|
| RoMAE-tiny | 0.062 (0.007) | 200.33 (0.001) |
| RoMAE-small | 0.0057 (0.002) | 200.33 (0.001) |
| RoMAE-base | 0.0031 (0.002) | 200.33 (0.002) |

Table 2: Results on Tiny ImageNet across various versions of RoMAE-small

| Model | F-score ($\pm$ std) |
|---|---|
| RoMAE (no [CLS]) | 0.500 (0.011) |
| RoMAE ([CLS]) | 0.475 (0.006) |
| RoMAE (absolute) | 0.479 (0.010) |

We observe a clear difference between the model that uses the [CLS] token and the one that does not. When supplied with the learnable token, RoMAE is able to reconstruct the original absolute position almost perfectly. Larger models seem to achieve a better MSE, although all sizes perform well. For all experimental details and an additional experiment on absolute position reconstruction we refer to Appendix D.5 and Appendix B.1 respectively.

### 5.2 Tiny ImageNet

To investigate the effect of positional embedding and the learned [CLS] token on RoMAE, we train three versions of RoMAE on Tiny ImageNet [31]; RoPE with the [CLS] token, RoPE without the [CLS] token, and absolute sinusoidal positional embeddings [58] with the [CLS] token. When fine-tuning RoMAE without the [CLS] token, we place the classification head on top of the mean of the output embeddings, otherwise we place the head on top of the [CLS] token. The final configuration with absolute positional embeddings is very similar to MAE, making for a good comparison. We use a patch size of (16, 16) and mask 75% of the input, similar to MAE. After pre-training each model for 200 epochs, we fine-tune for another 15. We also follow the procedure outlined by MAE to compute the pre-training loss using normalized patch values instead of pixel values.

For full experimental details we refer to Appendix D.1. The results are shown in Table 2. Although all 3 models perform similarly, we highlight that RoMAE with RoPE and no [CLS] token performs slightly better than both RoPE with [CLS] and absolute positional embeddings. This could be because of the models translation invariance (Proposition 4.1). The difference could also come from placing the classification head on top of the mean of output embeddings instead of the [CLS] token. Overall, our results indicate that RoMAE performs at least as well as MAE on images.

### 5.3 Audio benchmark

Table 3: Results for the ESC-50 benchmark for audio datasets, for the RoMAE-small model.

| Model | Accuracy |
|---|---|
| SSAST (Librispeech) | 80.0 |
| RoMAE-small (Librispeech) | **83.2** |
| SSAST (AudioSet-20k) | 82.2 |
| RoMAE-small (AudioSet-20k) | **84.7** |

We chose to test RoMAE's ability to classify audio files, after a self-supervised pre-training on unlabeled audio datasets, inspired by the SSAST pretraining strategy [21]. SSAST is the first Vision transformer-based model that introduces a self-supervised pretraining strategy, supporting arbitrary patch size, for the audio representation learning.

As our pretraining datasets, we used a modified version of the Audioset [20] and Librispeech [36] datasets. AudioSet is a 2017 multi-label audio event classification dataset. Using a carefully structured hierarchical ontology of 635 audio classes guided by manual curation and expert knowledge, the authors collected data from human labelers to probe the presence of specific audio classes, including, for example, human sounds, animal sounds, music, natural sounds, in 2 million 10-second segments of YouTube videos. However, access to the original 2 million audio clips is fraught with difficulty, as a consistent subset of YouTube videos is no longer available. In order to conduct a reproducible experiment, we decided to use as training set the balanced training AudioSet-20k made available on Hugging Face.[2] We thus pretrain RoMAE using two different data sets: AudioSet-20k and the Librispeech dataset. For the Audioset dataset we used the training/validation split provided by the downloaded dataset, while for Librispeech, downloaded and preprocessed using the scripts provided on the SSAST Github repo, we used a 70/30 split.

In order to apply our model to the audio datasets, we first transform the audio waveforms to Mel spectrograms. First, the input audio waveform of length $t$ seconds is converted into a sequence of 128-dimensional log-Mel filterbank[3] (fbank) features computed with a 25ms Hamming window

[2] https://huggingface.co/datasets/agkphysics/AudioSet

[3] The Mel Filterbank transform computes weighted averages of bins to provide spectral power estimates on a logarithmic frequency scale, which is more affine to human hearing resolution.

every 10ms. This results in a $128 \times 100\ t$ spectrogram. For the pretraining step, we followed the patchification strategy adopted in [21] and we split the spectrogram into a sequence of $N$ $(16 \times 16)$ patches, where $N = 12(100t - 16)/10$ is the number of patches and the effective input sequence length for the model. We refer to Appendix D.2 for the list of hyperparameters adopted for the pretraining and finetuning of the model. We would like to highlight here that the pretraining was run for 150 epochs, without the `[CLS]` token, while we choose to adopt a mask ratio equal to 0.75, that is very similar to the masking fraction associated with the SSAST-patch 400 model.

For the finetuning audio classification benchmark, we used the ESC-50 dataset [39], consisting of 2000 5-second environmental audio recordings classified into 50 classes. The current best results on ESC-50 are accuracies of 86.5 and 94.7 obtained with supervised training (on AudioSet-2M) respectively by SOTA-S and SOTA-P models. The SSAST model, which is the only ViT model adopting self-supervised training for this task, achieved accuracies of 82.2 and 84.6 when trained, respectively, on AudioSet-20K and AudioSet-2M (as reported in Table 2 of [21]), showing that the size and richness of the pretraining dataset impacts, in a non-negligible way, the performance of that model on the finetuning tasks. Since we did not have sufficient computational resources to pretrain RoMAE on AudioSet-2M, we chose to compare our model performance with that of the SSAST model pretrained on AudioSet-20K and Librispeech, respectively. We report the benchmark results in Table 3, showing that RoMAE performs better than the SSAST model in these two cases.

### 5.4 Irregular Time-series Classification

Table 4: Light curve classification results on ELAsTiCC.

| Method | F-score |
|---|---|
| Transformer [58, 7] | 0.5256 |
| ATAT [7] | 0.6270 |
| RoMAE-tiny-shallow | 0.7106 |
| RoMAE-tiny | **0.8029** |

Table 5: Regression MSE $\times 10^{-3}$ (mean $\pm$ std) on the Pendulum dataset. We use a custom RoMAE model size.

| Model | Regression MSE ($\times 10^{-3}$) |
|---|---|
| ODE-RNN [44] | 7.26 (0.41) |
| RKN-$\Delta_t$ [5] | 5.09 (0.40) |
| ContiFormer [11] | 4.63 (1.07) |
| CRU [45, 51] | 3.94 (0.21) |
| S5 [51] | 3.41 (0.27) |
| RoMAE | **3.32 (0.13)** |

**DESC ELAsTiCC Challenge** The DESC ELAsTiCC Challenge is a multi-variate irregular time-series dataset consisting of $\sim$1.8M simulated light curves and 36 classes of astronomical objects. Each light curve consists of 6 irregularly sampled channels (called 'bands'). To embed this data in RoMAE, we follow the procedure described in Section 4.2. This results in a 2 dimensional positional embedding, where one dimension embeds the time, and the second embeds the channel index. Although ELAsTiCC has additional metadata for each light curve, we compare performance only across raw light curves. We train RoMAE-tiny by conducting full pre-training for 200 epochs with a masking ratio of 75%, then fine-tuning for 25 epochs. For full details and a visualization of the data we refer to Appendix D.6.

Table 4 shows our results using two sizes of RoMAE, and compares with ATAT [7], a Transformer architecture specialized for ELAsTiCC. Despite the latter's specialization, RoMAE-tiny-shallow (with a comparable number of parameters as ATAT) improves over ATAT by about .08 F-score. The larger RoMAE-tiny achieves an improvement of .18 F-score. A key reason for RoMAE's better performance might be that ATAT does not conduct any pre-training. In the case of RoMAE-tiny, the larger scale of the model also likely plays a role.

**UEA Multivariate Time-series Archive** We evaluate RoMAE on a variety of datasets from the UEA Multivariate Time-series Archive [3]. To make the datasets irregular we follow the procedure outlined by Kidger et. al. [30], dropping 30% of the observations. Because all variates are present at each time-step, the only irregular dimension is time. Therefore, to embed this data in RoMAE, we combine all variates per time-step into one embedding. This is a much simpler setup than the one used for the ELAsTiCC dataset in Section 5.4. For each dataset we conduct pre-training for 400 epochs. When fine-tuning, we found it necessary to change hyper-parameters between different datasets. For more details on our experimental setup we refer to Appendix D.3.

Table 6: Accuracy across various datasets from the UEA Time-series Archive.

| | Model | | | | |
|---|---|---|---|---|---|
| Dataset | TST [65] | mTAN [46] | S5 [51] | ContiFormer [11] | RoMAE-tiny |
| BM | 0.9667 | **0.9917** | 0.9833 | 0.9750 | **0.9917** |
| CT | 0.9742 | 0.9529 | 0.9610 | 0.9833 | **0.9882** |
| EP | **0.9589** | 0.9203 | 0.9074 | 0.9324 | 0.9517 |
| HB | 0.7398 | **0.7789** | 0.7333 | 0.7561 | 0.7447 |
| LSST | 0.5520 | 0.5307 | **0.6389** | 0.6004 | 0.6225 |

Table 7: Results for the interpolation experiments discussed in Section 5.5.

| Synthetic (MSE) | | Spirals (RMSE) | |
|---|---|---|---|
| HeTVAE [35] | $\mathbf{0.223 \pm 0.070}$ | | |
| RoMAE-tiny | $0.233 \pm 0.007$ | Transformer [58] | $1.37 \pm 0.14$ |
| PhysioNet (MSE) | | Latent ODE [10] | $2.09 \pm 0.22$ |
| | | ContiFormer [11] | $0.49 \pm 0.06$ |
| HeTVAE [35] | $0.562 \pm 0.022$ | RoMAE-tiny | $\mathbf{0.0183 \pm 0.007}$ |
| RoMAE-tiny | $\mathbf{0.467 \pm 0.021}$ | | |

Table 6 presents the mean accuracy from 3 full training runs (pretraining + finetuning), as well as published results from a variety of models. Overall, RoMAE-tiny performs similarly or better than the comparators, and is able to handle the various datasets without issues. All the datasets trained on are relatively small, with some being on the order of hundreds of samples. Therefore, this experiment also shows how data-efficient RoMAE can be.

**Irregular Time-series Regression: Pendulum Dataset** The Pendulum dataset [51] is an irregular time-series dataset consisting of irregularly sampled images of a pendulum. To embed the images in RoMAE, we use a patch size of $(1, 24, 24)$ for (time, height, width). This corresponds to 1 embedding per time-step/image. RoMAE is trained directly on regression without any pre-training, predicting the sine and cosine of the angle of the pendulum which follows a non-linear dynamical system. We use a custom size for RoMAE which contains only 2 layers and an MLP hidden size of 30. This is much smaller than RoMAE-tiny and provides for a fairer comparison with other models trained on this dataset. For additional information on the dataset and full experimental details we refer to Appendix D.4. After training on 20 different seeds, the mean MSE and standard deviation of RoMAE on the Pendulum dataset are reported in Table 5. RoMAE outperforms specialized Transformer based models such as ContiFormer [12], as well as state space models such as S5 [51].

### 5.5 Interpolation

We evaluate RoMAE on three interpolation tasks with increasing dimensionality and sampling irregularity. *(i) Spiral*: A 2D synthetic benchmark of 300 noisy Archimedean spirals as in Ref. [12]; *(ii) Synthetic*: The 50-step univariate task from Ref. [48] and *(iii) PhysioNet*: 48-hour ICU records containing 41 clinical variables [49]. For *(i)*, each spiral is discretized into 75 evenly-spaced time steps. To create irregular time-series data, time points are randomly selected from the first half of each spiral, which are used for interpolation. For *(ii)*, the interpolation task is between a random subsample including between 3 and 10 points per trajectory. For *(iii)*, we follow the 50% masking protocol of [35]: half of the time, rows with at least one observation are hidden and must be reconstructed from context. We provide additional experimental details in the Appendix, respectively D.7, D.8, and D.9. All experiments are conducted to ensure a direct comparison with the results of the Transformer [58], LatentODE [9], and ContiFormer [12] for *(i)*, and HeTVAE for *(ii)* and *(iii)*. Across scales, our RoMAE-tiny configuration consistently scores competitively with respect to benchmarked MSE/RMSE, as seen in Table. 7. We lastly remark on RoMAE's ability to retain progressively higher frequency modes for interpolation with time series data in Appendix B.4.

## 6 Discussion

Table 7 shows that one tiny/small RoMAE model, pre-trained once with a generic masked-autoencoder objective and *no* task-specific architectural tuning, matches or surpasses specialised baselines across three increasingly difficult interpolation datasets. On the 2D *Spiral* benchmark, we improve by an order of magnitude over the previous best result by ContiFormer, which we attribute to MAE tubelet-masking enforcing long-range reasoning. For the 50-step *Synthetic* task we obtain $0.233 \pm 0.007$ MSE, comparable to HeTVAE's $0.223 \pm 0.070$ but with markedly tighter standard deviation (five seeds). On the irregular, 41-channel *PhysioNet-2012* ICU data we achieve $0.570 \pm 0.014$ MSE, within half a standard deviation of HeTVAE's heteroscedastic decoder. We have verified by retraining HeTVAE on the the PhysioNet interpolation task that it indeed excels on densely sampled variables, such as heart-rate, whereas RoMAE maintains more balanced interpolation across the sparsely observed channels; consequently, while the aggregate MSE over all the features is similar, RoMAE delivers a simpler, single-stage model without task-specific adaptations.

These results indicate that RoMAE pre-training can be universally beneficial for continuous time interpolation over irregular time-series data. The RoMAE architecture is able to scale from low-dimensional position time embeddings (50,1), to large (2880,41) points with extremely sparse observations differing across features, without refinement of the inherent architecture, suggesting that MAE-style models can serve as strong, task-agnostic baselines for continuous-time interpolation. RoMAE's classification results on datasets sizes ranging from a few hundreds to millions show how pre-training enables RoMAE to be both scalable and data-efficient.

Our tests on position reconstruction demonstrate that care must be taken when working with learned embeddings. Given how prominent the `[CLS]` token is when working with Transformer Encoders, our results are relevant for a multitude of models, including RoFormer [54].

**Limitations:** RoPE in RoMAE has some additional computational overhead if the positions are different with each forward pass, e.g., with any continuous irregular time-series. If positions stay constant, however, as in images, the overhead becomes negligible. For details on the additional compute incurred by continuous RoPE, see Appendix B.2. RoMAE is also not well suited for very long sequences, as it uses standard Attention which has $O(n^2)$ memory complexity with regards to sequence length. Lastly, RoMAE's ability to perform on extrapolation tasks is limited, as discussed in Sec. B.3.

**Broader and Societal Impact:** We believe that RoMAE's flexibility can extend representation learning to scientific fields where it may not have been easily adopted thus far, e.g., similar to what we showed with the ELAsTiCC Challenge. The potential for the misuse of RoMAE is similar to that of MAE and other foundational models.

## 7 Conclusion

This paper introduced RoMAE, an extension to the Masked Autoencoder architecture that allows it to accept multi-dimensional data with continuous positions as input. We investigated the theoretical implications of using relative positional embeddings for an MAE model, showing how it changes the difficulty of the masked pre-training task. We also showed how the model can use a learned `[CLS]` token to recover absolute position. Results across a large variety of modalities and tasks have shown that RoMAE achieves excellent performance in modalities requiring continuous position while maintaining the performance of MAE on images and audio. Future work will include building a more robust theoretical understanding of the implications of using RoPE. Because RoPE is compatible with various Attention implementations, one could also adapt RoMAE to use a linear Attention variant, which would allow it to work with much larger input sequences. Finally, we envisage the exploration of the many potential modalities that RoMAE could work with that were not tested here.

## Acknowledgments

RT and AS acknowledge funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 Project FAIR "Future Artificial Intelligence Research". This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22]. RT is partially supported by the Fondazione ICSC, Spoke 3 "Astrophysics and Cosmos Observations", Piano Nazionale di Ripresa e Resilienza Project ID CN00000013 "Italian Research Center on High-Performance Computing, Big Data and Quantum Computing" funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S (M4C2-19)" - Next Generation EU (NGEU). We also acknowledge the use of computational resources provided by the Italian AREA Science Park supercomputing platform ORFEO in Trieste. GC is supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Postdoctoral Fellowship Programme, SMASH co-funded under the grant agreement No. 101081355. The operation (SMASH project) is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund.

## References


[1] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00676. URL https://doi.org/10.1109/ICCV48922.2021.00676.

[2] L. J. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL http://arxiv.org/abs/1607.06450.

[3] A. J. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, and E. J. Keogh. The UEA multivariate time series classification archive, 2018. *CoRR*, abs/1811.00075, 2018. URL http://arxiv.org/abs/1811.00075.

[4] F. Barbero, A. Vitvitskyi, C. Perivolaropoulos, R. Pascanu, and P. Velickovic. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=GtvuNrk58a.

[5] P. Becker, H. Pandya, G. H. W. Gebhardt, C. Zhao, C. J. Taylor, and G. Neumann. Recurrent kalman networks: Factorized inference in high-dimensional deep feature spaces. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 544–552. PMLR, 2019. URL http://proceedings.mlr.press/v97/becker19a.html.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[7] Cabrera-Vives, G., Moreno-Cartagena, D., Astorga, N., Reyes-Jainaga, I., Förster, F., Huijse, P., Arredondo, J., Muñoz Arancibia, A. M., Bayo, A., Catelan, M., Estévez, P. A., Sánchez-Sáez, P., Álvarez, A., Castellanos, P., Gallardo, P., Moya, A., and Rodriguez-Mancini, D. Atat: Astronomical transformer for time series and tabular data. *A&A*, 689:A289, 2024. doi: 10.1051/0004-6361/202449475. URL https://doi.org/10.1051/0004-6361/202449475.

[8] H. Chen, Y. Han, F. Chen, X. Li, Y. Wang, J. Wang, Z. Wang, Z. Liu, D. Zou, and B. Raj. Masked autoencoders are effective tokenizers for diffusion models. *CoRR*, abs/2502.03444, 2025. doi: 10.48550/ARXIV.2502.03444. URL https://doi.org/10.48550/arXiv.2502.03444.

[9] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6571–6583, 2018.

[10] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6572–6583, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

[11] Y. Chen, K. Ren, Y. Wang, Y. Fang, W. Sun, and D. Li. Contiformer: Continuous-time transformer for irregular time series modeling. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 -*

*16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9328208f88ec69420031647e6ff97727-Abstract-Conference.html.

[12] Y. Chen, Q. Wang, and Y. e. Fu. Continuous-time Transformer for Irregular Time-series Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[13] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[16] S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. doi: 10.1016/J.NEUNET.2017.12.012. URL https://doi.org/10.1016/j.neunet.2017.12.012.

[17] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao. EVA-02: A visual representation for neon genesis. *Image Vis. Comput.*, 149:105171, 2024. doi: 10.1016/J.IMAVIS.2024.105171. URL https://doi.org/10.1016/j.imavis.2024.105171.

[18] C. Feichtenhofer, H. Fan, Y. Li, and K. He. Masked autoencoders as spatiotemporal learners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/e97d1081481a4017df96b51be31001d3-Abstract-Conference.html.

[19] S. Gao, T. Koker, O. Queen, T. Hartvigsen, T. Tsiligkaridis, and M. Zitnik. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[21] Y. Gong, C. Lai, Y. Chung, and J. R. Glass. SSAST: self-supervised audio spectrogram transformer. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10699–10709. AAAI Press, 2022. doi: 10.1609/AAAI.V36I10.21315. URL https://doi.org/10.1609/aaai.v36i10.21315.

[22] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752, 2023. doi: 10.48550/ARXIV.2312.00752. URL https://doi.org/10.48550/arXiv.2312.00752.

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

[24] D. Hendrycks and K. Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL http://arxiv.org/abs/1606.08415.

[25] B. Heo, S. Park, D. Han, and S. Yun. Rotary position embedding for vision transformer. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part X*, volume 15068 of *Lecture Notes in Computer Science*, pages 289–305. Springer, 2024. doi: 10.1007/978-3-031-72684-2_17. URL https://doi.org/10.1007/978-3-031-72684-2_17.

[26] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. URL http://arxiv.org/abs/1207.0580.

[27] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger. Deep networks with stochastic depth. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 646–661. Springer, 2016. doi: 10.1007/978-3-319-46493-0_39. URL https://doi.org/10.1007/978-3-319-46493-0_39.

[28] P. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer. Masked autoencoders that listen. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b89d5e209990b19e33b418e14f323998-Abstract-Conference.html.

[29] P. Jeevan and A. Sethi. Resource-efficient hybrid x-formers for vision. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 3555–3563. IEEE, 2022. doi: 10.1109/WACV51458.2022.00361. URL https://doi.org/10.1109/WACV51458.2022.00361.

[30] P. Kidger, J. Morrill, J. Foster, and T. J. Lyons. Neural controlled differential equations for irregular time series. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

[31] Y. Le and X. S. Yang. Tiny imagenet visual recognition challenge. 2015. URL http://vision.stanford.edu/teaching/cs231n/reports/2015/pdfs/yle_project.pdf.

[32] Z. Li, S. Li, and X. Yan. Time series as images: Vision transformer for irregularly sampled time series. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9a17c1eb808cf012065e9db47b7ca80d-Abstract-Conference.html.

[33] Y. Liu, H. Zhang, C. Li, X. Huang, J. Wang, and M. Long. Timer: Generative pre-trained transformers are large time series models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=bYRYb7DMNo.

[34] Z. Lu, Z. Wang, D. Huang, C. Wu, X. Liu, W. Ouyang, and L. Bai. Fit: Flexible vision transformer for diffusion model. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=jZVen2JguY.

[35] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt. Early Recognition of Sepsis with Heteroscedastic Temporal Variational Autoencoders. In *International Conference on Machine Learning (ICML)*, pages 7781–7792, 2021.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[37] H. Patel, R. Qiu, A. Irwin, S. Sadiq, and S. Wang. EMIT - event-based masked auto encoding for irregular time series. In E. Baralis, K. Zhang, E. Damiani, M. Debbah, P. Kalnis, and X. Wu, editors, *IEEE International Conference on Data Mining, ICDM 2024, Abu Dhabi, United Arab Emirates, December 9-12, 2024*, pages 370–379. IEEE, 2024. doi: 10.1109/ICDM59182.2024.00044. URL https://doi.org/10.1109/ICDM59182.2024.00044.

[38] B. Peng, J. Quesnelle, H. Fan, and E. Shippole. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

[39] K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1015–1018, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334594. doi: 10.1145/2733373.2806390. URL https://doi.org/10.1145/2733373.2806390.

[40] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[42] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[43] B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bitton, M. Bhatt, C. Canton-Ferrer, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950, 2023. doi: 10.48550/ARXIV.2308.12950. URL https://doi.org/10.48550/arXiv.2308.12950.

[44] Y. Rubanova, T. Q. Chen, and D. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5321–5331, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

[45] M. Schirmer, M. Eltayeb, S. Lessmann, and M. Rudolph. Modeling irregular time series with continuous recurrent units. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 19388–19405. PMLR, 2022. URL https://proceedings.mlr.press/v162/schirmer22a.html.

[46] S. N. Shukla and B. M. Marlin. Multi-time attention networks for irregularly sampled time series. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

[47] S. N. Shukla and B. M. Marlin. Heteroscedastic temporal variational autoencoder for irregularly sampled time series. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Az7opqbQE-3.

[48] S. N. Shukla and B. M. Marlin. Heteroscedastic temporal variational autoencoder for irregularly sampled time series. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Az7opqbQE-3.

[49] I. Silva, B. Moody, D. Scott, L. Celi, R. Mark, and G. Clifford. The PhysioNet/Computing in Cardiology Challenge 2012: Predicting In-Hospital Mortality from ICU Data. In *Computing in Cardiology*, pages 245–248, 2012.

[50] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 computing in cardiology*, pages 245–248. IEEE, 2012.

[51] J. T. H. Smith, A. Warrington, and S. W. Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

[52] Z. Song, Q. Lu, H. Zhu, D. Buckeridge, and Y. Li. Trajgpt: Irregular time-series representation learning for health trajectory analysis. *CoRR*, abs/2410.02133, 2024. doi: 10.48550/ARXIV.2410.02133. URL https://doi.org/10.48550/arXiv.2410.02133.

[53] N. Stroh. Trackgpt–a generative pre-trained transformer for cross-domain entity trajectory forecasting. *arXiv preprint arXiv:2402.00066*, 2024.

[54] J. Su, M. H. M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/J.NEUCOM.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

[55] K. Su, Q. Wu, P. Cai, X. Zhu, X. Lu, Z. Wang, and K. Hu. RI-MAE: rotation-invariant masked autoencoders for self-supervised point cloud representation learning. In T. Walsh, J. Shah, and Z. Kolter, editors, *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 7015–7023. AAAI Press, 2025. doi: 10.1609/AAAI.V39I7.32753. URL https://doi.org/10.1609/aaai.v39i7.32753.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

[57] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[59] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. Videomae V2: scaling video masked autoencoders with dual masking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14549–14560. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01398. URL https://doi.org/10.1109/CVPR52729.2023.01398.

[60] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun. Transformers in time series: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6778–6786. ijcai.org, 2023. doi: 10.24963/IJCAI.2023/759. URL https://doi.org/10.24963/ijcai.2023/759.

[61] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[62] W. Xiong, J. Liu, I. Molybog, H. Zhang, P. Bhargava, R. Hou, L. Martin, R. Rungta, K. A. Sankararaman, B. Oguz, M. Khabsa, H. Fang, Y. Mehdad, S. Narang, K. Malik, A. Fan, S. Bhosale, S. Edunov, M. Lewis, S. Wang, and H. Ma. Effective long-context scaling of foundation models. In K. Duh, H. Gómez-Adorno, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4643–4663. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.260. URL https://doi.org/10.18653/v1/2024.naacl-long.260.

[63] M. Xu, X. Men, B. Wang, Q. Zhang, H. Lin, X. Han, and W. Chen. Base of rope bounds context length. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9f12dd32d552f3ad9eaa0e9dfec291be-Abstract-Conference.html.

[64] Z.-Q. J. Xu, Y. Zhang, and T. Luo. Overview frequency principle/spectral bias in deep learning. *Communications on Applied Mathematics and Computation*, 7(3):827–864, 2025.

[65] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff. A transformer-based framework for multivariate time series representation learning. In F. Zhu, B. C. Ooi, and C. Miao, editors, *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2114–2124. ACM, 2021. doi: 10.1145/3447548.3467401. URL https://doi.org/10.1145/3447548.3467401.

[66] B. Zhang and R. Sennrich. Root mean square layer normalization. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

## 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 abstract and introduction states that we showcase RoMAE's abilities on irregular and multivariate time-series (shown in Section 5.4 and specifically 5.4), images (shown in Section 5.2 and 5.4), audio (details shown in Appendix 5.3), where we provide performances of state-of-the-art baselines. The investigation of RoMAE's ability to reconstruct embedded positions and analysis of relative position property are shown in Section 5.1 and 5.2.

   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: Limitations are discussed in Section 4 and 6, as well as in Appendix B.3.

   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: [Yes]

Justification: See Appendix C.

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: All necessary details for reproducibility are provided in Section 5 and Appendix.

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: [Yes]

   Justification: Link to code is provided in Section 6: `https://anonymous.4open.science/w/RoMAE-FF56/`. Experiment/evaluation code and checkpoints will also be made public.

   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: See Section 5 and Appendix.

   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: All experiments report performance as mean and standard deviation obtained on several runs as described in Section 5 and Appendix, except for the DESC ELasTiCC Challenge, as this is a significantly larger dataset.

   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: See Section 5.

   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: –

   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: We discuss how our model can be applied positively in downstream deployments and draw parallels with other foundation models regarding negative impacts; see Section 6.

    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: No LLMs or image generator are presented. The datasets used here are all publicly available, and the tasks addressed should not be at risk of misuse.

    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: All datasets used in this paper are properly cited. We have also cited the code assets our model relies on both in the paper and in our code.

    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: [Yes]

    Justification: We provide the code for our model and our training utilities. We also specify all training details in the paper.

    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:

    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:

    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:

    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.