Synergizing Multigrid Algorithms with Vision Transformer: A Novel Approach to Enhance the Seismic Foundation Model

Huiwen Wu [1] *, **Shuo Zhang** [2,3] †, **Yi Liu** [1], **Hongbin Ye** [1]

[1] Research Center for Scientific Data Hub, Zhejiang Laboratory, 310001, Hangzhou, China;
[2] State Key Laboratory of Mathematical Sciences (SKLMS) and State Key Laboratory of Scientific and Engineering Computing (LSEC), Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100190, Beijing, China.
[3] School of Mathematical Sciences, University of Chinese Academy of Sciences, 100049, Beijing, China.
whw@zhejianglab.org, szhang@lsec.cc.ac.cn, liuyi4@zhejianglab.org, zjuhongbinye@gmail.com

Abstract

Due to the rapid advancement and homogenization of Artificial Intelligence (AI) technology development, transformer-based foundation models have revolutionized scientific applications, such as drug discovery, materials research, and astronomy. However, seismic data presents unique characteristics that require specialized processing techniques for pretraining foundation models in seismic contexts with high- and low-frequency features playing crucial roles. Existing Vision Transformer (ViT) with sequential image tokenization fails to efficiently and effectively capture both high- and low-frequency seismic information because they ignore the intrinsic structural patterns of seismograms. This work introduces **ADATG**, a novel adaptive two-grid training strategy with Hilbert encoding, explicitly tailored for seismogram data and leveraging the hierarchical structures inherent in seismic data. Specifically, our approach employs spectrum decomposition to separate high- and low-frequency components, and hierarchical Hilbert encoding to represent the data effectively. Moreover, inspired by the frequency principle, we propose an adaptive training strategy that initially emphasizes coarse-level information and then progressively refines the model's focus on fine-level features. Extensive experiments demonstrate the effectiveness and efficiency of our method. This research highlights the importance of data encoding and training strategies informed by the distinct characteristics of high- and low-frequency features in seismic images, ultimately enhancing the pretraining of visual seismic foundation models.

Extended version — https://arxiv.org/abs/2511.13800

1 Introduction

The rapid development of transformer-based models has enabled a wide range of AI-driven applications in complex scientific tasks. This includes ligand coupling for drug discovery (Zhang et al. 2023), materials discovery (Pyzer-Knapp et al. 2025), and mathematical proof (Xin et al. 2024). Recently, significant attention has been paid to training seismic foundation models (Sheng et al. 2025; Si et al. 2024; Liu et al. 2024; Li et al. 2024) using large-scale, globally sourced seismic data. These pre-trained seismic models can be effectively utilized in various downstream tasks, including seismic facies classification, geobody segmentation, seismic image denoising, and full seismic inversion (Sheng et al. 2025). The development and training of seismic foundation models are essential for advancing both scientific research and practical engineering applications in geophysics.

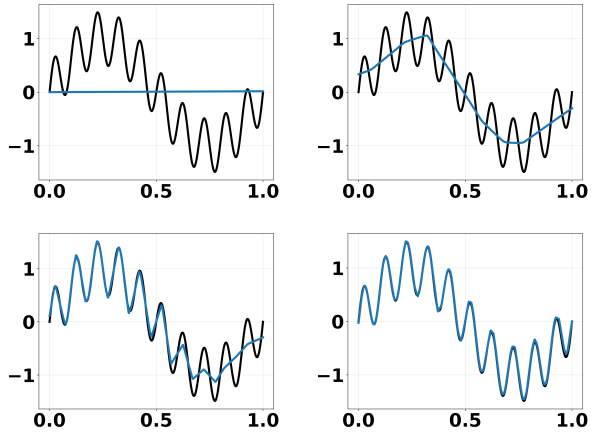


Figure 1: A toy example illustrating the Frequency Principle. From left to right and top to bottom, the panels show the model's prediction (blue curve) at training epochs 0, 100, 1,000, and 10,000, respectively, compared against the ground-truth signal (gray curve).

Due to the intrinsic characteristics of seismic data, a seismogram can be decomposed into high-frequency and low-frequency components. The high-frequency components are typically associated with rapid changes in ground motion and shorter wavelengths (Yilmaz 2001; Sheriff and Geldart 1995). In contrast, low-frequency components correspond to slower changes in ground motion and are linked to longer wavelengths (Pratt 1999; Virieux and Operto 2009). These high-frequency and low-frequency components play distinct but complementary roles in seismic analysis. From the perspective of waveform complexity, high frequencies contribute to the sharpness of seismic phases—such as P- and S-wave arrivals. In contrast, low frequencies predom-

*Corresponding author
†Corresponding author

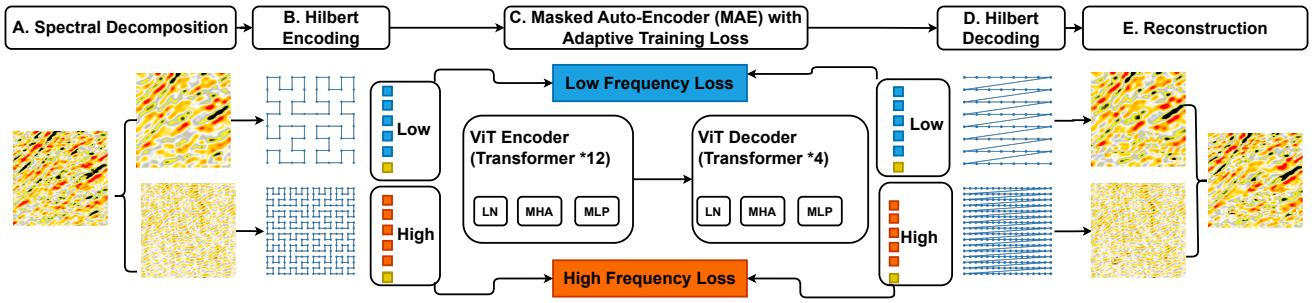


Figure 2: Pipeline of Seismic Foundation Model Pretraining with **ADATG**. (A) Input seismogram spectral decomposition via Fourier transform, (B) Frequency-adaptive Hilbert encoding (fine/coarse grids for high/low frequencies), (C) ViT training with adaptive MAE strategy, (D) Frequency-adaptive Hilbert decoding with inverse transformer matrix, (E) Merged reconstruction of difference frequency components.

inantly shape the overall envelope of signals, particularly surface waves (Aki and Richards 2002; Stein and Wysession 2003). In event discrimination tasks, the high-frequency to low-frequency energy ratio can help distinguish between tectonic earthquakes and explosions, as explosions tend to generate more high-frequency energy (Fichtner 2010; Tarantola 2005).

Thus, it is crucial to develop a unified foundation model effectively capturing both high-frequency and low-frequency components simultaneously.

Training foundation models to effectively leverage both high- and low-frequency features presents several challenges. First, in the traditional Vision Transformer (ViT) model, images are divided into 16×16 tokens, which are processed sequentially (Figure 4 (a)). This sequential encoding can disrupt intrinsic patterns within the seismogram. Second, because the seismogram contains both high- and low-frequency information, both of which are crucial for downstream tasks, existing ViT models often fail to capture the fine details represented by the high-frequency components and perform inadequately when reconstructing low-frequency components. Lastly, motivated by the frequency principle (Figure 5b) in which models preferentially learn low-frequency components before high-frequency ones during training — has been well-established in prior work (Xu et al. 2019; Xu, Zhang, and Luo 2025). Our experiments demonstrate that this principle also governs ViT, as detailed in Section 4.5. The existing ViT training paradigm which ignores this fact fails to acquire low- and high- frequency features accurately.

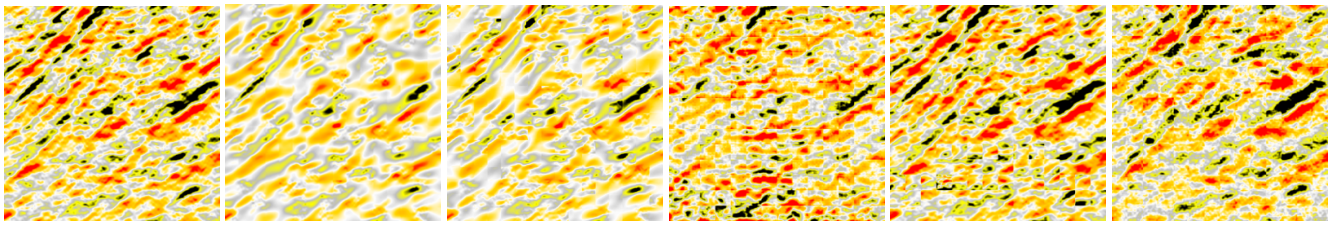
To address the challenges outlined above, we have enhanced the training approach for seismic foundation models in several key ways. First, we developed a Hilbert encoder for ViT to encode seismogram data based on their intrinsic patterns. Next, inspired by Multigrid methods sprits (Wesseling 1995; Xu and Zikatanov 2017; Chen and Wu 2018), we decompose the high- and low-frequency features of the input seismogram using discrete Fourier transforms and process them with different-scale Hilbert encoding. For low-frequency features, we use the standard Hilbert encoder; for high-frequency features, we use a refined Hilbert encoder. Finally, motivated by this observation, we propose a two-grid adaptive frequency decomposition strategy: the model prioritizes high-frequency features in the early training stages before shifting focus to low-frequency refinement later (Section 3.6). The main contributions are listed below.

- We propose a frequency decomposition method via a discrete Fourier transform to decompose the seismogram into high- and low-frequency components.
- We design a novel two-grid Hilbert encoding methods with coarse-level Hilbert encoding for the low-frequency feature and fine-level Hilbert encoding for high-frequency feature.
- Experimental results verifies the Frequency Principle for Transformer architecture. We further design the adaptive training strategy inspired by the Frequency Principle.

2 Related Work

Seismic Foundation Models. Recently, we encountered a few training foundation models for seismic data and transferred the learned knowledge to downstream tasks. For example, in **SFM** (Sheng et al. 2025), the authors employ self-supervised learning to pretrain a Transformer-based seismic foundation model to produce all-purpose seismic features that can be applied to various downstream tasks. In SeisCLIP (Si et al. 2024), the authors propose aa seismology foundation model trained through contrastive learning from multimodal data. In SeisLM (Liu et al. 2024), the authors propose a foundation seismic foundation model with input data from seismic waveform signals and trained using self-supervised contrastive loss, similar to language models. However, since most seismic foundation models process the time-frequency seismic spectrum as input data, **SFM** (Sheng et al. 2025) is the exception that treats the seismogram as an image and processes it using computer vision techniques. Here, we focus on the image processing approach and compare it with **SFM** (Sheng et al. 2025).

Masked Image Modeling. The adaptation of Transformers to computer vision began with Dosovitskiy et al. (Dosovitskiy 2020), who tokenized images into 16×16 patches



(a) **Original** (b) **SFM** (c) **HE-ViT** (d) **Ran-TG** (e) **ADATG-HH** (f) **ADATG-NH**

Figure 3: Reconstructed images using different pretrained Seismic Foundation Models (SFM). From left to right: the original image, reconstruction using the base Vision Transformer (ViT) architecture (Sheng et al. 2025), the Hilbert Encoding ViT (**HE-ViT**), the randomized Two-grid method (**Ran-TG**), and two variants of **ADATG**: one incorporating both high and low frequency components with Hilbert Encoding (**ADATG-HH**), and another using only high-frequency components (**ADATG-NH**). All methods are applied under identical evaluation conditions to enable direct visual comparison.

for sequence modeling. This inspired Masked Image Modeling (MIM) approaches, notably the Masked Autoencoder (MAE) (He et al. 2022), which randomly masks input patches and reconstructs the missing content using an asymmetric encoder-decoder: the encoder operates only on visible patches, while a lightweight decoder recovers the full image from encoded features and mask tokens. Subsequent works have extended MAE in various directions—Liu et al. (Liu et al. 2023) proposed MixMAE, which mixes visible patches across images and trains dual decoders for joint reconstruction, and Han et al. (Han et al. 2024) developed an efficient MAE variant for large-scale models. In contrast to these image-domain methods, we integrate spectral decomposition into the MAE framework of (He et al. 2022) to enable effective self-supervised pre-training on large-scale seismogram data.

Space-Filling Curves. Space-filling curves map multi-dimensional data into 1D sequences while preserving locality. The Hilbert curve, in particular, has been widely adopted for spatial encoding: Chen (Chen, Wang, and Shi 2007) proposed constant-complexity encoding/decoding via its recursive structure; Moon et al. (Moon et al. 2001) derived closed-form cluster counts for arbitrary query regions; and Bhupati (Bhupati et al. 2019) used it to process fMRI activation maps for classification. Leveraging the strong spatial coherence in seismograms, we employ the Hilbert curve to linearize and encode vision tokens in our pre-training framework.

3 Methodology

3.1 Spectral Decomposition

Definition 1 (Discrete Fourier Transform in Matrix Form (Trefethen and Bau 1997))**.** *Let* $\omega = \exp(-2\pi/Ni)$ *and* $\omega^{jk} = \exp(-jk2\pi/Ni)$*. Then* $\omega^N = 1$*. Then the* $N \times N$ *Fourier matrix is defined as*

$$\mathbf{F}_N = \begin{bmatrix} \omega^{0\cdot 0} & \omega^{0\cdot 1} & \dots & \omega^{0\cdot(N-1)} \\ \omega^{1\cdot 0} & \omega^{1\cdot 1} & \dots & \omega^{1\cdot(N-1)} \\ \dots & \dots & \dots & \dots \\ \omega^{(N-1)\cdot 0} & \omega^{(N-1)\cdot 1} & \dots & \omega^{(N-1)\cdot(N-1)} \end{bmatrix}.$$

Then the Discrete Fourier Transform (DFT) in matrix form can be defined as

$$\hat{\mathbf{X}} = \mathbf{DFT}(X) = \mathbf{F}_N \mathbf{X},$$

and the inverse discrete Fourier transform (iDFT) in matrix form is

$$\mathbf{X} = \mathbf{iDFT}(\hat{\mathbf{X}}) = \mathbf{F}_N^{-1} \hat{\mathbf{X}}.$$

Then we apply the DFT to the input seismogram image.

$$\hat{\mathbf{X}} = \mathbf{DFT}(\mathbf{X}); \quad \hat{\mathbf{X}}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i e^{-i2\pi ik/n}. \tag{1}$$

Next, we decompose the high-frequency and low-frequency component in the Fourier space with a given threshold k_0,

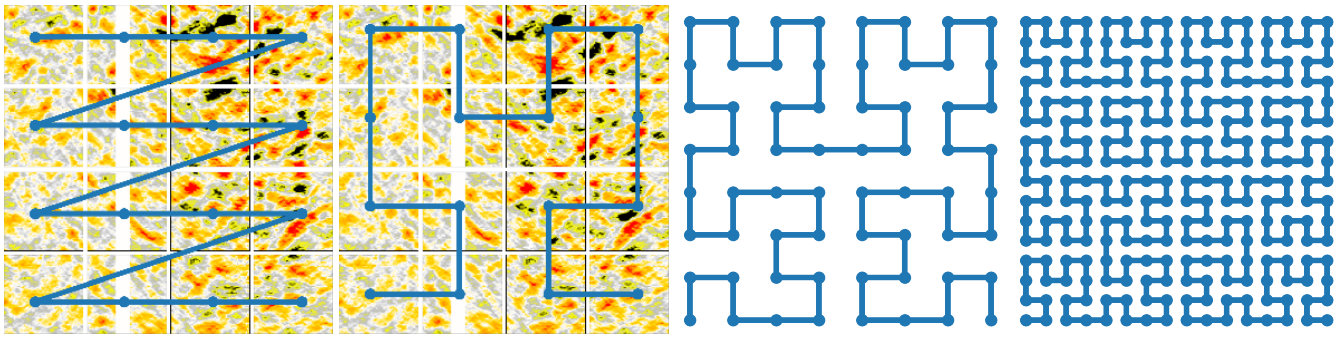
$$\hat{\mathbf{X}}_{\text{low}} = \sum_{k \le k_0} \hat{\mathbf{X}}_k; \quad \hat{\mathbf{X}}_{\text{high}} = \sum_{k > k_0} \hat{\mathbf{X}}_k. \tag{2}$$

Then we apply iDCT to convert the high-frequency and low-frequency components of the seismogram into the physical domain, separately.

$$\mathbf{X}_{\text{low}} = \mathbf{iDFT}(\hat{\mathbf{X}}_{\text{low}}); \quad \mathbf{X}_{\text{high}} = \mathbf{iDFT}(\hat{\mathbf{X}}_{\text{high}}). \tag{3}$$

3.2 Vision Transformer (ViT)

In this section, we introduce the traditional ViT (Dosovitskiy 2020) in the pre-training model of the seismic foundation. Let's consider an input image with dimensions $H \times W \times C$, where H, W, and C represent the height, width, and number of channels, respectively. In the traditional ViT framework (Dosovitskiy 2020), the first step involves reshaping the input image into a sequence of flattened two-dimensional patches, resulting in a shape of $N \times (P^2 \times C)$. Here, N is the total number of patches and P^2 is the area of a single patch. The convolutional operator used in convolutional neural networks (CNN) effectively captures both locality and translationally equivalent information (Krizhevsky, Sutskever, and Hinton 2012). In contrast, in the base ViT architecture, only the final Multi-Layer Perception (MLP) captures local and translationally equivalent information, while the self-attention layer focuses on global information (Dosovitskiy 2020). This design allows ViT to exhibit stronger generalization capabilities, although they may capture the local texture of images less effectively.



(a) **Base Encoding** (b) **Hilbert Encoding** (c) **Coarse Grid** (d) **Fine Grid**

Figure 4: Hilbert Encoding (Left Two) and Twogrid Hilbert Encoding (Right Two).

3.3 Hilbert Encoding

Importantly, seismogram images present significant patterns characterized by both high- and low-frequency information. The high frequency components are typically related to shallow or local events, whereas the low frequency components may correspond to deeper events (Ringler et al. 2022; Lindsey and Martin 2021). Therefore, relying solely on global or local information when analyzing seismogram images can be limiting. To effectively capture local features in the seismogram, we introduce Hilbert encoding based on a space-filling curve. First of all, we introduce the space filling curve.

Definition 2. *Space Filling Curve (Bially 1969) A space-filling curve is a continuous function* $f : [0,1] \to [0,1]^n$ *such that for every* $y \in [0,1]^n$*, there exists a point* $t \in [0,1]$ *where f(t) =y.*

A space-filling curve has three key properties, including continuity, surjectivity, and fractal nature. The last property enables iterative construction. The Hilbert curve is a special example of a space-filling curve and we define the iterative generation as below.

Definition 3. *Hilbert Curve (Butz 1969) The Hilbert curve is generated by the following formula iteratively.*

- *Base Case, Order 1 The first order Hilbert curve* $\mathcal{H}_1 = [(0,0),(0,1),(1,1),(1,0)]$.
- *Recursive Step, Order n The curve* $\mathcal{H}_n$ *is constructed by connecting four copies of* $(n-1)-th$ *order Hilbert curve* $\mathcal{H}_{n-1}$*, each rotated and reflected to ensure continuity.*

$$\mathcal{H}_{n+1} = [\mathrm{Rotate}(\mathcal{H}_\mathrm{n}, \frac{\pi}{2}], \mathcal{H}_\mathrm{n}, \mathcal{H}_\mathrm{n}, \mathrm{Rotate}(\mathcal{H}_\mathrm{n}, \frac{3\pi}{2})].$$

The inverse of a Hilbert curve, denoted $\mathcal{H}^{-1}$, can be defined as the operation that reverses the permutation sequence generated by the Hilbert curve $\mathcal{H}$. Using the Hilbert encoding (Chen, Wang, and Shi 2007), one can further change the sequence of the patched sequence. Specifically, in traditional ViT, after patchification, we group the embeddings in a right-to-left, up-to-down sequence; see Figure 4a. With Hilbert encoding, the sequence of embeddings is constructed based on the Hilbert curve sequence; see Figure 4b.

3.4 Twogrid Hilbert Encoding

Let X_{high} and X_{low} be the high- and low-frequency components obtained from the spectral decomposition defined in Eq. (3). Select the order of the Hilbert curve $n_1, n_2 \in \mathbb{N}^+$ and $n_1 < n_2$. Let $\mathcal{H}_{n_1}$ be the Hilbert curve of order n_1 and $\mathcal{H}_{n_2}$ be the Hilbert curve of order n_2. For example, in Figure 4, we show the coarse grid of order 3 in Figure 4c and the fine grid of order 4 in Figure 4d. Next, we use the Hilbert curve on the coarse grid of order n_1 to encode the low-frequency component, while the Hilbert curve on the fine grid of order n_2 to encode the high-frequency component, respectively.

$$E_{\text{low}} = \mathcal{H}_{n_1}(X_{\text{low}}), \quad E_{\text{high}} = \mathcal{H}_{n_2}(X_{\text{high}}). \tag{4}$$

After obtaining the partial embedding, we input the low- and high-frequency data into the foundation model $\mathcal{M}$.

3.5 Frequency Decomposed Twogrid Approach

In this section, we outline the algorithm for training twogrid MAEs in Algorithm 1. Specifically, with the frequency decomposition process described in Sec. 3.1, one can decompose the input seismogram into two components, high frequency and low frequency, respectively. Next, we used a single neural network to reconstruct the high- and low-frequency components, respectively. The model structure is the small size Transformer with 256 heads, 12 Transformer blocks encoder and 4 Transformer blocks decoder. Let X_{high} represent the high-frequency component of the input seismogram image, while X_{low} denote the decomposed low-frequency component. Let $\mathcal{M}$ be the trained foundation models. Accordingly, the loss function for the high-frequency model is

$$\mathcal{L}_{\text{high}} = \|X_{\text{high}} - \mathcal{H}_{n_2}^{-1}\mathcal{M}(\mathcal{H}_{n_2}(X_{\text{high}}))\|_2. \tag{5}$$

Similarly, the loss function for the low-frequency model is

$$\mathcal{L}_{\text{low}} = \|X_{\text{low}} - \mathcal{H}_{n_1}^{-1}\mathcal{M}(\mathcal{H}_{n_1}(X_{\text{low}}))\|_2. \tag{6}$$

The reconstructed loss is the weighted sum of high-frequency and low frequency ℓ_2 loss, i.e.,

$$\mathcal{L}_{\text{twogrid}} = \alpha\mathcal{L}_{\text{high}} + (1-\alpha)\mathcal{L}_{\text{low}}, \tag{7}$$

where $\alpha \in [0,1]$. If we use $\alpha = 0$, we learn only with the high-frequency components. For $\alpha = 1$, the model is trained solely on low-frequency components.

Algorithm 1: **ADATG**:Adaptive Frequency-Decomposed Two-Grid Masked Auto-Encoder Training

Require: Seismogram image set $\mathcal{I}$, learning rate η, epochs E, batch-size B

Require: Coarse/fine Hilbert orders n_1, n_2, spectral threshold k_0

Ensure: Pretrained foundation model $\mathcal{M}$

Initialize $\mathcal{H}_{n_1}$, $\mathcal{H}_{n_2}$ {Coarse/fine Hilbert curves}
for each epoch $e \in \{0, \ldots, E-1\}$ **do**
  $\mathcal{B} \leftarrow \text{Sample}(\mathcal{I}, B)$ {Random minibatch}
  $X_{\text{high}}, X_{\text{low}} \leftarrow \text{FreqDecomp}(\mathcal{B}, k_0)$ {Eq. (1)-(3)}
  $E_{\text{high}} \leftarrow \mathcal{H}_{n_2}(X_{\text{high}})$ {Fine-grid encoding}
  $E_{\text{low}} \leftarrow \mathcal{H}_{n_1}(X_{\text{low}})$ {Coarse-grid encoding}
  $\mathcal{L} \leftarrow \text{AdaptiveMAELoss}(E_{\text{high}}, E_{\text{low}})$ {Eq. (8)}
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ {Parameter update}
  $\hat{I}_{\text{high}}, \hat{I}_{\text{low}} \leftarrow \text{Reconstruct}(E_{\text{high}}, E_{\text{low}})$
end for

3.6 Adaptive Frequency Decomposed Twogrid Approach

Motivated by the frequency principle (Rahaman et al. 2019; Xu, Zhang, and Xiao 2019; Xu et al. 2019; Zhang et al. 2019; Luo et al. 2019), a deep neural network (DNN) tends to learn a target function starting at low frequencies and progressing to high frequencies during training. In this section, we propose an adaptive training strategy specifically for training two-grid masked auto-encoder (MAE) models. The loss function is defined as follows:

$$\mathcal{L}_{\text{adap}} = \alpha_t \mathcal{L}_{\text{high}} + (1 - \alpha_t)\mathcal{L}_{\text{low}}, \tag{8}$$

where $\alpha_t = \alpha t/T$ is a linear decay function that depends on the ratio of the current iteration step to the total number of training steps, t/T. Consequently, $1 - \alpha_t$ represents a linear increase. According to this formulation, the neural network initially focuses more on the low-frequency components. As training progresses, it gradually shifts its attention to the high-frequency components, aligning with the frequency principle discussed earlier. In conclusion, we summarize **ADATG** methods in Algorithm 1.

4 Experimental Analysis

In this section, we implement several experiments to answer the following research questions.

- **RQ.1** How does Hilbert-ViT outperform the base version of ViT, and what specific improvements does it achieve?
- **RQ.2** What is the choice for threshold of high- and low-frequency components?
- **RQ.3** What roles do high-frequency and low-frequency components play in the analysis and interpretation of the original input seismogram?
- **RQ.4** How does the frequency principle inform and guide the development of an adaptive training strategy?
- **RQ.5** What measurable improvements does the adaptive training strategy provide?

4.1 Setup

Data Preparation This section outlines the data preparation process for training the foundation model, following the procedures described in SFM (Sheng et al. 2025). The original three-dimensional seismic datasets are sourced from the United States Geological Survey (USGS) (U.S. Geological Survey 2025), the South Australian Resources Information Gateway (SARIG) (Government of South Australia 2025), and the Society of Exploration Geophysics (SEG) (Society of Exploration Geophysicists 2025). **SFM** (Sheng et al. 2025) performs two-dimensional seismic slicing in both inline and cross-line directions, resulting in over 2,200,000 segments, each measuring 244×244 pixels. To enhance data diversity, we select 111,110 samples for training and 7,814 samples for testing. Additionally, we improve the input data resolution from 224×224 to 256×256 using reflection padding for the two-grid Hilbert encoding.

Training Details We illustrate the model structure in Figure 2 (C), featuring an encoder with 12 Transformer blocks and a decoder with 4 Transformer blocks. The total parameter size is 84.57 MiB, with 81.82 MiB of trainable parameters. Key hyperparameters for **ADATG** include a batch size (B) of 336, a mask ratio of 0.75, 1,600 training epochs (E), a learning rate (η) of 1.5×10^{-5}, and a weight decay of 0.05. Experiments were conducted on a server with four NVIDIA A100 GPUs, each having 80 GB of memory.

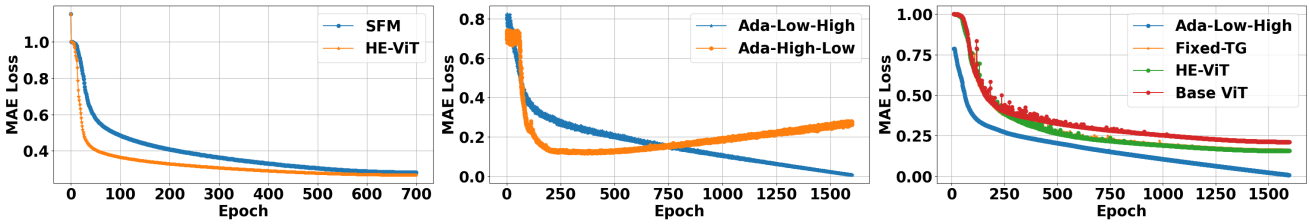
Comapred Methods **SFM** refers to the baseline ViT methods from (Sheng et al. 2025). **HE-ViT** is one grid ViT with Hilbert Encoding; **Fixed-TG** is a fixed two-grid ViT using Hilbert Encoding; **Ran-TG** is a two-grid ViT fine grid randomly sampled. **High** indicates training on high-frequency components only, while **Low** focuses on low-frequency components. Lastly, **ADATG-HH** is **ADATG** with both fine and coarse grid encoding, and **ADATG-NH** has only fine grid encoding.

4.2 Improvement by Hilbert Encoding (RQ.1)

In this section, we show the improvement of the Hilbert encoding compared to **SFM** in the pre-training stage of the foundation model in Figure 5a. We observe that the Hilbert encoding (**HE-ViT**) accelerates the training process with a faster convergence and arrives loss around 0.3104 within the first 200 epochs from From Figure 5a (a). As training progresses, the **SFM** arrives with a similar performance as **HE-ViT** and finally reaches the similar convergence point at epoch 700. The final loss of **SFM** at epoch 700 is 0.2817 , while the final loss of the **HE-ViT** at the same epoch is 0.2673. With Hilbert Encoding, one gains the loss descend of 0.0144. However, although based on Hilbert encoding ViT, we obtain faster convergence and smaller loss compared to **SFM**, **HE-ViT** fails to improve the metrics such as MSE, PSNR, and SSIM for the test samples. Thus, we discuss the two-grid approach for better Hilbert Encoding.

4.3 Spectral Decomposition Threshold (RQ.2)

By spectral Decomposition described in Section 3.1, one can obtain both informative high- and low-frequency components for the seismogram. Furthermore, we record the



(a) Comparison of **SFM** and **HE-ViT**. (b) Frequency Principle for ViT. (c) Improvement of Adaptive Strategy.

Figure 5: Training Dynamics.

Methods	**SFM** (Sheng et al. 2025)	**HE-ViT**	**Fixed-TG**	**Ran-TG**
MAE Loss ↓	0.1886	0.1535	0.1634	0.1665
MSE ↓	1.6383 ± 0.3932	2.0629 ± 0.0965	1.8542 ± 0.0966	1.9132 ± 0.0791
PSNR ↑	18.4969 ± 2.1720	15.7355 ± 1.7805	16.2393 ± 1.8034	13.4243 ± 1.3693
SSIM ↑	0.1308 ± 0.0745	0.0620 ± 0.0425	0.0826 ± 0.0413	0.0620 ± 0.0404
MS-SSIM ↑	0.2943 ± 0.2054	0.0731 ± 0.0693	0.0885 ± 0.0797	0.0327 ± 0.0510
Methods	**High**	**Low**	**ADATG-HH**	**ADATG-NH**
MAE Loss ↓	0.0667	0.2701	**0.0068**	0.1291
MSE ↓	1.6409 ± 0.2595	0.4583 ± 0.4349	**0.3988 ± 0.1544**	1.4363 ± 0.3771
PSNR ↑	6.2952 ± 2.6248	16.6024 ± 1.6867	**27.3535 ± 2.890**	19.9746 ± 2.1178
SSIM ↑	0.0807 ± 0.0567	0.0286 ± 0.0149	**0.7362 ± 0.0540**	0.2640 ± 0.0782
MS-SSIM ↑	0.0441 ± 0.0537	0.0471 ± 0.0645	**0.3890 ± 0.3582**	0.3131 ± 0.2605

Table 1: Overall performance of different foundation models pre-training.

Threshold k_0	4	8	16	32	64
Low	32.81	86.03	173.07	228.05	253.14
High	254.22	240.49	181.42	105.56	36.60
Original	256.63	256.63	256.63	256.63	256.63

Table 2: Distribution of frequency-component energy under ℓ_2-norm, based on 998 training samples.

ℓ_2 norm for high- and low-frequency compounds in Table 2, which confirms the observation. When the threshold is small, for example, $k_0 = 4$, the high frequency occupies the main energy. However, for a large threshold $k_0 = 64$, most energy is concentrated in the low frequency component. For $k_0 = 16$, the energy for low frequency is 173.07 while for high frequency is 181.42. The high- and low-frequency components show similar energy. Then we choose the frequency threshold as $k_0 = 16$ through the experimental parts.

4.4 High and Low Frequency Components (RQ.3)

In this section, we present the results for learning both high-frequency and low-frequency components. We utilize the fast Fourier transform to effectively separate these two components. For our implementation, we choose a spectral decomposition threshold of 16. Independent reconstructions of the high-frequency and low-frequency models are illustrated in Figures 6e and 6f.

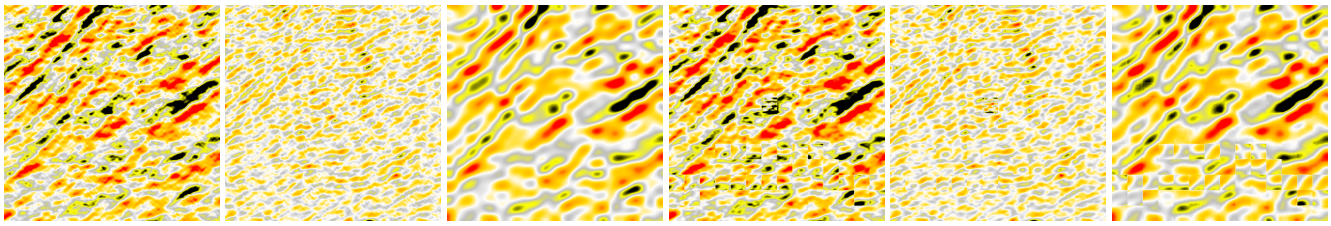
The high-frequency components are associated with rapid changes in ground motion and correspond to shorter wavelengths. In contrast, low-frequency components relate to slower variations in ground motion and are linked to longer wavelengths (Yilmaz 2001; Sheriff and Geldart 1995). High-frequency components are valuable for fine-structure identification, high-resolution imaging, and lithological analysis. Meanwhile, low-frequency components are useful for deep structure detection, velocity model construction, and full waveform inversion (Virieux and Operto 2009; Pratt 1999). By integrating high-frequency and low-frequency information, we can enhance multiscale imaging and improve the output image quality via training of MAE task (Table 1).

4.5 Frequency Principle (RQ.4)

In this section, we experimentally verify the frequency principle for the ViT architecture. We use the same structure as the ViT model for training **ADATG**, as shown in Fig. 2 part C. In contrast to **ADATG**, we include a comparative experiment that begins with training with the high-frequency components and then transitions to the low-frequency components at a linear rate. The loss function is defined as follows:

$$\mathcal{L}_{\text{reverse}} = (1 - \alpha_t)\mathcal{L}_{\text{high}} + \alpha_t \mathcal{L}_{\text{low}}, \tag{9}$$

where $\alpha_t = \alpha t/T$ serves as a linear decay coefficient for the high-frequency reconstructed loss, and $1 - \alpha_t$ indicates the linear increase coefficient for the low-frequency reconstructed loss. To distinguish between models, we denote **ADATG** with two grid encodings and an adaptive training loss (Equation (8)) as **ADATG-HH** (Ada-High-Low), and **ADATG** with two grid encodings and a reverse adaptive loss (Equation (9)) as **ADATG** (Ada-Low-High).



(a) **Original** (b) **High** (c) **Low** (d) **Recons.** (e) **Recons.-High** (f) **Recons.-Low**

Figure 6: Decomposition and Reconstruction of High- and Low-Frequency Components. The left part shows the original image (a), its high-frequency component (b), and its low-frequency component (c). The right part displays the reconstructed image (d), with the reconstructed high-frequency (e) and low-frequency (f) components.

Metrics	MAE Loss ↓	MSE ↓	PSNR ↑	SSIM ↑	MS-SSIM ↑
ADATG-HH (Ada-Low-High)	**0.0068**	**0.3988**	**27.3525**	**0.7362**	**0.3890**
ADATG-HH (Ada-High-Low)	0.2769	4.1129	15.1371	0.0351	0.0454
ADATG-NH	0.1291	1.4363	19.9747	0.2640	0.3131

Table 3: Comparison of adaptive methods. Best results are indicated in bold.

Figure 5b illustrates the trend of loss using these two adaptive strategies. The strategy labeled as 'Ada-High-Low' refers to the approach where we initially focus entirely on high-frequency components and then gradually shift our attention to low-frequency components in a linear manner. Conversely, 'Ada-Low-High' describes the reverse strategy, where we start by concentrating on low-frequency components and then transition to high-frequency components gradually. As shown in Figure 5b, the 'Ada-Low-High' approach causes the loss to decrease significantly within a few epochs before increasing again after reaching its lowest point. In contrast, the 'Ada-High-Low' strategy decreases more slowly compared to 'Ada-Low-High' but continues to exhibit a decreasing trend even after reaching the end of training at 1,600 epochs. The intersection between the two curves occurs around 800 epochs. In conclusion, the observations presented in Figure 5b support the frequency principle for ViT blocks. Specifically, the model tends to learn low-frequency information at the beginning of training and gradually shifts to high-frequency information.

4.6 Improvement of Adaptive Strategy (RQ.5)

In this section, we discuss the improvements made to the adaptive strategy. Figure 5c compares the performance of adaptive two-grid methods using the low to high strategy **ADATG-HH** (Ada-Low-High) with **SFM**, **HE-ViT**, and **Fixed-TG**, all evaluated over a long training period of 1,600 epochs. As shown in Table 1, **SFM** achieves a loss of approximately 0.1886 after sufficient training epochs. The performance of the **HE-ViT** and **Fixed-TG** is quite similar, with final Mean Absolute Errors (MAE) of 0.1535 and 0.1634 respectively. In particular, **Fixed-TG** outperforms the Randomized Two-Grid methods. Using the adaptive training strategy, the MAE loss of **ADATG** (Ada-Low-High) is significantly reduced to around 0.0068, marking a decrease of up to 96.39% compared to **SFM**. Furthermore, we observe that this method converges more quickly and has a lower initial MAE loss. The blue curves representing Ada-Low-High are consistently positioned below all other colored curves. Moreover, we show the detailed comparison for **ADATG-NH**, denoting **ADATG** with a plain encoding for the low-frequency component. Although **ADATG-NH** shows good visualization in Figure 3, it performs worse than the evaluation metrics in Table 3. This may be induced by the visualization is mainly influenced via high-frequency information, while the low frequency plays a more important role in evaluation scores.

5 Conclusion

Driven by the intrinsic nature of seismograms, which contain both high- and low-frequency components, we developed a two-grid approach (**ADATG**) to enhance the training efficiency of our seismic foundation model. First, we decompose the input seismogram into its high-frequency and low-frequency components. Next, we apply hierarchical Hilbert encoding to both components to preprocess the data for the ViT model. We calculate the training loss as a linear combination of the losses from the high-frequency and low-frequency components. We then improve the two-grid training method by implementing an adaptive construction of the training loss. Finally, we conduct extensive experiments to demonstrate the effectiveness of our proposed methods. This approach can also be applied to various computer vision scenarios where high-frequency and low-frequency features differ significantly, yet both play an essential role. In the future, we plan to expand our model to pre-train a seismic foundation model using 3D seismograms as input. We will also explore how this pre-trained model can improve downstream tasks, including full waveform inversion. By doing so, we aim to provide more effective feature representations and create an initial model that closely approximates the actual scenario in full waveform inversion task.

Acknowledgments

This work is supported by the **Zhejiang Province Key Research and Development Plan** (Grant No. 2025SSYS0004, Grant No. 2024C01036) and the **National Natural Science Foundation of China** (Grant No. 92370205, Grant No. 62571529, and Grant No. 12271512).

References

Aki, K.; and Richards, P. G. 2002. *Quantitative Seismology*. Sausalito, California: University Science Books, 2nd edition. This foundational textbook discusses the relationship between frequency content and the sharpness of seismic phases, emphasizing how high-frequency components are critical for resolving detailed features in seismic waveforms.

Bhupati, L. R. T.; et al. 2019. *Classification of fMRI Brain Activation Maps by Using Space Filling Curves*. Ph.D. thesis.

Bially, T. 1969. Space-filling curves: Their generation and their application to bandwidth reduction. *IEEE Transactions on Information Theory*, 15(6): 658–664.

Butz, A. R. 1969. Convergence with Hilbert's space filling curve. *Journal of Computer and System Sciences*, 3(2): 128–146.

Chen, L.; and Wu, H. 2018. A Preconditioner based on Non-uniform Row Sampling for Linear Least Squares Problems. *arXiv preprint arXiv:1806.02968*.

Chen, N.; Wang, N.; and Shi, B. 2007. A new algorithm for encoding and decoding the Hilbert order. *Software: Practice and Experience*, 37(8): 897–908.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fichtner, A. 2010. *Full Seismic Waveform Modelling and Inversion*. Springer.

Government of South Australia. 2025. SARIG: Resource and Energy Georeference Databases. South Australian Resources Information Gateway.

Han, Y.; Chen, H.; Yao, L.; Li, K.; and Yin, J. 2024. MAT-VIT: A Vision Transformer with MAE-Based Self-Supervised Auxiliary Task for Medical Image Classification. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2046–2052. IEEE.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, S.; Yang, X.; Cao, A.; Wang, C.; Liu, Y.; Liu, Y.; and Niu, Q. 2024. SeisT: A foundational deep learning model for earthquake monitoring tasks. *IEEE Transactions on Geoscience and Remote Sensing*.

Lindsey, N. J.; and Martin, E. R. 2021. Fiber-optic seismology. *Annual Review of Earth and Planetary Sciences*, 49(1): 309–336.

Liu, J.; Huang, X.; Zheng, J.; Liu, Y.; and Li, H. 2023. MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6252–6261.

Liu, T.; Münchmeyer, J.; Laurenti, L.; Marone, C.; de Hoop, M. V.; and Dokmanić, I. 2024. SeisLM: a Foundation Model for Seismic Waveforms. *arXiv preprint arXiv:2410.15765*.

Luo, T.; Ma, Z.; Xu, Z.-Q. J.; and Zhang, Y. 2019. Theory of the frequency principle for general deep neural networks. *arXiv preprint arXiv:1906.09235*.

Moon, B.; Jagadish, H. V.; Faloutsos, C.; and Saltz, J. H. 2001. Analysis of the clustering properties of the Hilbert space-filling curve. *IEEE Transactions on knowledge and data engineering*, 13(1): 124–141.

Pratt, R. G. 1999. Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model. *Geophysics*, 64(3): 888–901.

Pyzer-Knapp, E. O.; Manica, M.; Staar, P.; Morin, L.; Ruch, P.; Laino, T.; Smith, J. R.; and Curioni, A. 2025. Foundation models for materials discovery–current state and future directions. *Npj Computational Materials*, 11(1): 61.

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

Ringler, A. T.; Anthony, R. E.; Davis, P.; Ebeling, C.; Hafner, K.; Mellors, R.; Schneider, S.; and Wilson, D. C. 2022. Improved resolution across the Global Seismographic Network: A new era in low-frequency seismology. *The Seismic Record*, 2(2): 78–87.

Sheng, H.; Wu, X.; Si, X.; Li, J.; Zhang, S.; and Duan, X. 2025. Seismic foundation model: A next generation deep-learning model in geophysics. *Geophysics*, 90(2): IM59–IM79.

Sheriff, R. E.; and Geldart, L. P. 1995. *Exploration Seismology*. Cambridge University Press.

Si, X.; Wu, X.; Sheng, H.; Zhu, J.; and Li, Z. 2024. SeisCLIP: A seismology foundation model pre-trained by multi-modal data for multi-purpose seismic feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*.

Society of Exploration Geophysicists. 2025. Open Data on the SEG Wiki. A collection of open data resources for exploration geophysics.

Stein, S.; and Wysession, M. 2003. *An Introduction to Seismology, Earthquakes, and Earth Structure*. Malden, MA: Blackwell Publishing. This book provides a detailed discussion of seismic wave propagation and how high-frequency components enhance the resolution of seismic phase arrivals.

Tarantola, A. 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM.

Trefethen, L. N.; and Bau, D. 1997. *Numerical Linear Algebra*. Philadelphia, PA: SIAM. ISBN 978-0-89871-361-9.

U.S. Geological Survey. 2025. The National Archive of Marine Seismic Surveys (NAMSS). A search tool for marine seismic survey data provided by the U.S. Geological Survey.

Virieux, J.; and Operto, S. 2009. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6): WCC1–WCC26.

Wesseling, P. 1995. Introduction to multigrid methods. Technical report.

Xin, H.; Guo, D.; Shao, Z.; Ren, Z.; Zhu, Q.; Liu, B.; Ruan, C.; Li, W.; and Liang, X. 2024. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*.

Xu, J.; and Zikatanov, L. 2017. Algebraic multigrid methods. *Acta Numerica*, 26: 591–721.

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

Xu, Z.-Q. J.; Zhang, Y.; Luo, T.; Xiao, Y.; and Ma, Z. 2019. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*.

Xu, Z.-Q. J.; Zhang, Y.; and Xiao, Y. 2019. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, 264–274. Springer.

Yilmaz, Ö. 2001. *Seismic Data Analysis: Processing, Inversion, and Interpretation of Seismic Data*. Society of Exploration Geophysicists.

Zhang, X.; Zhang, O.; Shen, C.; Qu, W.; Chen, S.; Cao, H.; Kang, Y.; Wang, Z.; Wang, E.; Zhang, J.; et al. 2023. Efficient and accurate large library ligand docking with KarmaDock. *Nature Computational Science*, 3(9): 789–804.

Zhang, Y.; Xu, Z.-Q. J.; Luo, T.; and Ma, Z. 2019. Explicitizing an implicit bias of the frequency principle in two-layer neural networks. *arXiv preprint arXiv:1905.10264*.