Efficient Seismic Data Interpolation via Sparse Attention Transformer and Diffusion Model

Xiaoli Wei, Chunxia Zhang, *Member, IEEE*, Baisong Jiang, Anxiang Di, Deng Xiong, Jiangshe Zhang, Mingming Gong, *Member, IEEE*

***Abstract*—Seismic data interpolation is a critical preprocessing step for improving seismic imaging quality and remains a focus of academic innovation. To address the computational inefficiencies caused by extensive iterative resampling in current plug-and-play diffusion interpolation methods, we propose the diffusion-enhanced sparse attention transformer (Diff-spaformer), a novel deep learning framework. Our model integrates transformer architectures and diffusion models via a Seismic Prior Extraction Network (SPEN), which serves as a bridge module. Full-layer sparse multi-head attention and feed-forward propagation capture global information distributions, while the diffusion model provides robust prior guidance. To mitigate the computational burden of high-dimensional representations, self-attention is computed along the channel rather than the spatial dimension. We show that using negative squared Euclidean distance to compute sparse affinity matrices better suits seismic data modeling, enabling broader contribution from amplitude feature nodes. An adaptive ReLU function further discards low or irrelevant self-attention values. We conduct training within a single-stage optimization framework, requiring only a few reverse diffusion sampling steps during inference. Extensive experiments demonstrate improved interpolation fidelity and computational efficiency for both random and continuous missing data, offering a new paradigm for high-efficiency seismic data reconstruction under complex geological conditions.**

***Index Terms*—Seismic data interpolation, denoising diffusion model, transformer, negative squared Euclidean distance, sparse attention**

I. Introduction

SEISMIC data acquired in exploration surveys often suffer from spatial irregularity due to economic constraints and environmental factors, while complete regular seismic records are mandatory to the foundation for successful imaging and inversion in industry [1]. Interpolation serves as a critical step to enhance the resolution and signal coherence of migrated seismic images and reduce artifacts caused by missing data, as it is capable of reconstructing complete wavefields.

Traditional methods infer missing values by exploring the mathematical or physical properties of seismic data, leveraging spatial relationships and variation patterns of known data points. Predictive filter-based approaches employ spatial correlations between adjacent traces to estimate missing values through adaptive filters optimized to minimize prediction errors [2], [3], [4]. Wave equation-based methods reconstruct missing seismic data by inverting subsurface velocity models or wavefield information from available data, followed by forward simulations to generate complete wavefields that fill data gaps through physics-driven interpolation [5], [6]. Transform-based methods project missing seismic data into mathematical transformation domains (e.g., 1D or multidimensional Fourier transform [7, 8], [9], sparsity-enhanced Curvelet transform [10], [11], and Radon transform [12]) and exploit sparse signal priors to recover gaps through domain-specific optimization. Rank-reduction methods interpolate seismic data by leveraging low-rank structures and linear correlations in transform domains, recovering missing traces via matrix completion or nuclear norm optimization [13, 14].

The above traditional methods rely on rigid assumptions and prior knowledge, limiting their adaptability. Data-driven approaches leverage adaptive learning to bypass these constraints, emerging as popular solutions in contemporary seismic interpolation. The symmetric encoder-decoder architecture of U-Net establishes a stable mapping mechanism for seismic data processing [15], [16], [17], making it a dominant framework for interpolation tasks [18], [19], [20]. Recent U-Net enhancements integrate physics-aware loss functions [21], partial convolution layers (adaptive feature focusing via validity-guided attention) [22], and wavelet-based downsampling (preserving high-frequency components) [23]. GAN-based methods improve seismic interpolation via adversarial learning [24], [25], where discriminator-guided training drives the generator toward data distribution alignment, enhancing geological fidelity. Multi-branch [26] and multi-stage interpolation methods [27], [28], [29] hierarchically integrate diverse feature extraction paths with stepwise optimization, enhancing accuracy through adaptive fusion of seismic attributes and phased refinement. Dual-domain conditional generative adversarial network (DD-CGAN) [30] utilizes joint supervision of spatiotemporal and time-frequency domains in seismic data feature restoration, balancing energy distribution optimization in the time-frequency domain with waveform variation preservation in the spatiotemporal domain. The Coarse-to-Fine model [31] employs depth-varied feature extractors and differentiated supervision for progressive low/high amplitude interpolation. Attention-based methods improve feature saliency by capturing global context and long-range depen-

Corresponding author: Chunxia Zhang. E-mail: cxzhang@mail.xjtu.edu.cn.

Xiaoli Wei, Chunxia Zhang, Baisong Jiang, Anxiang Di, Jiangshe Zhang are with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China.

Deng Xiong is with the Geophysical Technology Research and Development Center, BGP, Zhuozhou, Hebei, 072751, China

Mingming Gong is with the School of Mathematics and Statistics, Faculty of Science, The University of Melbourne, Melbourne, VIC 3010, Australia.

This research was supported by the National Natural Science Foundation of China (No. 12371512), the Science & Technology Research and Development Project of CNPC (Grant No. 2021ZG03), and the Key Acquisition, Processing and Interpretation Techniques for Seismic Data Processing of CNPC (Grant No. 01-03-2023)

dencies [32], [33], [34], addressing the locality limitation of convolutions, which is crucial for reconstructing continuous missing seismic data. The transformer enhances this advantage with multi-head attention [35], enabling parallel multi-scale seismic feature interactions. Physics-informed and physics-guided models significantly enhance solution stability in data-driven frameworks through convex constraint enforcement. One of the most well-known convex optimization frameworks, projection onto convex sets (POCS), has been extensively studied and combined with deep neural networks to solve interpolation inverse problems [36], [37]. Physics-Informed Neural Networks (PINN) offer a new paradigm that explicitly constrains the continuity of the interpolation process using plane wave differential equations, promoting reconstructions naturally similar to the available data [38], [39].

Recently, diffusion models for interpolation have become a research focus [40], [41], bridging deep learning with iterative frameworks. Using Langevin dynamics and a predefined variance schedule, missing seismic data is mapped into a noisy distribution flow, enabling a deep neural network to learn nonlinear mappings across noise levels. However, optimizing valid data usage in reverse conditional interpolation remains a key challenge. While resampling and conditional generation strategies have shown promise in the prior study [42], the plug-and-play approach falls short in complex seismic data reconstruction. Subsequent work has thus incorporated self-correction mechanisms [43], posterior distribution through Langevin dynamics [44], and constrained training paradigms [45], [46], [47] to enhance interpolation accuracy and stability. However, these methods inevitably face a common challenge: prolonged inference time due to multiple sampling iterations. Furthermore, the presence of various high-performance feature extractors, especially the remarkable ability of multi-head attention mechanisms to effectively model and manage both random and continuous missing data scenarios, deserves significant attention. This insight motivates the integration of transformer architectures with diffusion models. In this paper, we propose the diffusion-enhanced sparse attention transformer (Diff-spaformer), which synergistically combines multi-scale transformer networks for efficient long-sequence modeling with a diffusion-based prior generator to maintain distribution consistency, enabling efficient interpolation of missing seismic data. Our method outperforms some existing diffusion-based approaches in both computational efficiency and accuracy. The main contributions are summarized below.

- We propose a novel framework that integrates transformer networks with diffusion models through a bridge component called the Seismic Prior Extraction Network (SPEN). Full-layer sparse multi-head attention and feed-forward propagation synergistically capture global information distributions, while the diffusion model provides strong prior guidance to enhance reconstruction quality.
- To reduce the computational complexity of high-dimensional space representations, we reorient the self-attention computation from the spatial dimension to the channel dimension. We adopt the negative squared Euclidean distance as similarity functions to compute this sparse affinity matrix, facilitating more feature nodes to make contributions. The improved sparse transformer block can model the global pixel relationships with higher efficiency and effectiveness.
- We use a simple yet effective ReLU function as the activation function to adaptively remove the self-attention values with low/no correlation.
- Extensive experimental results show that only adopting a single-stage unified multi-component optimization, our method can deliver superior performance while avoiding excessive computational overhead.

The content of this paper is outlined as follows. Section II introduces the background. Section III proposes the methodological framework and describes the detailed components. Section IV presents experimental results. In Section V, we have conducted ablation studies, and Section VI concludes this paper.

II. Background

Denoising Diffusion Probabilistic Models (DDPM), as defined in [48], achieve complex distribution modeling by learning a progressive noise-adding and denoising process of data distributions. The progressive noising process constitutes a predefined T-step forward process, and the denoising process executes a T-step reverse process to construct desired data samples from noises.

A. Forward Process

In the forward process, the noise-adding operations between adjacent timesteps are designed as a Markov chain $\boldsymbol{x}_0 \rightarrow \boldsymbol{x}_1 \rightarrow \ldots \rightarrow \boldsymbol{x}_{T-1} \rightarrow \boldsymbol{x}_T$ (where $\boldsymbol{x}_t \in \boldsymbol{R}^n$). The state transition from the previous state $\boldsymbol{x}_{t-1}$ to the current state $\boldsymbol{x}_t$ is mathematically defined by a conditional Gaussian distribution parameterized with scheduled variance coefficients β_t as

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) := \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{1}$$

i.e.,

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

The reparameterization technique integrates these multi-step stochastic processes into the following closed-form expression as

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right), \tag{3}$$

where $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$.

B. Reverse Process

Each step in the reverse process achieves state transition through a conditional probability distribution $p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right)$. According to Bayes theorem, it can be derived from Eq. (1) and Eq. (3) as

$$p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right), \sigma_t^2 \mathbf{I}\right), \tag{4}$$

where

$$\boldsymbol{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right)$$

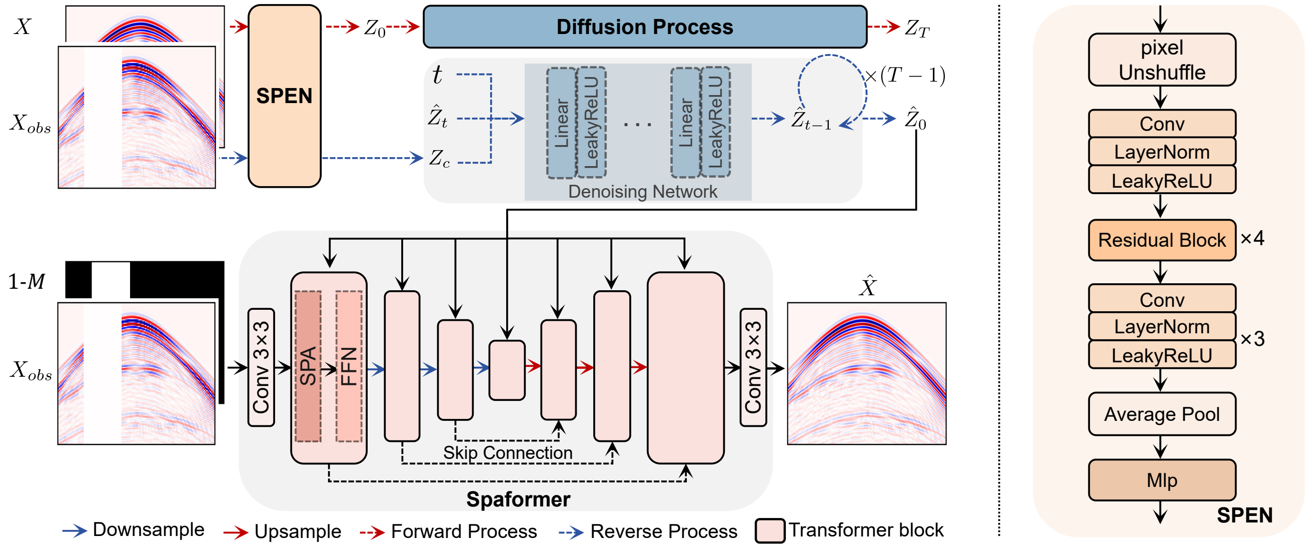


Fig. 1. The overall framework of Diff-spaformer. It comprises three components, i.e., SPEN, diffusion process, and Spaformer. SPEN encodes seismic prior information and generates latent features through the diffusion model. The Spaformer module receives the concatenation of the missing data $\boldsymbol{X}_{obs}$ and the missing mask 1-M as the input while integrating seismic prior knowledge extracted from SPEN at each encoding-decoding layer. This hierarchical fusion mechanism ultimately outputs the interpolated data $\hat{\boldsymbol{X}}$.

and

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t.$$

The θ-parameterized neural network estimates the noise component $\boldsymbol{\epsilon}_t$ to recover the noiseless observation from the noisy input $\boldsymbol{x}_t$, i.e., $\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$. The Gaussian form of the target distribution in Eq. (4) simplifies the measurement of the Kullback-Leibler (KL) divergence, enabling efficient variational training through the optimization of the evidence lower bound (ELBO). The simplified training objective used in DDPM is formulated as

$$\mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}\left[\left\|\epsilon - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|_2^2\right]. \tag{5}$$

where $\|\cdot\|_2$ symbolizes L2 norm and t is uniformly distributed over the interval from 1 to T.

III. Methodology

Given the complete seismic data denoted as $\boldsymbol{X} \in \mathbb{R}^{n_s \times n_t}$ (where n_s and n_t represent the spatial and temporal dimensions, respectively) and the degraded observation data $\boldsymbol{X}_{obs}$ containing missing traces, the interpolation process is typically modeled as an inverse problem:

$$\boldsymbol{X}_{obs} = \boldsymbol{M} \odot \boldsymbol{X}, \quad \text{with} \quad \boldsymbol{M}[i,:] = \begin{cases} \boldsymbol{1}, & \text{if } i \text{ is valid} \\ \boldsymbol{0}, & \text{else} \end{cases} \tag{6}$$

where $\odot$ represents the element-wise multiplication and $\boldsymbol{M}$ denotes the binary mask matrix with continuous or random missing traces. Our proposed Diff-spaformer model aims to learn the nonlinear relationship between degraded observations and fully sampled seismic data through the parameterized mapping process:

$$\hat{\boldsymbol{X}} = f_\theta(\boldsymbol{X}_{obs}, \boldsymbol{Z}), \tag{7}$$

where $\boldsymbol{Z}$ represents the seismic prior feature. Fig. 1 demonstrates the specific workflow of our proposed Diff-spaformer model. Spaformer is built on the U-Net architecture, equipped with transformer modules, and the diffusion process further provides compressed prior information to guide Spaformer. SPEN serves as a bridge between Spaformer and the diffusion process. Section III-A, Section III-B, and Section III-C will introduce the details of the three parts in turn.

A. Seismic Prior Extraction Network (SPEN)

SPEN operates externally to the Spaformer architecture, focusing on capturing inherent seismic characteristics like amplitude patterns and spectral properties. This auxiliary module supplements the primary encoder-decoder framework by integrating domain-specific structural constraints, thereby optimizing reconstruction accuracy through enhanced distribution modeling of seismic signals. We adopt the prior network design used in previous restoration works [49], as shown on the right side of Fig. 1. Given the inherent high spatial continuity of seismic data properties (e.g., amplitude structure and frequency distribution), we implement spatial-to-depth transformation through PixelUnshuffle downsampling. This operation effectively preserves complete signal energy while extracting multi-scale features through spatial dimension reorganization, circumventing detail loss inherent in conventional pooling methods through channel expansion rather than spatial compression. The network employs a convolutional-residual structure for stable feature extraction, where convolutional layers capture local spatial correlations and residual layers enhance gradient propagation stability through cross-layer connections. Then, average pooling is applied to the feature maps for spatial compression, reducing dimensionality while retaining global statistical features and suppressing local noise. Finally, a multilayer perceptron (MLP) performs nonlinear transformations and high-order feature fusion on the compressed features. The SPEN generates compressed prior information $\boldsymbol{Z}_0 \in \mathbb{R}^{C'}$ that represents the energy distribution of seismic signals as

$$\boldsymbol{Z}_0 = \mathrm{SPEN}\left(\boldsymbol{X}\right). \tag{8}$$

At the same time, conditional prior information $\boldsymbol{Z}_c \in \mathbb{R}^{C'}$ for the diffusion process is provided by encoding features from the observed data as follows:

$$\boldsymbol{Z}_c = \mathrm{SPEN}\left(\boldsymbol{X}_{obs}\right). \tag{9}$$

In diffusion model training, the original data prior $\boldsymbol{Z}_0$ and conditional prior $\boldsymbol{Z}_c$ are critical for optimizing the iterative process. The original distribution captures the intrinsic probability characteristics of the source data, while the conditional prior encodes domain-specific constraints or observed patterns from incomplete datasets through feature extraction mechanisms. This dual-input way enables the model to learn conditional distributions guided by known information during optimization.

B. *Diffusion Process*

We exploit the T-steps diffusion forward process $\boldsymbol{Z}_0 \rightarrow \boldsymbol{Z}_1 \rightarrow \ldots \rightarrow \boldsymbol{Z}_{T-1} \rightarrow \boldsymbol{Z}_T$ by executing the predefined progressive noise-adding operation

$$q\left(\boldsymbol{Z}_t \mid \boldsymbol{Z}_0\right) = \mathcal{N}\left(\boldsymbol{Z}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{Z}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right), \tag{10}$$

where $\bar{\alpha}_t$ has been defined in Eq. (3) and $\beta_t \in (0,1)$ follows a linear increasing schedule. The reverse process starts from T-th time step and $\boldsymbol{Z}_T$ is randomly sampled from the standard Gaussian distribution. Since the seismic priors $\boldsymbol{Z}_0$ have been encoded by SPEN into compact features, the complex distribution of the original seismic data is effectively projected into a low-dimensional latent space. These latent space features serve as the strong valid priors, effectively guiding the diffusion model's training and allowing us to capture key noise patterns using a simple and small network without complex architectures. As shown in Fig. 1, this noise-matching network $\boldsymbol{\epsilon}_\theta$ is only composed of multiple linear layers. During the reverse process, it uses the concatenated inputs of t, $\hat{\boldsymbol{Z}}_t$, and $\boldsymbol{Z}_c$ to directly predict the raw data corresponding to the current time step t as

$$\hat{\boldsymbol{Z}}_0^t = \epsilon_\theta\left(\left[t, \hat{\boldsymbol{Z}}_t, \boldsymbol{Z}_c\right]\right), \tag{11}$$

where t is embedded by using $t = \frac{t}{T}$. The final compact prior feature $\hat{\boldsymbol{Z}}_0$ can be obtained by performing iterative generation

$$\hat{\boldsymbol{Z}}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\boldsymbol{Z}}_0^t + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\hat{\boldsymbol{Z}}_t, \tag{12}$$

where the noise term is excluded by reducing stochasticity to enhance generation efficiency and stability, particularly in our scenarios with simpler model architectures and smaller scales. Due to the long sampling time T in traditional diffusion model training, loss function computation and gradient backpropagation are typically carried out using randomly sampled sequences. However, with the smaller and simpler structure in our case, we can compute the loss over all time steps during each training iteration as

$$\mathcal{L}_{\text{diff}} = \frac{1}{T-1}\sum_{t=0}^{T-1}\|\boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_t\|_1, \tag{13}$$

where $\|\cdot\|_1$ represents the L1 norm, facilitating the optimization to obtain an even sharper distribution.

In the inference phase, SPEN directly encodes missing data $\boldsymbol{X}_{obs}$ to produce conditional latent features $\boldsymbol{Z}_c$, subsequently generating prior latent features $\hat{\boldsymbol{Z}}_0$ via the noise matching and diffusion sampling in Eqs. (11) and (12).

C. *Spaformer*

The Spaformer module serves as the core component of our Diff-spaformer model to establish an end-to-end network mapping from missing data to complete data. As Fig. 1 shows, the concatenated input of observed seismic data $\boldsymbol{X}_{obs}$ and its corresponding missing mask 1-M is first processed by a 3×3 convolutional layer for channel dimension alignment. The encoding-decoding hierarchy utilizes U-Net's structural paradigm enhanced with transformer blocks, where each block integrates a sparse attention (SPA) mechanism for contextual feature modeling and a feedforward network (FFN) for feature transformation and enhancement. This hierarchical architecture balances signal fidelity and feature expression efficiency through progressive processing from global to local. Crucially, seismic prior knowledge is systematically integrated through cross-layer feature fusion mechanisms, ensuring consistent domain-specific constraint enforcement throughout the network. The final reconstruction phase employs a 3×3 convolutional layer to generate interpolated seismic data outputs $\hat{\boldsymbol{X}}$ while maintaining dimensional consistency with the input space. Specifically, suppose the input feature and output feature are $\boldsymbol{F}$ and $\hat{\boldsymbol{F}}$, respectively, we combine the SPA module with the FFN module using the following residual connection:

$$\hat{\boldsymbol{F}} = \mathrm{FFN}\left(\boldsymbol{F}', \hat{\boldsymbol{Z}}_0\right) + \boldsymbol{F}', \text{ where } \boldsymbol{F}' = \mathrm{SPA}\left(\boldsymbol{F}, \hat{\boldsymbol{Z}}_0\right) + \boldsymbol{F}. \tag{14}$$

The SPA module processes the normalized input feature $\boldsymbol{F}$ along with the prior latent feature $\hat{\boldsymbol{Z}}_0$, and adds its output to the original input $\boldsymbol{F}$ via a residual connection to obtain $\boldsymbol{F}'$. Then, the normalized feature $\boldsymbol{F}'$, combined with the prior latent feature $\hat{\boldsymbol{Z}}_0$, serves as the input to the FFN module. The FFN output is added back to its input through another residual connection to generate $\hat{\boldsymbol{F}}$. This dual residual structure preserves the original feature information and enables the network to learn the residual mapping, effectively mitigating the vanishing gradient problem.

During model optimization, the L1 norm serves as the primary term in the loss function. To implement differentiated error weighting between valid traces and missing traces, we develop the weighted composite loss framework as

$$\mathcal{L}_{\text{rec}} = \lambda_1\|\left(1-\boldsymbol{M}\right)\odot\left(\hat{\boldsymbol{X}}-\boldsymbol{X}\right)\|_1 + \lambda_2\|\boldsymbol{M}\odot\left(\hat{\boldsymbol{X}}-\boldsymbol{X}\right)\|_1, \tag{15}$$

where $\lambda_1, \lambda_2 \geq 0$ to regulate the proportional weights of different loss components. This proposed formulation achieves higher error sensitivity in missing regions and enhances the model's capacity to detect and respond to discrepancies within data gaps through dynamically adjusted penalty mechanisms, while maintaining balanced optimization for valid signal regions. The final loss formulation can be expressed as

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{diff}}, \tag{16}$$

where the coefficient $\lambda_3 \geq 0$. This ensures that the optimization process prioritizes specific error terms according to their assigned weights and balances the trade-off between different objectives.

D. SPA Module

The SPA module integrates the compact seismic prior information extracted from SPEN, achieving multi-scale feature modeling from global to local scales. This process incorporates several key improvements, including dynamic feature calibration mechanisms, sparse self-attention, and affinity matrix computation based on negative squared Euclidean distance. We will systematically elaborate on these modules.

1) Dynamic Feature Calibration Mechanism

As Fig. 2 shows, the dynamic feature calibration mechanism enables the adaptive fusion of external prior information $\hat{\boldsymbol{Z}}_0$ with the main feature flow. Let $\boldsymbol{F} \in \mathbb{R}^{c \times h \times w}$ be the input feature (c is the channel dimension, and h and w represent spatial dimensions). As Fig. 2 shows, we perform feature alignment on the seismic prior $\hat{\boldsymbol{Z}}_0$ through the linear transformation as

$$\hat{\boldsymbol{Z}} = \boldsymbol{W}_z \cdot \hat{\boldsymbol{Z}}_0, \tag{17}$$

where $\boldsymbol{W}_z \in \mathbb{R}^{c \times C'}$ is the weight matrix. After reshaping $\hat{\boldsymbol{Z}}$, the channel attention mechanism is applied to generate gating weights:

$$\boldsymbol{G} = \sigma\left(\text{Conv}_2\left(\text{ReLU}\left(\text{Conv}_1\left(\text{AvgPool}\left(\boldsymbol{F} + \hat{\boldsymbol{Z}}\right)\right)\right)\right)\right), \tag{18}$$

where σ denotes a Sigmoid activation function. Then, we adopt the following gated residual connections to achieve feature fusion:

$$\hat{\boldsymbol{F}}_G = \boldsymbol{F} + \boldsymbol{G} \odot \hat{\boldsymbol{Z}}, \tag{19}$$

where $\odot$ denotes the Hadamard product, and this operation allows the network to dynamically adjust the contribution of the seismic prior. Compared to traditional methods such as direct addition or channel concatenation, the dynamic feature calibration mechanism can more effectively realize adaptive fusion of cross-modal features.

2) Sparse Self-attention

We divide the input feature $\hat{\boldsymbol{F}}_G$ into different subspaces (n_{head} in total) through multiple sets of independent linear projections as

$$\boldsymbol{Q}_i = \hat{\boldsymbol{F}}_G \boldsymbol{W}_i^Q, \boldsymbol{K}_i = \hat{\boldsymbol{F}}_G \boldsymbol{W}_i^K, \boldsymbol{V}_i = \hat{\boldsymbol{F}}_G \boldsymbol{W}_i^V, i = 1, \ldots, n_{\text{head}} \tag{20}$$

where the trainable parameter matrices $\boldsymbol{W}_i^Q, \boldsymbol{W}_i^K, \boldsymbol{W}_i^V \in \mathbb{R}^{c \times c_n}$ $(c_n = c/n_{\text{head}})$. Then, the multi-head self-attention [50] can be constructed by

$$\begin{gathered}\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_{n_{\text{head}}}\right) \boldsymbol{W}^O, \\ \text{where } \text{head}_i = \text{Attention}\left(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\right),\end{gathered} \tag{21}$$

where the output parameter matrix $\boldsymbol{W}^O \in \mathbb{R}^{c \times c}$ effectively integrates information from multiple heads to avoid redundancy. Each head independently performs the attention calculation, allowing the model to simultaneously focus on different types of dependencies within the feature. It parallelizes multi-scale feature learning to enhance the model's ability to capture complex dependencies while maintaining computational efficiency.

Given $\boldsymbol{Q}_i, \boldsymbol{K}_i \in \mathbb{R}^{c_n \times N}$ $(N = w \times h,$ usually $c_n \ll N)$ within the same head, the pairwise affinity matrix between them is $\boldsymbol{S}_i = s\left(\boldsymbol{Q}_i^T, \boldsymbol{K}_i\right)$, typically $\boldsymbol{S}_i \in \mathbb{R}^{N \times N}$. This operation gives rise to $\mathcal{O}\left(N^2 c_n\right)$ computations and $\mathcal{O}\left(N^2\right)$ memory costs. Unlike general global attention mechanisms that solely operate on compressed features (e.g., ANet [32]), computational expenses become unacceptable when applied to the high-resolution features in the shallow layers of our model. To solve this issue, we reformulate attention mechanisms across channel dimensions. Specifically, global dependencies are captured through channel-wise dot-product operations between query Q_i and key K_i projections as

$$\boldsymbol{S}_i = s\left(\boldsymbol{Q}_i, \boldsymbol{K}_i^T\right) \text{ with } \boldsymbol{S}_i \in \mathbb{R}^{c_n \times c_n}. \tag{22}$$

The computational complexity and memory cost have been reduced to $\mathcal{O}\left(N c_n^2\right)$ and $\mathcal{O}\left(c_n^2\right)$, respectively, effectively improving the computational efficiency.

In conventional attention mechanisms, similarity scores are normalized using the softmax function, creating a dense weight matrix. While this enables dynamic attention allocation, it can lead to scattered weights, causing the model to attend to noise or weakly related elements, which impedes key information extraction. To address this issue, we adopt a sparse attention mechanism proposed in [51] as

$$\boldsymbol{W}_i = \frac{1}{\omega} \text{ReLU}\left(\boldsymbol{S}_i\right). \tag{23}$$

where ω is a learnable scaling parameter. The Rectified Linear Unit (ReLU) activation function is applied to conduct sparse similarity score calculation, and its nonlinear characteristics filter out negative values (low correlation or noise signals), retaining only positive values and directly suppressing weak correlations. This approach eliminates the reliance on softmax normalization and instead achieves sparsity through activation thresholding. The generated sparse attention map focuses on strongly related elements, reducing the interference of irrelevant information and lowering computational complexity.

Finally, the sparse attention map achieves feature-weighted aggregation by performing matrix multiplication with the Value matrix $\boldsymbol{V}_i$ per head. Specifically, the mathematical formulation for this process is:

$$\text{head}_i = \boldsymbol{W}_i \boldsymbol{V}_i. \tag{24}$$

We concatenate all heads and restore it to the original dimension to generate the multi-head feature $\text{Concat}\left(\text{head}_1, \ldots, \text{head}_{n_{\text{head}}}\right)$ (i.e., $\hat{\boldsymbol{F}}$ in Fig. 2).

3) Negative Squared Euclidean Distance

The similarity definition in Eq. (22) is the core technological approach for constructing feature projections. Its essence lies in quantifying the strength of the association

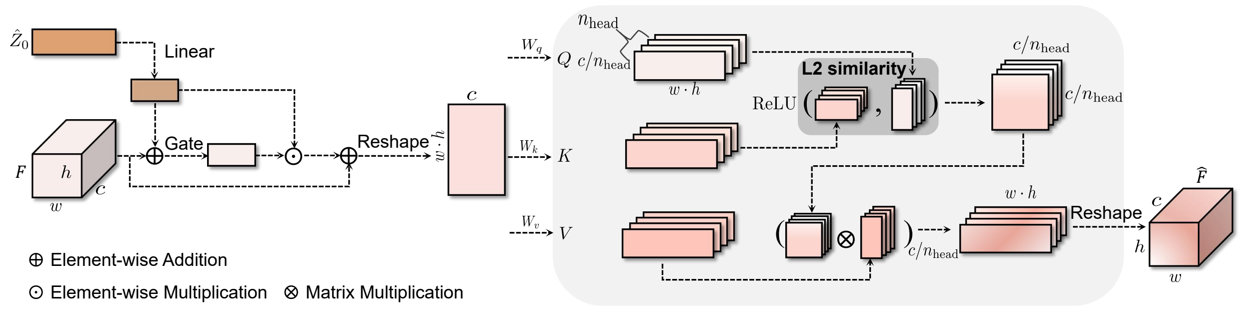


Fig. 2. SPA module

between feature vectors. Cosine similarity is a widely used metric that quantifies the similarity between vectors based on their directional alignment in the vector space. Its core principle involves evaluating the degree of similarity between two non-zero vectors by calculating the cosine of the angle between them. However, studies have shown that for tasks requiring strong global pixel attention, the optimization should focus on balancing both directional similarity and magnitude information rather than neglecting one in favor of the other [52]. In seismic data interpolation, accurate recovery of high-frequency details depends on magnitude information to represent structural similarity. The variation in high-frequency signal magnitude reflects strata heterogeneity and acoustic impedance differences. Thus, constructing the seismic similarity measure requires considering both the absolute magnitude and phase characteristics of the waveform.

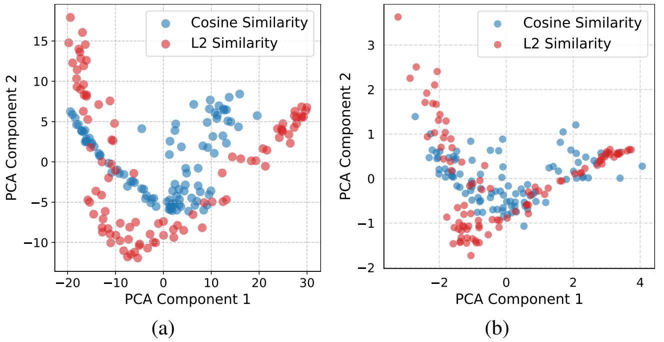
Let $\boldsymbol{S}_{i,jk} = s\left(\boldsymbol{Q}_{i,j}, \boldsymbol{K}_{i,k}^T\right)$ represent the similarity between the query feature vector $\boldsymbol{Q}_{i,j}$ (at the j-th position) and the key feature vector $\boldsymbol{K}_{i,k}^T$ (at the k-th position). The negative squared Euclidean distance for similarity functions is defined as

$$s\left(\boldsymbol{Q}_{i,j}, \boldsymbol{K}_{i,k}^T\right) = -\left\|\boldsymbol{Q}_{i,j} - \boldsymbol{K}_{i,k}^T\right\|_2^2, \qquad (25)$$

whose computational complexity is only slightly higher than that of cosine similarity. For simplicity, we refer to it as L2 similarity. Its core is to measure the Euclidean distance between vectors, converting the distance into a similarity score through a negative sign (the smaller the distance, the higher the score). Unlike the cosine similarity, L2 similarity is not dominated by the vector magnitudes, thus avoiding the interference of seismic amplitude differences on attention weights. Fig. 3(a) and Fig. 3(b) visualize the projections of the affinity matrix $\mathrm{Concat}\left(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{n_{\text{head}}}\right)$ and the output features $\mathrm{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ from the multi-head attention module after dimensionality reduction using Principal Component Analysis (PCA). It can be observed that, compared to cosine similarity, L2 similarity exhibits a more uniform distribution in the 2D projection, avoiding the clustering concentration caused by the direction sensitivity of cosine similarity. This characteristic ensures more balanced contribution weights across seismic data points, thereby maintaining the balance between clusters.

E. FFN

The SPA module achieves collaborative optimization of local dependencies and global multi-scale features in feature extraction through the structured sparse design and enhanced information extraction. The FFN performs nonlinear enhancement and position-independent processing on the features output by the self-attention mechanism, further enhancing the representational capacity of the features. As Fig. 4 shows, we adopt the same dynamic feature calibration mechanism proposed in Section III-D1 to achieve the adaptive fusion of external prior information $\hat{\boldsymbol{Z}}_0$, also denoting the fused feature as $\hat{\boldsymbol{F}}_G$. We then use the basic structure of the FFN from the transformer architecture [49], which consists of two linear transformations and a nonlinear activation function, and the formula is defined as



(a) (b)

Fig. 3. (a) PCA projection of the affinity matrix $\mathrm{Concat}\left(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{n_{\text{head}}}\right)$ under different similarity functions. (b) PCA projection of $\mathrm{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ under different similarity functions.

$$\hat{\boldsymbol{F}} = \boldsymbol{W}_{f2}\hat{\boldsymbol{F}}_G \odot \mathrm{GELU}\left(\boldsymbol{W}_{f1}\hat{\boldsymbol{F}}_G\right), \qquad (26)$$

where $\boldsymbol{W}_{f1}$ and $\boldsymbol{W}_{f2}$ represent the two 3×3 convolution operations, and $\hat{\boldsymbol{F}}_G$ undergoes 1×1 convolutions at both ends to expand and restore the channel dimensions. It strengthens local detail feature extraction by combining dynamic gating and convolution operations while maintaining the global modeling capability of the Transformer.

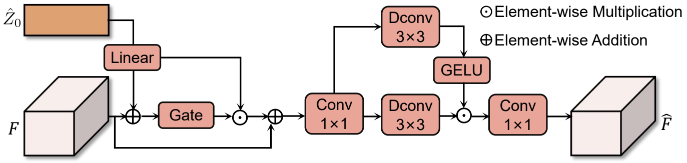


Fig. 4. FFN module

This Diff-spaformer framework optimizes the interplay between conditional feature encoding, diffusion-based prior generation, and multi-scale feature integration, ensuring efficient and accurate reconstruction. Algorithm 1 and Algorithm 2 outline the procedures for the training and inference phases of our model, respectively.

Algorithm 1 Training process of Diff-spaformer

```
Input: Training ground truth seismic data $\{\boldsymbol{X}^i\}_{i=1}^n$ with total number $n$; Initializing the SPEN, Spaformer, and the noise-matching network;
Diffusion steps $T$; batch size $K$; the number of epochs $N$.
Randomly initialize Diff-spaformer;
for $i = 1, \dots, N$ do
    Sample batch data $\{\boldsymbol{X}^i\}_{i=1}^K$ from training data;
    Generate the binary missing masks $\{\boldsymbol{M}^i\}_{i=1}^K$;
    Construct missing data $\{\boldsymbol{X}_{obs}^i\}_{i=1}^K$ according to Eq. 6;
    Get the compressed prior feature $\boldsymbol{Z}_0$ form Eq. 8 and the conditional prior information $\boldsymbol{Z}_c$ form Eq. 9;
    Create the $T$-steps forward process based on Eq. 10;
    for $t = T, \dots, 1$ do
        Get $\{\hat{\boldsymbol{Z}}_0^{i,t}\}_{i=1}^K$ from the noise matching process according to Eq. 11;
        Get $\{\hat{\boldsymbol{Z}}_{t-1}^i\}_{i=1}^K$ from reverse process according to Eq. 12;
    end for
    Get predictions $\{\hat{\boldsymbol{X}}^i\}_{i=1}^K$ from the Spaformer network;
    Update the Diff-spaformer network with $\mathcal{L}$ in Eq. 16;
end for
```

Algorithm 2 Inference process of Diff-spaformer

```
Input: Missing seismic data $\boldsymbol{X}_{obs}$; Corresponding binary missing mask $\boldsymbol{M}$; Trained Diff-spaformer model.
Get the conditional prior information $\boldsymbol{Z}_c$ form Eq. 9;
Sample $\boldsymbol{Z}_T$ from the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$;
$\hat{\boldsymbol{Z}}_T = \boldsymbol{Z}_T$
for $t = T, \dots, 1$ do
    Generate $\hat{\boldsymbol{Z}}_0^t$ from the noise matching process based on Eq. 11;
    Get $\hat{\boldsymbol{Z}}_{t-1}$ from reverse process according to Eq. 12;
end for
Get predictions $\hat{\boldsymbol{X}}$ from the Spaformer network;
Output: Interpolated data $\widetilde{\boldsymbol{X}} = \boldsymbol{M} \odot \boldsymbol{X}_{obs} + (1 - \boldsymbol{M}) \odot \hat{\boldsymbol{X}}$.
```

IV. Experiments

A. Evaluation Metrics

The fidelity evaluation of interpolated seismic data is quantitatively assessed through three key metrics, including Mean Squared Error (MSE), Signal-to-Noise Ratio (SNR), and Peak Signal-to-Noise Ratio (PSNR). For the interpolated seismic data $\{\widetilde{\boldsymbol{X}}^i\}_{i=1}^n$ and the ground truth $\{\boldsymbol{X}^i\}_{i=1}^n$, MSE quantifies the average squared deviation as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\widetilde{\boldsymbol{X}}^i - \boldsymbol{X}^i)^2. \tag{27}$$

This metric measures the dispersion of reconstruction errors, with values approaching zero indicating superior reconstruction accuracy and interpolation fidelity. The squared term amplifies significant deviations while preserving dimensional alignment with source data. SNR evaluates the signal preservation capability in decibels (dB) through power ratio analysis as

$$\text{SNR} = 10 \log_{10} \frac{\|\boldsymbol{X}\|_F^2}{\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\|_F^2}, \tag{28}$$

where $\|\cdot\|_F$ refers to the Frobenius norm. PSNR extends SNR by incorporating the dynamic range of seismic signals and is defined as

$$\text{PSNR} = 10 \log_{10} \frac{\text{MAX}_{\boldsymbol{X}}^2}{\text{MSE}}, \tag{29}$$

where $\text{MAX}_{\boldsymbol{X}}$ represents the maximum amplitude of $\boldsymbol{x}_{\text{gt}}$. Higher SNR and PSNR values correspond to superior interpolation quality. We use the Structural Similarity Index (SSIM) [53] to evaluate the structural similarity between two seismic datasets by comparing their local statistical features. The calculation formula of SSIM is as follows

$$\text{SSIM}(\boldsymbol{X}, \widetilde{\boldsymbol{X}}) = \frac{\left(2\mu_{\boldsymbol{X}}\mu_{\widetilde{\boldsymbol{X}}} + C_1\right)\left(2\sigma_{\boldsymbol{X}\widetilde{\boldsymbol{X}}} + C_2\right)}{\left(\mu_{\boldsymbol{X}}^2 + \mu_{\widetilde{\boldsymbol{X}}}^2 + C_1\right)\left(\sigma_{\boldsymbol{X}}^2 + \sigma_{\widetilde{\boldsymbol{X}}}^2 + C_2\right)},$$

where $\mu_{\boldsymbol{X}}(\mu_{\widetilde{\boldsymbol{X}}})$, $\sigma_{\boldsymbol{X}}(\sigma_{\widetilde{\boldsymbol{X}}})$, and $2\sigma_{\boldsymbol{X}\widetilde{\boldsymbol{X}}}$ denote the mean (luminance), variance (contrast), and covariance (structure), respectively. C_1 and C_2 are constants introduced to avoid division by zero, typically set to very small values (e.g., $C_1 = 1e-4$, $C_2 = 1e-4$). In practical applications, SSIM is computed by sliding a window (e.g., a 3×3 pixel window) across the seismic data, calculating the value for each block, and then averaging the results over the entire data to capture local spatial characteristics. The resulting similarity value ranges from -1 to 1, with values closer to 1 indicating greater structural similarity. This multi-metric framework enables comprehensive characterization of reconstruction accuracy, noise suppression effectiveness, amplitude preservation fidelity, and structural similarity in seismic data interpolation tasks.

B. Data Set

We validate the proposed model on three publicly available datasets, including the synthetic dataset Model94 and Society of Exploration Geophysicists (SEG) C3, and the field dataset Mobil Avo Viking Graben Line 12 (MAVO), all of which are commonly used for seismic data reconstruction tasks.

The Model94 dataset contains 277 shot-gathers, of which 198 consecutive and complete gathers (each comprising 480 traces with 15 m intervals) are selected following the operation in [54]. Each trace comprises 2,000 time samples at a 4 ms sampling rate. To mitigate partial temporal vacancies, we crop the temporal dimension to 1,000 samples. The data are randomly divided into 118, 40, and 40 shot-gathers for training, validation, and testing, respectively. Due to limited data volume, the dataset is augmented by duplication, yielding 30,000 training, 6,000 validation, and 6,000 test patches through randomized cropping on shot-gathers.

The SEG C3 dataset includes 45 shots, each with a 201 × 201 receiver grid (dx, dy = 20 m) and 625 samples per

TABLE I
COMPARISON OF VARIOUS METHODS ON THE TEST SET OF DIFFERENT DATASETS WITH RANDOM MISSING TRACES. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Dataset	SEG C3				MAVO				Model94			
Model	MSE↓	SNR↑	PSNR↑	SSIM↑	MSE↓	SNR↑	PSNR↑	SSIM↑	MSE↓	SNR↑	PSNR↑	SSIM↑
DD-CGAN[30]	2.284e-04	30.458	36.412	0.952	2.639e-04	30.254	35.786	0.955	2.421e-04	29.588	36.160	0.978
cWGAN-GP[25]	9.708e-05	34.175	40.129	0.981	1.971e-04	31.522	37.054	0.967	1.390e-04	31.997	38.569	0.987
PConv-UNet[22]	6.810e-05	35.715	41.669	0.986	1.221e-04	33.601	39.133	0.977	1.870e-04	30.711	37.283	0.984
ANet[32]	1.204e-04	33.240	39.194	0.977	1.889e-04	31.705	37.237	0.969	1.562e-04	31.493	38.065	0.986
Coarse-to-Fine[31]	7.260e-05	35.437	41.391	0.985	1.595e-04	32.440	37.972	0.971	1.187e-04	32.684	39.256	0.988
SeisFusion*[46]	9.603e-05	34.222	40.176	0.980	1.841e-04	31.817	37.349	0.965	1.650e-04	31.252	37.824	0.980
SeisDDIMCR [43]	4.777e-05	37.255	43.209	0.989	1.127e-04	33.950	39.482	0.976	8.043e-05	34.374	40.946	0.991
Ours	**3.763e-05**	**38.291**	**44.245**	**0.992**	**1.027e-04**	**34.353**	**39.885**	**0.978**	**4.440e-05**	**36.954**	**43.526**	**0.996**

trace (dt = 8 ms). Due to the limited valid seismic events in the bottom part of time samples, we select parts of the data for our experiment, resulting in a final data size of $45\times201\times201\times300$. Following the training, validation, and test set generation rules used in [43], for each seismic shot, we randomly selected 20 gathers for validation and 20 gathers for testing along the inline direction, with the remaining 161 gathers used for training. Then, on each slice, patches with both the time and trace dimensions of size 128 are randomly cropped. Ultimately, the total numbers of patches for validation, testing, and training are 30,000, 6,000, and 6,000, respectively.

The MAVO dataset consists of a 1001×120 receiver grid and 1500 time samples per trace, recorded with a time interval of 4 ms and a trace interval of 25 m. We exclude some of the later received data and construct $1001\times120\times1000$ data for experiments. Following [43], we randomly divide all the gathers into three parts, i.e., 801 for training, 100 for validation, and the remaining 100 for testing. Then, patches are randomly cropped from each part, resulting in 20,000 training patches, 4,000 validation patches, and 4,000 testing patches. Each patch keeps time and trace dimensions of 256 and 112, respectively.

Before being fed into models, all seismic patches are first normalized to the range $[0, 1]$ using min-max normalization.

C. Comparison Method

We compare our model with 7 methods, including two GAN-based approaches, i.e., the dual-domain optimized DD-CGAN [30] and conditional Wasserstein generative adversarial networks with gradient penalty (cWGAN-GP) [25], partial convolution-based U-Net (PConv-UNet) [22], CNN guided by attention mechanism (ANet) [32], two-stage Coarse-to-Fine model [31], and diffusion-based models SeisFusion [46] and SeisDDIMCR [43]. SeisDDIMCR and SeisFusion share the same backbone network. SeisFusion* (differentiated by the asterisk) implements single-network noise matching to eliminate dual-network computational costs. This critical divergence from the original dual-network architecture is explicitly marked in the notation. All other hyperparameter configurations follow their respective architectures.

D. Implementation Details

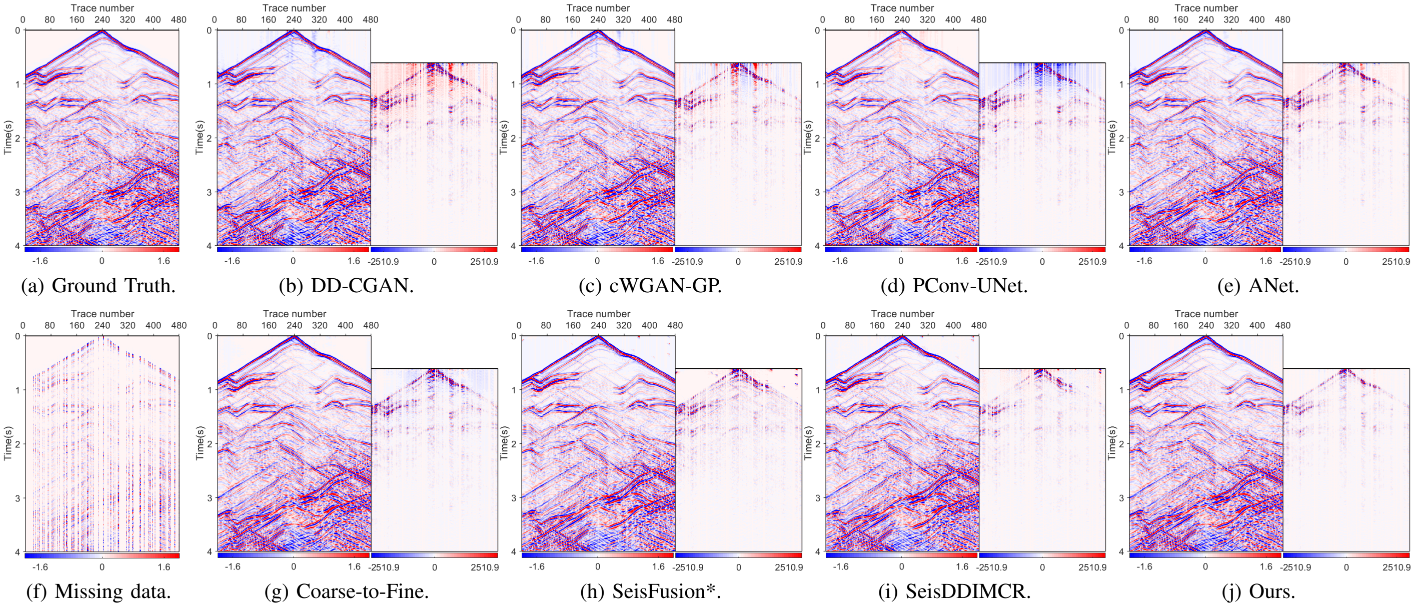
In the diffusion model configuration, the number of diffusion steps is set to 4, and the noise variance coefficient β increases linearly from 0.1 to 0.99. The Spaformer has a four-layer channel dimension, increasing from 64, 64×2, 64×4, to 64×8. The dimension C' of the compressed prior information is set to 64×4. The weight coefficients of the loss function are set to $\lambda_1 = 6.0$, $\lambda_2 = 1.0$, and $\lambda_3 = 6.0$. We execute Algorithm 1 using the Adam (Adaptive Moment Estimation) gradient descent optimization algorithm to train the model on different datasets. We train every model separately for two missing data scenarios, including random and continuous missing types. The random missing rate is set between 0.2 and 0.8, and the continuous missing rate ranges from 0.1 to 0.6. To preserve edge integrity, we prevent continuous missing cases from occurring near boundary traces. Masks are randomly generated during each training iteration to ensure sufficient diversity in training sample pairs. The training consists of 100 epochs with a batch size of 20. The learning rate follows a piecewise constant decay strategy, starting at $1e-4$, and decreasing by a factor of 10 after 50 epochs. The model parameters for the comparison methods are set to be consistent with those in the original paper, while the training strategy is aligned with that of our model. However, the total number of training epochs differs between the comparison methods and our approach, with the comparison methods undergoing more training epochs than our method. We conduct total experiments on PyTorch $1.12.1$ and NVIDIA A100 Tensor Core GPU.

E. Experimental Results

Random missing traces and continuous missing traces are two common manifestations of seismic data gaps. High-density random missing traces can lead to data sparsity, introducing aliasing artifacts and spectral leakage, and models need to implicitly learn anti-aliasing capabilities. Continuous missing traces rely more on global semantic reasoning, and it is crucial to reduce distribution discrepancies between missing and non-missing regions. We implement Algorithm 2 to validate the performance of our model. Similarly, the comparative methods are also tested on the same dataset.

1) Random Missing Traces

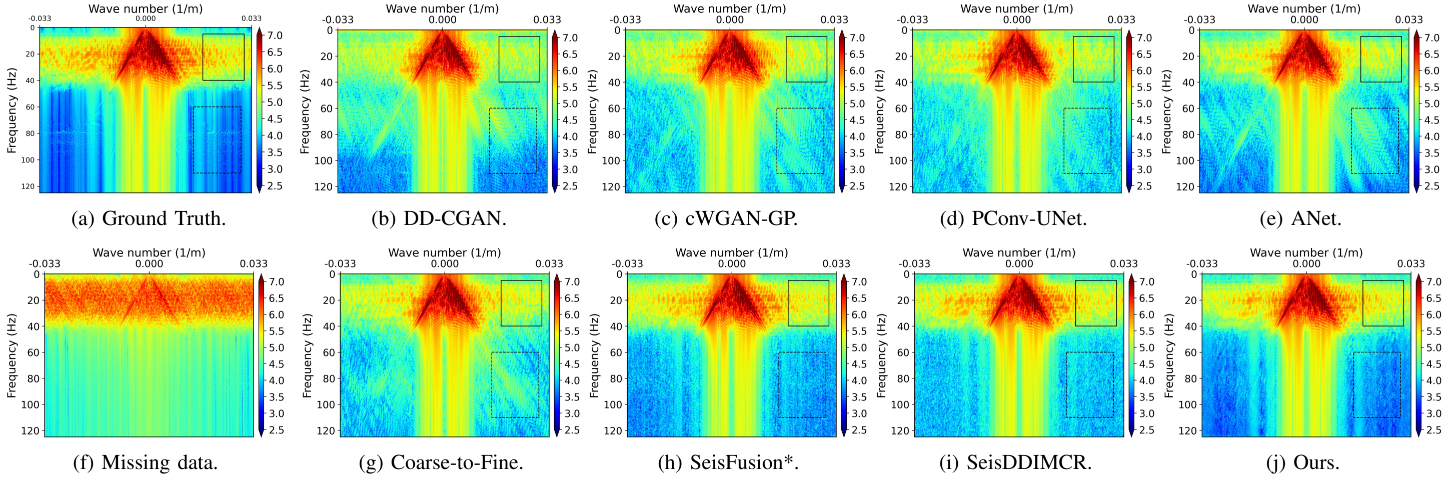
The test random missing rate is consistent with the training set, also ranging from 0.2 to 0.8. Tab. I summarizes the



(a) Ground Truth. (b) DD-CGAN. (c) cWGAN-GP. (d) PConv-UNet. (e) ANet.

(f) Missing data. (g) Coarse-to-Fine. (h) SeisFusion*. (i) SeisDDIMCR. (j) Ours.

Fig. 5. Interpolation results of the Model94 complete test slice with a 79.2% random missing ratio on different methods. We restore the seismic data to its original amplitude range and apply the gain method to enhance the visibility of weak amplitude details. The reconstruction residuals are displayed on the right panel for better comparison. Residuals are calculated and presented directly from raw data without any gain processing.



(a) Ground Truth. (b) DD-CGAN. (c) cWGAN-GP. (d) PConv-UNet. (e) ANet.

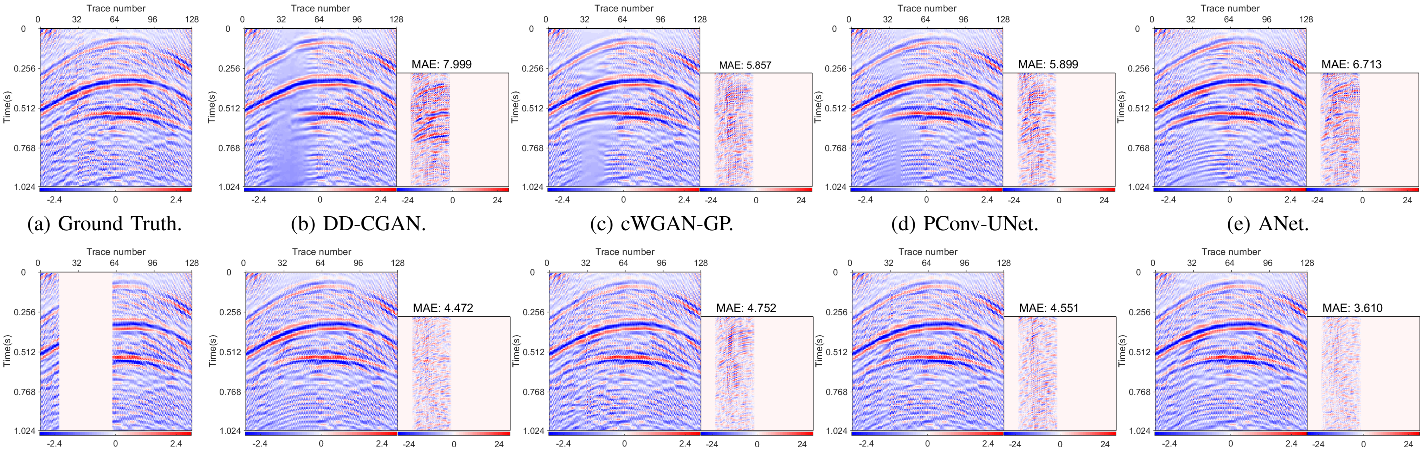
(f) Missing data. (g) Coarse-to-Fine. (h) SeisFusion*. (i) SeisDDIMCR. (j) Ours.

Fig. 6. The f-k spectra of Model94 test data interpolation results with 79.2% random missing traces on different methods.

quantitative interpolation evaluation results on the three test sets. Compared with comparative methods, our method consistently demonstrates superior performance across all evaluation metrics. It should be noted that, compared to SeisDDIMCR, our method achieves a more than 1 dB improvement in SNR on the SEG C3 and Model94 test sets, demonstrating its ability to bypass the highly iterative interpolation process while maintaining reconstruction fidelity. Fig. 5 visually compares the interpolation results and residuals of highly sparse random missing traces (missing rate: 79.2%) on the Model94 complete test slice. Reconstructing complete data from only 20.8% of the original observations requires joint modeling of global and local correlations while balancing anti-aliasing constraints and high-frequency reconstruction. First, except for Coarse-to-Fine and our proposed method, all other methods exhibit some degree of artifact residue in high-amplitude regions. Second, the residual plots clearly reflect the amplitude errors between the reconstructed signals and the ground truth signals at the same scale. Compared to the other six methods, SeisDDIMCR and our method show noticeably cleaner residual plots. Furthermore, our method exhibits less residual in the middle of the spread (e.g., from trace number 160 to 320). Besides, to quantitatively compare the effectiveness of different methods in suppressing aliasing artifacts induced by highly sparse data, we visualize the corresponding f-k spectra in Fig. 6. We clearly observe spectral leakage and coherent noise contamination in the f-k spectrum of the missing data, which manifest as linear dipping co-phase axes in the f-k domain. Except for SeisFusion*, SeisDDIMCR, and our method, all other methods suffer from severe spectral leakage and exhibit noticeable linear inclined artifacts. The central energy concentration region (main signal) of our method is similar to other approaches, but it exhibits significantly lower noise in the low-frequency range (indicated by the black

TABLE II
COMPARISON OF VARIOUS METHODS ON THE TEST SET OF DIFFERENT DATASETS WITH CONSECUTIVE MISSING TRACES. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Dataset	SEG C3				MAVO				Model94			
Model	MSE↓	SNR↑	PSNR↑	SSIM↑	MSE↓	SNR↑	PSNR↑	SSIM↑	MSE↓	SNR↑	PSNR↑	SSIM↑
DD-CGAN[30]	7.545e-04	25.269	31.223	0.891	3.624e-04	28.877	34.409	0.938	1.683e-03	21.167	27.739	0.917
cWGAN-GP[25]	3.180e-04	29.022	34.976	0.935	2.153e-04	31.138	36.670	0.960	7.705e-04	24.560	31.132	0.951
PConv-UNet[22]	4.023e-04	28.000	33.954	0.933	1.577e-04	32.490	38.022	0.972	9.343e-04	23.723	30.295	0.945
ANet[32]	4.428e-04	27.584	33.538	0.935	2.028e-04	31.397	36.929	0.965	9.207e-04	23.787	30.359	0.949
Coarse-to-Fine[31]	2.244e-04	30.535	36.489	0.960	1.429e-04	32.918	38.450	0.972	7.585e-04	24.629	31.201	0.951
SeisFusion*[46]	3.982e-04	28.045	33.999	0.953	1.611e-04	32.398	37.930	0.973	1.082e-03	23.085	29.657	0.946
SeisDDIMCR [43]	1.601e-04	32.002	37.956	0.973	9.741e-05	34.582	40.114	0.979	5.913e-04	25.710	32.282	0.966
Ours	**1.369e-04**	**32.682**	**38.636**	**0.973**	**9.234e-05**	**34.814**	**40.346**	**0.979**	**4.418e-04**	**26.976**	**33.548**	**0.968**



(a) Ground Truth. (b) DD-CGAN. (c) cWGAN-GP. (d) PConv-UNet. (e) ANet.

(f) Missing data. (g) Coarse-to-Fine. (h) SeisFusion*. (i) SeisDDIMCR. (j) Ours.

Fig. 7. Interpolation results of the SEG C3 patch with a 35% continuous missing ratio on different methods. We restore seismic data to its native amplitude range and apply the gain method to enhance subtle waveforms. Reconstruction residuals (right panel) are evaluated directly from raw data, excluding gain processing to preserve authentic signal differences. MAE is marked above the residual plot.

dashed box) and reduced energy leakage, benefiting from the precise reconstruction of randomly missing traces. Our method achieves a frequency distribution more consistent with the ground truth (e.g., within the solid black box), recovering relatively more mid- and high-frequency components.

2) Continuous Missing Traces

The continuous missing rate in the test scenario remains between 0.1 and 0.6. Tab. II presents the interpolation metrics on the three test sets with continuous missing gaps. Compared with other approaches, our method demonstrates superior performance in both fidelity preservation and textural detail retention. Fig. 7 compares the interpolation differences between different methods by showing amplitude recovery results and residual maps for a local SEG C3 test patch with 35% missing gaps. DD-CGAN, cWGAN-GP, PConv-UNet, and ANet fail significantly in signal recovery, with clear deviations in waveform continuity and amplitude fidelity. Blurred boundary information leads to unreasonably abrupt changes or artifacts. In contrast, methods like Coarse-to-Fine, SeisFusion*, and SeisDDIMCR show improvements, generating visually reasonable interpolations, yet still have issues such as local amplitude anomalies. Our method, with the smallest Mean Absolute Error (MAE), produces the most consistent recovery with the real signal. Fig. 8 shows the amplitude wiggle plots of different methods for a single SEG C3 seismic trace with continuous missing gaps. Our method achieves the best global waveform matching, with the highest Pearson correlation coefficient (r=0.945). In the zoomed-in region, the amplitude variation of our method closely matches the ground truth. More recovery results for continuous missing interpolation in the x-t domain and f-k domain are provided in the supplementary material.

The continuous missing rate significantly impacts seismic trace interpolation performance. As systematically compared in Fig. 9(a) using SNR curves of the SEG C3 test dataset, all methods exhibit SNR degradation with increasing missing rates, and demonstrate distinct attenuation gradients. While the two diffusion-based models and our method exhibit comparable performance at low missing rates (<20%), a marked performance divergence emerges under high missing rates (>30%), particularly highlighting our approach's enhanced robustness in severe data-absence scenarios. Notably, SeisFusion* suffers drastic SNR deterioration (△SNR=-16.1 dB from 10% to 60% missing rates), revealing that plug-and-play conditional resampling strategies relying solely on available data inevitably induce error accumulation issues, particularly exacerbated under severe missing conditions. Our method demonstrates consistently superior interpolation performance across all missing rates.

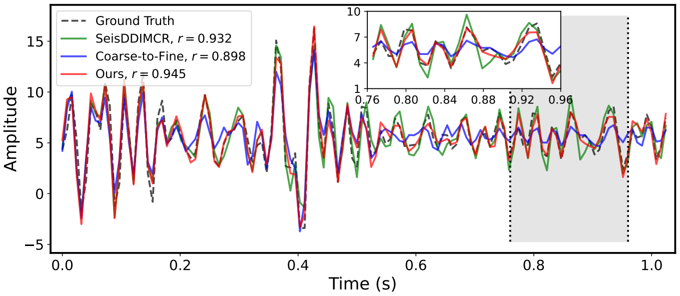
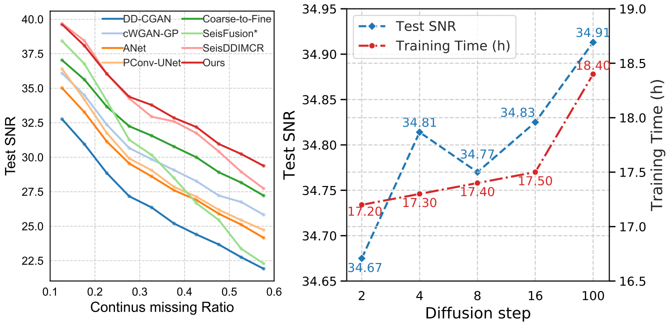


Fig. 8. The wiggle plot of continuous missing gap interpolation results for a single SEG C3 test trace using different methods. We select a representative missing trace (marked by the black dashed line) to evaluate interpolation performance while quantitatively calculating the Pearson correlation coefficient r between the reconstructed and true amplitudes.



(a) (b)

Fig. 9. (a) Curves of SNR versus the continuous missing ratio for different methods on the SEG C3 test dataset. (b) The comparison of test SNR and training time cost for our method under different diffusion steps on the MAVO dataset with continuous missing traces.

V. Ablation Study

This section conducts systematic ablation studies to validate the efficacy of individual components (Spaformer and SPA modules) in the proposed methodology. Further complexity and computational efficiency analyses substantiate the superiority of our approach. Besides, we provide the ablation study on the loss function in the supplementary materials.

A. Spaformer

The Spaformer constitutes the core of our framework as a U-Net-based end-to-end generator that integrates compressed seismic priors encoded through SPEN and generated via diffusion processes. Its architecture employs transformer blocks as fundamental units, incorporating multiscale feature construction through SPA modules and feature propagation or enhancement via FFN. Due to the structural complexity of SPA modules, their detailed analysis is reserved for Section V-B. This section systematically examines the ablation effects of SPEN, diffusion processes, FFN, and the adaptive gating operation of the dynamic feature calibration mechanism (commonly used in FFN and SPA). Tab. III quantifies the performance degradation observed on the MAVO test set with continuous missing traces when selectively disabling these components, revealing their distinct functional roles in maintaining waveform fidelity and texture generation capabilities. The first row of Tab. III presents the benchmark performance metrics of our complete model. To systematically assess the SPEN module's contributions, two ablation configurations are implemented. The first one is the non-shared SPEN architecture with independently initialized encoders for $\boldsymbol{X}_{obs}$ and $\boldsymbol{X}$ (Row 2 in Tab. III). The second configuration completely removes SPEN, thereby eliminating both prior learning and the diffusion process (as shown in Row 3 of Tab. III). Experimental results demonstrate hierarchical performance degradation. The non-shared configuration reduces test SNR by 0.16 dB due to impaired cross-path knowledge transfer, while complete SPEN removal causes more severe degradation of test SNR ($\triangle$SNR $= -0.33$ dB). This quantitative analysis reveals that shared SPEN parameters enable inter-path feature correlation, while the integrated prior learning contributes greater fidelity enhancement than standalone U-Net architecture, establishing SPEN's pivotal role in seismic signal reconstruction. The ablation study with diffusion process elimination (Row 4 in Tab. III) causes performance degradation (0.32 dB SNR drop). Fig. 9(b) illustrates the variation in model performance and training time with different diffusion steps on the MAVO dataset. This phenomenon demonstrates an inherent accuracy-efficiency trade-off, i.e., increasing diffusion steps from 2 to 100 steps progressively enhances test SNR (34.67→34.91 dB) yet linearly extends training duration (17.20→18.40 hours), necessitating systematic balancing in diffusion step configuration. The ablation study results in Row 5 of Tab. III highlight the critical role of FFN, as its removal significantly weakens feature enhancement, evidenced by a 2.4 dB SNR drop. The experimental results in Row 6 of Tab. III show that the removal of the adaptive gating mechanism causes a subtle SNR reduction, demonstrating its fine-grained modulation in dynamic weight allocation.

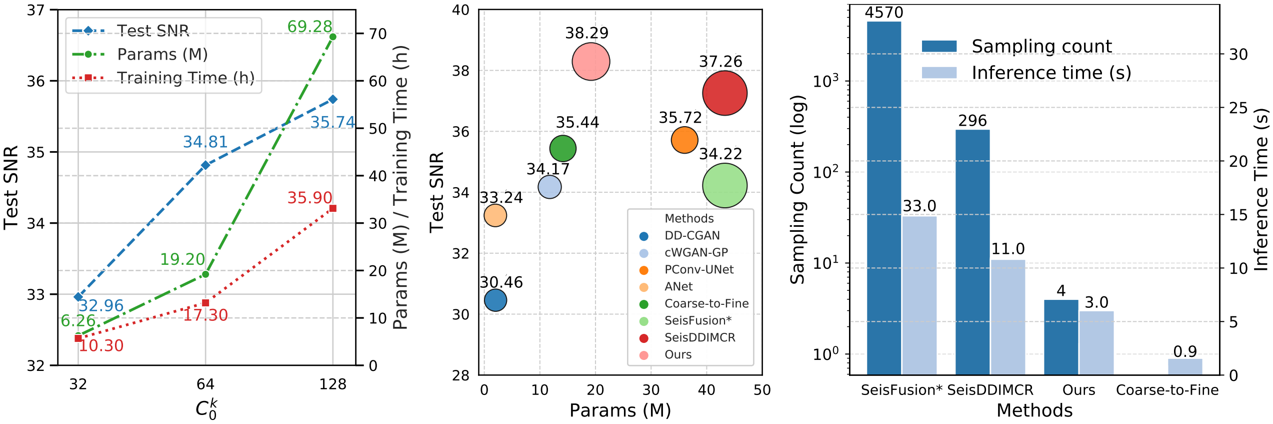
TABLE III

Ablation of different configurations of Spaformer.

SPEN	Diffusion Process	FFN	Gate	MSE	SNR	PSNR	SSIM
✓	✓	✓	✓	**9.234e-05**	**34.814**	**40.346**	**0.979**
Non-shared				9.532e-05	34.659	40.208	0.979
✗	✗			9.952e-05	34.489	40.021	0.978
	✗			9.944e-05	34.492	40.024	0.978
		✗		1.606e-04	32.410	37.942	0.969
			✗	9.434e-05	34.721	40.253	0.979

B. SPA

SPA serves as the core module for establishing similarity metrics and building global correlations. To validate the effectiveness of our proposed sparse attention calculation method based on L2 similarity, we compare the model's performance and calculation efficiency under different similarity measurement methods and key space dimensions (C_0^k, defined by the first-layer channel dimension of the U-Net). The channel dimensions of the four layers are C_0^k, $2C_0^k$, $4C_0^k$, and $8C_0^k$, respectively. The results are summarized in Tab. IV. Increasing C_0^k significantly reduces reconstruction error and improves quality, as higher dimensions capture more detailed structural information. However, as Fig. 10(a) shows, the performance gain diminishes with increasing dimensionality, with the improvement from 64 to 128 dimensions smaller than from 32 to 64. Both the total model parameters (Params (M)) and training cost exhibit superlinear growth with the



(a) (b) (c)

Fig. 10. (a) The comparison of the test SNR, number of parameters (Params), and training time cost for our method under different selections of C_0^k on the MAVO dataset with continuous missing traces. (b) The comparison of Params and test SNR for different methods on the SEG C3 dataset with random missing traces. The bubble size scales proportionally to the FLOPs, reflecting the computational cost of a model. (c) The comparison of the inference time (evaluated on a single-sample basis) and sampling count for different methods on the MAVO dataset with continuous missing traces. The sampling count axis is shown on a log scale.

increase of C_0^k. Additionally, it can also be observed that FLOPs grow nonlinearly with dimension, while the size of keys increases linearly in Tab. IV. Therefore, we adopt the moderate dimensionality of 64 to balance the performance and computational efficiency. Ablation experiments show that with equal feature dimensionality, L2 similarity consistently outperforms cosine similarity, and occupies similar computational (FLOPs) and memory costs (size of keys). This may be due to L2's sensitivity to feature magnitude, which aligns with the physical meaning of amplitude errors in seismic data, enabling more accurate matching without extra cost.

TABLE IV
ABLATION OF DIFFERENT CONFIGURATIONS OF SPA.

Similarity function	C_0^k	MSE	SNR	PSNR	SSIM	FLOPs(G)	Size of keys(MB)
L2 similarity	64	9.234e-05	34.814	40.346	0.979	6.26	3.41
Cosine similarity	64	9.794e-05	34.559	40.091	0.979	6.23	3.41
L2 similarity	32	1.414e-04	32.962	38.494	0.972	1.57	0.85
Cosine similarity	32	1.486e-04	32.746	38.278	0.971	1.56	0.85
L2 similarity	128	7.462e-05	35.740	41.272	0.983	24.97	13.63
Cosine similarity	128	7.569e-05	35.677	41.209	0.983	24.90	13.63

* The number of floating-point operations (FLOPs) is calculated solely for the similarity operation.

C. Computational Complexity

We evaluate our model from three aspects, i.e., parameter size, computational cost, and inference cost. As shown in Fig. 10(b), the joint visualization of the computational complexity and interpolation performance of random missing traces on the SEG C3 dataset reveals differences in the accuracy-complexity trade-offs on various models. Model sizes range from 2M (ANet and DD-CGAN) to 43M (SeisDDIMCR). Our method achieves the highest SNR with a moderate model size of 19M, outperforming the second-best SeisDDIMCR by 1.03 dB, thereby demonstrating superior efficiency and performance. Fig. 10(c) compares the inference efficiency of three diffusion architectures (SeisFusion*, SeisDDIMCR, and Ours) with the end-to-end baseline (Coarse-to-Fine) on the MAVO dataset. SeisFusion* and SeisDDIMCR suffer from heavy resampling overhead, limiting real-time processing capability. By leveraging transformer-based feature extraction and SPEN-guided prior projection, our method reduces the sampling process to just four steps, achieving an inference speed comparable to that of the end-to-end baseline.

VI. CONCLUSION

In this paper, we propose Diff-spaformer, a novel framework that integrates transformer architecture with diffusion processes to address the challenge of seismic data interpolation. SPEN effectively bridges the global modeling capability of sparse multi-head attention with the distribution consistency constraints of diffusion models. We demonstrate the superior performance of L2 similarity over traditional cosine similarity in seismic amplitude modeling. Sparse self-attention and adaptive ReLU filtering significantly reduce computational complexity while maintaining feature interaction efficiency. The framework employs a single-stage optimization and deterministic lightweight sampling strategy, improving both efficiency and interpolation quality. Experimental results show strong performance in handling random and continuous missing data, avoiding the computational burden of iterative resampling in plug-and-play diffusion interpolation models, with promising potential for field data applications. Future work will focus on 3D seismic data interpolation and domain generalization strategies to enhance robustness across exploration scenarios.

ACKNOWLEDGMENT

The authors thank Sandia National Laboratory, Mobil Oil Company, and Dr. Joe Dellinger at BP for their provision of open-access seismic data resources that facilitate the development and validation of the proposed methodology throughout this investigation.

References

[1] D. Trad, "Five-dimensional interpolation: Recovering from acquisition constraints," *Geophysics*, vol. 74, no. 6, pp. V123–V132, 2009.

[2] S. Spitz, "Seismic trace interpolation in the fx domain," *Geophysics*, vol. 56, no. 6, pp. 785–794, 1991.

[3] Y. Wang, "Seismic trace interpolation in the f-x-y domain," *Geophysics*, vol. 67, no. 4, pp. 1232–1239, 2002.

[4] Y. Chen, S. Fomel, H. Wang, and S. Zu, "5d dealiased seismic data interpolation using nonstationary prediction-error filter," *Geophysics*, vol. 86, no. 5, pp. V419–V429, 2021.

[5] J. Ronen, "Wave-equation trace interpolation," *Geophysics*, vol. 52, no. 7, pp. 973–984, 1987.

[6] S. Fomel, "Applications of plane-wave destruction filters," *Geophysics*, vol. 67, no. 6, pp. 1946–1960, 2002.

[7] M. D. Sacchi, T. J. Ulrych, and C. J. Walker, "Interpolation and extrapolation using a high-resolution discrete fourier transform," *IEEE Transactions on Signal Processing*, vol. 46, no. 1, pp. 31–38, 1998.

[8] P. Zwartjes and M. Sacchi, "Fourier reconstruction of nonuniformly sampled, aliased seismic data," *Geophysics*, vol. 72, no. 1, pp. V21–V32, 2007.

[9] P. Zwartjes and A. Gisolf, "Fourier reconstruction with sparse inversion," *Geophysical Prospecting*, vol. 55, no. 2, pp. 199–221, 2007.

[10] M. Naghizadeh and M. D. Sacchi, "Beyond alias hierarchical scale curvelet interpolation of regularly and irregularly sampled seismic data," *Geophysics*, vol. 75, no. 6, pp. WB189–WB202, 2010.

[11] R. Shahidi, G. Tang, J. Ma, and F. J. Herrmann, "Application of randomized sampling schemes to curvelet-based sparsity-promoting seismic data recovery," *Geophysical Prospecting*, vol. 61, no. 5, pp. 973–997, 2013.

[12] J. Wang, M. Ng, and M. Perz, "Seismic data interpolation by greedy local radon transform," *Geophysics*, vol. 75, no. 6, pp. WB225–WB234, 2010.

[13] S. Trickett, L. Burroughs, A. Milton, L. Walton, and R. Dack, "Rank-reduction-based trace interpolation," in *SEG International Exposition and Annual Meeting*, no. SEG-2010-3829, October 2010.

[14] J. Ma, "Three-dimensional irregular seismic data reconstruction via low-rank matrix completion," *Geophysics*, vol. 78, no. 5, pp. V181–V192, 2013.

[15] X. Wu, L. Liang, Y. Shi, and S. Fomel, "Faultseg3d: Using synthetic data sets to train an end-to-end convolutional neural network for 3d seismic fault segmentation," *GEOPHYSICS*, vol. 84, no. 3, pp. IM35–IM45, 2019.

[16] X. Wang, Y. Sui, and J. Ma, "Quadratic unet for seismic random noise attenuation," *Geophysics*, vol. 90, no. 2, pp. V43–V55, 2025.

[17] H. Wang, J. Zhang, X. Wei, L. Long, C. Zhang, and Z. Guo, "Upnet: Uncertainty-based picking deep learning network for robust first break picking," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, no. 5924214, 2024.

[18] S. Mandelli, F. Borra, V. Lipari, P. Bestagini, A. Sarti, and S. Tubaro, "Seismic data interpolation through convolutional autoencoder," in *SEG International Exposition and Annual Meeting*, October 2018, no. SEG-2018-2995428.

[19] F. Meng, Q. Fan, and Y. Li, "Self-supervised learning for seismic data reconstruction and denoising," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, no. 7502805, 2021.

[20] S. Tang, Y. Ding, H.-W. Zhou, and H. Zhou, "Reconstruction of sparsely sampled seismic data via residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.

[21] W. Fang, L. Fu, M. Zhang, and Z. Li, "Seismic data interpolation based on u-net with texture loss," *Geophysics*, vol. 86, no. 1, pp. V41–V54, 2021.

[22] S. Pan, K. Chen, J. Chen, Z. Qin, Q. Cui, and J. Li, "A partial convolution-based deep-learning network for seismic data regularization," *Computers & Geosciences*, vol. 145, no. 104609, 2020.

[23] N. Liu, L. Wu, J. Wang, H. Wu, J. Gao, and D. Wang, "Seismic data reconstruction via wavelet-based residual deep learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, no. 4508213, 2022.

[24] H. Kaur, N. Pham, and S. Fomel, "Seismic data interpolation using CycleGAN," in *SEG Technical Program Expanded Abstracts*, 2019, pp. 2202–2206.

[25] Q. Wei and X. Li, "Big gaps seismic data interpolation using conditional Wasserstein generative adversarial networks with gradient penalty," *Exploration Geophysics*, vol. 53, no. 5, pp. 477–486, 2022.

[26] M. M. Abedi, D. Pardo, and T. Alkhalifah, "Ensemble deep learning for enhanced seismic data reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, no. 5916311, 2024.

[27] T. He, B. Wu, and X. Zhu, "Seismic data consecutively missing trace interpolation based on multistage neural network training process," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, no. 7504105, 2022.

[28] B.-F. Wang, S.-C. Lin, and X.-Y. Chen, "Self-supervised simultaneous deblending and interpolation of incomplete blended data using a multistep blind-trace u-net," *Petroleum Science*, vol. 22, no. 3, pp. 1098–1109, 2025.

[29] A. Song, C. Wang, C. Zhang, J. Zhang, and D. Xiong, "Seismic data reconstruction via recurrent residual multiscale inference," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, no. 8029105, 2022.

[30] D. Chang, W. Yang, X. Yong, G. Zhang, W. Wang, H. Li, and Y. Wang, "Seismic data interpolation using dual-domain conditional generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 10, pp. 1856–1860, 2020.

[31] X. Wei, C. Zhang, H. Wang, Z. Zhao, D. Xiong, S. Xu, J. Zhang, and S.-W. Kim, "Hybrid loss-guided coarse-to-fine model for seismic data consecutively missing trace reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, no. 5923315, 2022.

[32] J. Yu and B. Wu, "Attention and hybrid loss guided deep learning for consecutively missing seismic data reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, no. 5902108, 2022.

[33] M. Ding, Y. Zhou, and Y. Chi, "Self-attention generative adversarial network interpolating and denoising seismic signals simultaneously," *Remote Sensing*, vol. 16, no. 2, p. 305, 2024.

[34] B. Wu, Q. Xie, and B. Wu, "Seismic impedance inversion based on residual attention network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, no. 4511117, 2022.

[35] Y. Guo, L. Fu, and H. Li, "Seismic data interpolation based on multi-scale transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, no. 7504205, 2023.

[36] H. Zhang, X. Yang, and J. Ma, "Can learning from natural image denoising be used for seismic data interpolation?" *Geophysics*, vol. 85, no. 4, pp. WA115–WA136, 2020.

[37] Y. Chen, S. Yu, and J. Ma, "A projection-onto-convex-sets network for 3d seismic data interpolation," *Geophysics*, vol. 88, no. 3, pp. V249–V265, 2023.

[38] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[39] F. Brandolin, M. Ravasi, and T. Alkhalifah, "PINNslope: seismic data interpolation and local slope estimation with physics informed neural networks," *Geophysics*, vol. 89, no. 4, pp. V331–V345, 2024.

[40] R. Durall, A. Ghanim, M. R. Fernandez, N. Ettrich, and J. Keuper, "Deep diffusion models for seismic processing," *Computers*

& Geosciences, vol. 177, no. 105377, 2023.

[41] Q. Liu and J. Ma, "Generative interpolation via a diffusion probabilistic model," *Geophysics*, vol. 89, no. 1, pp. V65–V85, 2024.

[42] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 461–11 471.

[43] X. Wei, C. Zhang, H. Wang, C. Tan, D. Xiong, B. Jiang, J. Zhang, and S.-W. Kim, "Seismic data interpolation via denoising diffusion implicit models with coherence-corrected resampling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, no. 5933217, 2024.

[44] C. Meng, J. Gao, Y. Tian, H. Chen, W. Zhang, and R. Luo, "Stochastic solutions for simultaneous seismic data denoising and reconstruction via score-based generative models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, no. 5919415, 2024.

[45] F. Deng, S. Wang, X. Wang, and P. Fang, "Seismic data reconstruction based on conditional constraint diffusion model," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, no. 7502305, 2024.

[46] S. Wang, F. Deng, P. Jiang, Z. Gong, X. Wei, and Y. Wang, "Seisfusion: Constrained diffusion model with input guidance for 3-d seismic data interpolation and reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, no. 5929815, 2024.

[47] X. Wang, Z. Wang, Z. Xiong, Y. Yang, C. Zhu, and J. Gao, "Reconstructing regularly missing seismic traces with a classifier-guided diffusion model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, no. 5906914, 2024.

[48] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[49] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 13 095–13 105.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[51] W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, and J. Wang, "Sparse self-attention transformer for image inpainting," *Pattern Recognition*, vol. 145, no. 109897, 2024.

[52] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," *Advances in neural information processing systems*, vol. 34, pp. 11 781–11 794, 2021.

[53] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[54] A. Song, C. Wang, C. Zhang, J. Zhang, D. Xiong, and X. Wei, "Regeneration-constrained self-supervised seismic data interpolation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, no. 5901610, 2023.