A Physics-aware Bayesian Vision Transformer for Seismic AVO Inversion: Towards an Embodied Structural Intelligence Framework with Structure-aware Uncertainty Modeling

Zhen Liu

zhenliu.erasmus@gmail.com

China University of Petroleum (East China) https://orcid.org/0000-0003-3528-5264

Junhua Zhang

China University of Petroleum (East China)

Yongrui Chen

China University of Petroleum (East China)

Deyong Feng

Sinopec Shengli Oilfield Company

Liang Qi

Sinopec Shengli Oilfield Company

Research Article

Keywords: Pre-stack AVO inversion, Physics-Informed Neural Networks (PINN), Bayesian PINN (BPINN), BPI-ViT, Vision Transformer, uncertainty quantification, structure-aware modeling, Zoeppritz equations, interpretability.

Posted Date: October 17th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7097139/v3

Additional Declarations: The authors declare no competing interests.

A Physics-aware Bayesian Vision Transformer for Seismic AVO Inversion: Towards an Embodied Structural Intelligence Framework with Structure-aware Uncertainty Modeling

Zhen Liu, Junhua Zhang, Yongrui Chen, Deyong Feng, and Liang Qi

***Abstract*—Traditional seismic inversion frameworks struggle to preserve spatial structure and to quantify model reliability. We present a next-generation pathway that progresses from a convolutional Physics-Informed Neural Network (PINN) to a Bayesian PINN (BPINN) with uncertainty modeling, and culminates in a Bayesian Physics-Informed Vision Transformer (BPI-ViT) that enables structure-level uncertainty quantification. In our formulation, PINN "training data" are equation-domain samples used to minimize physical residuals—supporting physics-driven, data-agnostic generalization—while BPI-ViT integrates multi-layer self-attention and Bayesian inference to transition from pixel-level optimization to structure-aware collaboration. Consistent evaluation on the Marmousi2 benchmark and validation on field-scale CO_2 EOR monitoring data show that BPI-ViT outperforms prior methods in target-horizon recovery, fault and anomaly detection, spatial continuity, and uncertainty quantification, while maintaining physical consistency. These results establish a structural-intelligent paradigm that moves seismic inversion beyond error minimization toward structure-aware, reliable, and cognitively informed modeling, and provide a foundation for future multi-physics and complex-geology applications.**

***Index Terms*—Pre-stack AVO inversion, Physics-Informed Neural Networks (PINN), Bayesian PINN (BPINN), BPI-ViT, Vision Transformer, uncertainty quantification, structure-aware modeling, Zoeppritz equations, interpretability.**

This work was supported by the project "Research on CO_2 Flooding Non-uniformity Time-lapse Seismic Monitoring Technology" (30200020-23-ZC0613-0023) of the Geophysical Research Institute, Sinopec Shengli Oilfield Company.

Z. Liu is with the National Key Laboratory of Deep Oil and Gas and the School of Geosciences, China University of Petroleum (East China), Qingdao 266580, Shandong, China (e-mail: zhenliu.erasmus@gmail.com).

J. Zhang is with the National Key Laboratory of Deep Oil and Gas and the School of Geosciences, China University of Petroleum (East China), Qingdao 266580, Shandong, China (e-mail: zjh@upc.edu.cn).

Y. Chen is with the National Key Laboratory of Deep Oil and Gas and the School of Geosciences, China University of Petroleum (East China), Qingdao 266580, Shandong, China (e-mail: 348308670@qq.com).

D. Feng is with the Geophysical Research Institute, Sinopec Shengli Oilfield Company, Dongying, 257022, Shandong, China (e-mail:martineredingerjuv74@gmail.com).

L. Qi is with the Geophysical Research Institute, Sinopec Shengli Oilfield Company, Dongying, 257022, Shandong, China (e-mail: vegac3120@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

I. INTRODUCTION

The classical Zoeppritz equations are widely used in AVO analysis, but their performance is highly dependent on the accuracy of the wavelet and prior models, making the inversion results sensitive to wavelet choice and subject to strong non-uniqueness[1]. Shuey (1985) proposed a small-angle approximation to the AVO equations, which simplified the computation, yet still required high-quality priors to obtain stable inversion results and was not suitable for complex geological settings[2]. Buland & Omre (2003) introduced a Bayesian linearized AVO method, which significantly improved the stability and confidence of inversion results, but still failed to provide high-resolution estimates for deep structures[3].

Raissi et al. (2019) proposed the physics-informed neural network (PINN), embedding physical governing equations into deep neural networks by constraining the PDE residuals in the loss function, thereby enabling a physics–data dual-driven inversion framework applicable to full waveform inversion in label-scarce scenarios[4]. Karniadakis et al. (2021) further reviewed the application of PINN to engineering problems, including wavefield modeling and boundary condition handling in oil and gas[5]. In the context of seismic inversion, Rasht Behesht et al. (2021) applied PINN to two-dimensional acoustic wavefield simulation and inversion, validating its effectiveness across different levels of structural complexity[6]. Wang et al. (2021) proposed PINNeik, employing PINN to solve the Eikonal equation for precise prediction of large-scale wave travel time fields[7]. While PINN substantially improves inversion accuracy and physical consistency in small-sample regimes, its predictions are point estimates lacking quantitative uncertainty evaluation[8]. Chen et al. (2022)[9] first introduced PINN to Rayleigh wave phase velocity tomography, optimizing neural network parameters under physical constraints to suppress noise impact and achieve consistency with traditional methods, significantly reducing data requirements. Chen et al. (2023)[10] proposed a PINN-based travel time tomography for elliptical anisotropic media, enabling efficient imaging of anisotropic velocity structures in the northeastern Tibetan Plateau and reducing dependence on observed travel times, thus providing a new tool for high-resolution imaging in

complex media.

In recent years, some studies have introduced geophysical knowledge into neural network frameworks to improve the accuracy of pre-stack AVO inversion. Biswas et al. (2019)[11] proposed a physics-guided convolutional neural network, integrating AVO forward modeling and seismic wavelet convolution within neural network training, thus enhancing the physical plausibility of multi-parameter joint inversion. This method achieved good accuracy for both pre-stack and post-stack data, but because its forward module was implemented externally and did not share the same differentiable computational graph with the neural network, it limited the deep integration of physical gradients. Wang et al. (2025)[12] proposed physics-constrained deep learning (PCNN), which predicts elastic parameters and independently calculates the AVO response, thereby enhancing physical consistency. This approach benefits reservoir prediction accuracy, especially in areas with sparse well control, but its forward modeling remains external and does not realize full physical information sharing in an end-to-end differentiable graph, leading to insufficient integration.

To address the lack of uncertainty quantification in PINN, BPINN was introduced. Wen et al. (2018) proposed the Flipout layer, enabling batch-level randomization by pseudo-independent weight perturbations and reducing gradient variance[13]. Liu et al. (2021)[14] proposed the B-PINNs framework, combining Bayesian inference with PINN via variational inference and Markov chain Monte Carlo, thus enabling uncertainty modeling for noisy data in both forward and inverse PDE problems and demonstrating robustness and generalizability. Li Peng et al. (2024) presented BPINN-VI in Geophysics, applying variational inference to North Sea seismic inversion and successfully outputting mean and standard deviation maps for elastic parameters, improving inversion stability and spatial continuity[15]. Nevertheless, BPINN's uncertainty map mainly manifests as pixel-level weight distributions, with limited capability for structural-level uncertainty interpretation. Gou et al. (2023)[16] proposed a BPINN-based joint inversion method for subsurface velocity and travel time fields, incorporating Gaussian variational inference and Stein variational gradient descent to enable model-independent velocity estimation and uncertainty quantification, thus providing confidence intervals for seismic imaging.

The Transformer architecture, owing to its self-attention mechanism, has been widely adopted in both NLP and computer vision. The Transformer was first introduced by Vaswani et al. (2017), leading to breakthroughs in natural language processing[17]. Dosovitskiy et al. (2021) then proposed the Vision Transformer (ViT), treating image patches as tokens for input, and demonstrating strong global dependency modeling in image classification tasks[18]. Wang et al. (2021) proposed the ViT-based seismic velocity inversion model SVIT, which showed excellent performance in velocity prediction[19]. Simultaneously, works such as SeisMAE (2024) introduced the Transformer into self-supervised seismic inversion tra`ining, further extending its potential[20]. However, current Transformer applications still lack integration with physical forward operators and lack interpretable attention–physics consistency mechanisms[21].

Despite the individual advances in accuracy, physical consistency, and generalizability, a triangular balance among "physical consistency + uncertainty + structural modeling" has not yet been achieved. To this end, we propose a novel BPI ViT architecture that fuses PINN, BPINN, and Transformer frameworks. By introducing an interpretable attention–physics consistency mechanism, we establish a multimodal framework for next-generation seismic AVO inversion, achieving— for the first time—structure-level uncertainty visualization, deep integration of physical constraints, and high-resolution elastic parameter estimation at both theoretical and experimental levels. Details of the network architecture and implementation are provided below.

II. METHOD

A.Principle of PINN

In this study, we design a PINN framework (Fig. 1a) that explicitly embeds seismic physical priors into a deep neural network architecture, enabling efficient joint inversion of P-wave velocity (VP), S-wave velocity (VS), and density (ρ). The PINN leverages the expressive capability of convolutional neural networks for both temporal and spatial feature extraction. Taking seismic angle gathers as input, the network predicts elastic property profiles in an end-to-end manner, while the loss function explicitly incorporates both the forward physical model and multiple regularization constraints, thereby balancing data fidelity and physical consistency.

At the architectural level, the PINN accepts seismic angle gathers as input, initially encoding temporal sequence information via a one-dimensional convolutional layer (kernel size 1×60, stride 2) to extract sequential features. He normal initialization, batch normalization, and dropout are employed to accelerate convergence and enhance robustness. The resulting convolutional feature maps are flattened into a one-dimensional vector, which is then passed through fully connected layers to predict a three-component vector of physical properties. The fully connected layers utilize sigmoid activation functions to ensure output stability, and the output vector is reshaped into a structured form of shape (3, outputsize), corresponding to predictions of VP, VS, and ρ, respectively.

Distinct from conventional black-box deep neural networks, the physical components of the PINN—including the forward modeling operator, prior model constraints, and smoothness regularization—are incorporated as tensors directly within the computational graph. These, together with the network predictions, comprise a fully differentiable, end-to-end framework. The physical constraints and neural network share the same computation graph, enabling unified automatic differentiation. Thus, the gradients of the physical information can be seamlessly propagated to update the neural network weights during backpropagation, resulting in a

complete and coherent PINN.

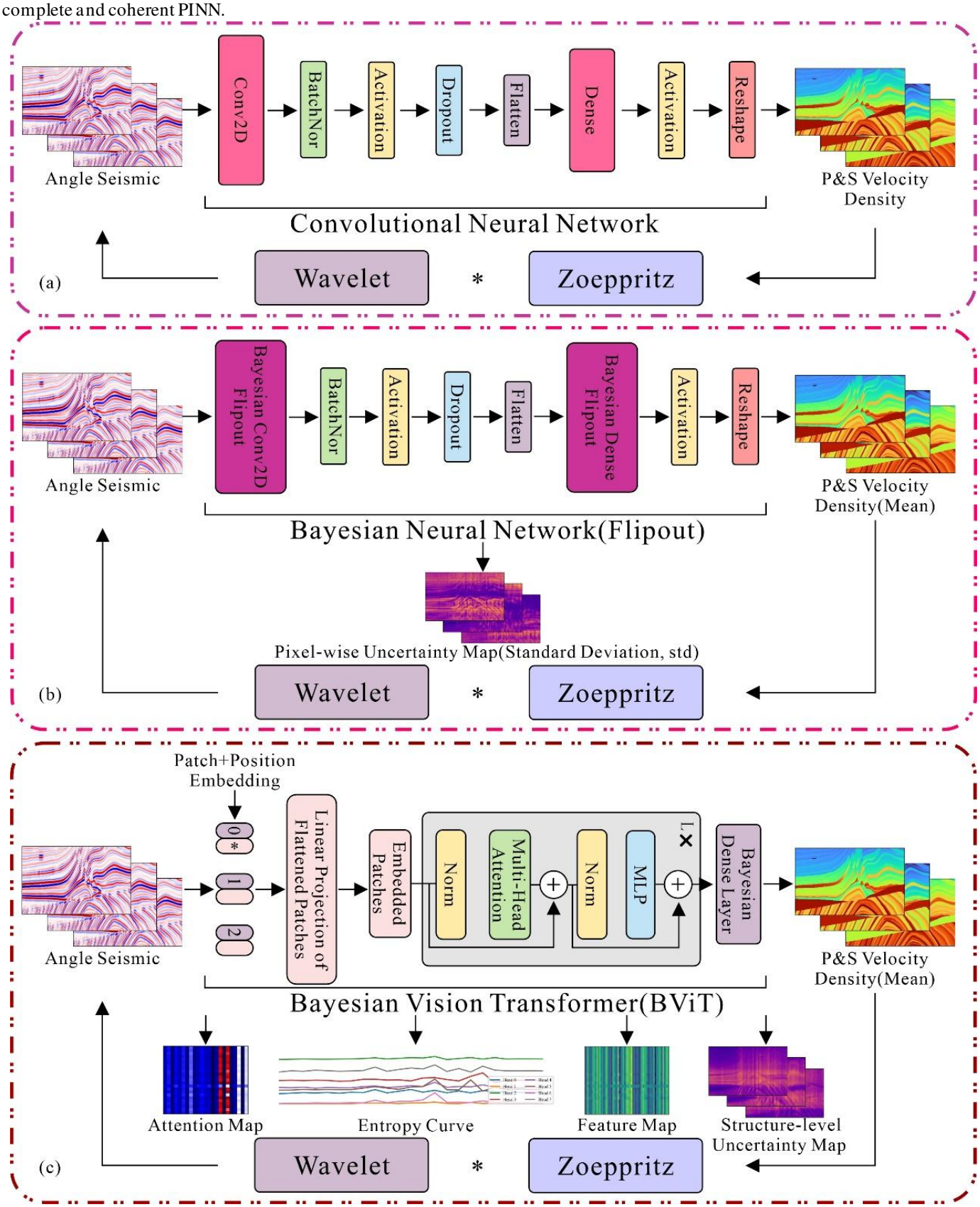


Fig. 1. Schematic comparison of three Physics-Informed Neural Network architectures. (a) Convolutional neural network PINN (CNN-PINN); (b) Bayesian PINN (BPINN) incorporating uncertainty quantification; (c) Bayesian Physics-Informed Vision Transformer (BPI-ViT) with interpretable outputs. All architectures embed physical constraints, including wavelet convolution and the Zoeppritz equations, and jointly predict P-wave velocity, S-wave velocity, and density.

In summary, the PINN model integrates convolutional feature extraction, fully connected prediction of elastic parameters, and physical consistency constraints within a tensorized, differentiable framework. This enables high-accuracy, interpretable, and physically consistent joint inversion of elastic parameters, harnessing the complementary strengths of data-driven deep learning and geophysical prior knowledge.

B.Principle of BPINN

Building upon the PINN framework, we further incorporate the concept of Bayesian neural networks to develop the Bayesian Physics-Informed Neural Network (BPINN, Fig. 1b), which enables explicit quantification of uncertainty in the parameter space. Structurally, BPINN retains the same input–output definitions and core feature extraction processes as PINN, but introduces Bayesian Flipout layers to impose probabilistic priors on the neural network parameters, thereby endowing the predictions with distributional characteristics and mitigating structural bias due to overfitting.

The core architecture of BPINN consists of a set of one-dimensional Bayesian convolutional Flipout layers and Bayesian fully connected Flipout layers. The convolutional module extracts local features from seismic angle gathers along the temporal axis using 1×60 kernels, while narrow Gaussian priors (mean 0, standard deviation 0.1) are assigned to the kernel weights and biases. The Kullback–Leibler (KL) divergence between the prior and posterior distributions is incorporated into the loss function to achieve Bayesian regularization. The fully connected layers, also implemented as Flipout layers, treat the predicted elastic parameters as samples from probability distributions rather than deterministic values, thereby enhancing robustness to outliers and high levels of noise.

Importantly, the physical constraint terms in BPINN are also embedded as tensors within the computational graph, including data misfit, first- and second-order smoothness regularizations, and prior model constraints. Because the Bayesian layers and the physical forward modeling components share the same backpropagation process, all physical information is propagated via automatic differentiation in a unified, end-to-end optimization workflow.

Compared to standard PINN, BPINN introduces an additional KL divergence regularization term into the loss function to measure the divergence between the priors and posteriors of the Flipout layer weights, thereby maximizing the model's ability to express uncertainty while fitting the observed data and physical constraints. The KL divergence is computed element-wise and scaled for each independent network parameter, which prevents excessive KL values for individual weights and ensures stable training.

In summary, BPINN retains the physical prior constraints of PINN while introducing Bayesian probabilistic modeling, enabling distributed quantification of uncertainty in elastic parameter predictions. This provides a stronger theoretical foundation for representing complex noise and non-uniqueness in seismic inversion scenarios.

C.Principle of BPI-ViT

Building upon the PINN and BPINN frameworks, we further introduce the Vision Transformer (ViT) architecture and, for the first time, integrate Bayesian layers into a standard ViT backbone, resulting in a novel Bayesian Vision Transformer (BViT) tailored for seismic inversion. BViT enables structural-level uncertainty quantification and opens new avenues for explainable scientific AI. In this study, "BViT" refers specifically to the core Vision Transformer network with integrated Bayesian layers, whereas "BPI-ViT" (Bayesian Physics-Informed Vision Transformer; Fig. 1c) denotes the full physics-informed inversion framework that embeds additional physical constraints (such as the Zoeppritz equations and wavelet convolution) atop the BViT.

The core design of BPI-ViT proceeds as follows: First, a patch embedding layer partitions the input seismic angle gathers into a series of local patches, with each patch encoded via a one-dimensional convolution and projected into a high-dimensional embedding space. This preserves local continuity of seismic waveforms while providing structured input for subsequent multi-head self-attention mechanisms. The model then applies multi-layer, multi-head self-attention modules within the Transformer encoder blocks to model long-range dependencies among patches, thereby enhancing the capacity to capture global seismic waveform features. This process also preserves multiple interpretable intermediate representations—such as attention weights, feature maps, and structural-level uncertainty measures (see Fig. 1c). On this basis, BPI-ViT incorporates Bayesian probabilistic modeling and a KL divergence annealing strategy, which further improve spatial continuity and noise robustness in the inversion of complex geological structures.

In summary, the BPI-ViT architecture synergistically combines the global feature modeling capacity of Transformers, the uncertainty quantification of Bayesian neural networks, and the physical consistency constraints of PINN. This results in a new multi-component joint inversion framework characterized by structural interpretability, robustness, and enhanced reliability.

D.Physical Information Constraints

To ensure physical consistency in seismic inversion, the Zoeppritz equations are incorporated as physical prior constraints within the PINN series models. Specifically, the predicted elastic parameters (V_P, V_S, and ρ) are used to compute the Zoeppritz P–P reflection coefficients, thereby embedding physically interpretable reflection behavior directly into the network training process and ensuring the physical plausibility of model predictions.

The Zoeppritz equations adopted in this study are implemented in a numerically simplified form and can be expressed as:

$$R_{PP} = \frac{Z_2 \cos\theta_2 - Z_1 \cos\theta_1}{Z_2 \cos\theta_2 + Z_1 \cos\theta_1} \quad (1)$$

Here, $Z_i = \rho_i V_{Pi}$ denotes the acoustic impedance of the iii-th layer, while θ_1 and θ_2 represent the incident and transmitted

angles, respectively. This formulation rigorously incorporates the critical angle constraints resulting from angle conversion and enables batch inversion across multiple angle components within the angle domain. As a complete formulation of the Zoeppritz equations, this physical model directly characterizes the continuity of reflection amplitudes at different incidence angles, thereby providing a more accurate and physically consistent constraint.

At the implementation level, the predicted tensor of the three elastic parameters is first used to compute a sequence of seismic reflection coefficients via the Zoeppritz equations. These coefficients are subsequently convolved with a Ricker wavelet to generate synthetic seismic records that match the characteristics of observed seismic data. During training, the synthetic and observed seismic records are aligned and their misfit serves as the data fidelity term in the loss function. The entire forward modeling process is fully tensorized and embedded within the differentiable computational graph of the neural network, ensuring that physical constraints are integrated into the gradient backpropagation through automatic differentiation, thus enabling end-to-end physically consistent training.

Additionally, to accurately capture reflection behavior across a range of seismic incidence angles (e.g., 0°–40°), the model computes Zoeppritz reflection coefficients in parallel for all relevant angle components, thereby enhancing prediction accuracy for wide-angle seismic data. To mitigate numerical instability from concentrated seismic wavelet energy, a standard Ricker wavelet is employed:

$$w(t) = (1 - 2\pi^2 f^2 t^2) e^{-\pi^2 f^2 t^2} \tag{2}$$

The Ricker wavelet serves as the source wavelet, providing convolutional smoothing of the Zoeppritz-derived reflection coefficients to ensure both computational efficiency and numerical stability.

In summary, the physical information constraint module in this study integrates tensorized Zoeppritz equations, Ricker wavelet convolution, and multi-angle reflection coefficient modeling. This approach achieves a fusion of physical consistency and computational differentiability, supplying interpretable and backpropagatable seismic forward modeling constraints for PINN, BPINN, and BPI-ViT deep neural network frameworks.

E.loss function

For the PINN, BPINN, and BPI-ViT physics-informed neural networks, we design a unified and extensible loss function framework that comprehensively integrates physical consistency, smoothness, and prior constraints throughout the multi-parameter seismic inversion process. The overall loss function consists of four components: the data misfit loss, smoothness regularization, prior model constraint, and the Kullback–Leibler (KL) divergence regularization from the Bayesian network.

First, the data misfit term quantifies the residual between the synthetic seismic records—generated by passing the network-predicted elastic parameters through the Zoeppritz equations and convolving with the wavelet—and the observed seismic data. The formal expression is given as:

$$L_{data} = \frac{1}{N}\sum_{i=1}^{N} \left\| d_i^{pred} - d_i^{obs} \right\|^2 \tag{3}$$

Here, d_i^{pred} denotes the synthetic seismic record generated by the forward modeling of the predicted parameters, and d_i^{obs} denotes the observed seismic record.

Secondly, to prevent sharp oscillations in the predicted results, first- and second-order difference operators are applied as smoothness regularization to the sequence of elastic parameters:

$$L_{grad} = \frac{1}{N}\sum_{i=1}^{N} \|\nabla_x m_i\|^2, \quad L_{grad2} = \frac{1}{N}\sum_{i=1}^{N} \|\nabla_x^2 m_i\|^2 \tag{4}$$

Here, m_i denotes the predicted elastic parameter for the i-th sample.

Thirdly, to ensure consistency between the model predictions and geological prior knowledge, a prior model constraint term is incorporated:

$$L_{prior} = \frac{1}{N}\sum_{i=1}^{N} \left\| m_i - m_i^{prior} \right\|^2 \tag{5}$$

Here, m_i^{prior} represents the smoothed initial (prior) model.

For BPINN and BPI-ViT networks, considering the learnable parameter distributions within their Bayesian structures, an additional KL divergence regularization term is incorporated into the total loss:

$$L_{KL} = D_{KL}\big(q(w) \parallel p(w)\big) \tag{6}$$

A linear annealing strategy is adopted to dynamically adjust the KL weight, preventing the KL term from dominating during the early stages of training and hindering convergence. The annealing of the KL weight can be expressed as:

$$\alpha_{KL}(t) = \frac{t}{T} \times \alpha_{max} \tag{7}$$

Here, t denotes the current training epoch, T is the total number of epochs, and α_{max} is the maximum KL coefficient.

In summary, the total loss function for all three network types in this study can be uniformly expressed as:

$$L_{total} = \lambda_d L_{data} + \lambda_s\big(L_{grad} + L_{grad2}\big) + \lambda_p L_{prior} + \alpha_{KL} L_{KL} \tag{8}$$

Here, λ_d, λ_s, and λ_P control the weights of the data misfit, smoothness regularization, and prior constraint terms, respectively, while α_{KL} denotes the annealing coefficient for KL regularization. For the PINN network, α_{KL} is set to zero, so only the first three loss components are activated; for BPINN and BPI-ViT, the full KL divergence term is retained to regularize the weight probability distributions.

With this structure, the framework effectively integrates data-driven residual minimization, consistency with physical prior models, and smoothness of the elastic parameter fields. Furthermore, uncertainty in the predictions is quantified under the Bayesian framework, thereby enhancing the stability and reliability of the inversion results.

III. NUMERICAL EXAMPLES

A.Model Description and Synthetic Data Generation

For consistency and comparability, this study adopts the

Marmousi2 model as the standard test scenario, as in previous work. The Marmousi2 model is widely recognized in the geophysical community as a benchmark for evaluating the performance of seismic wave simulation, inversion, and imaging algorithms, due to its complex geological structures, heterogeneous velocity variations, and multi-scale discontinuities. It simulates a typical marine sedimentary basin environment, comprising multiple faults, unconformities, and high-velocity bodies, and is therefore regarded as a rigorous and realistic test for challenging inversion scenarios.

In this study, we constructed velocity and density profiles based on the Marmousi2 model (Fig. 2a–c). Given that the computational complexity of network training grows linearly with data size, and considering that major interfaces remain resolvable at this resolution, the model was downsampled by a factor of ten to improve computational efficiency. Forward modeling was then performed using the Zoeppritz equations to generate pre-stack angle gathers at 10°, 20°, and 30°. The reflection wavelet used is a Ricker wavelet with a central frequency of 25 Hz, a sampling interval of 2 ms, and a total length of 64 samples, corresponding to a 128 ms time window. This wavelet exhibits favorable time–frequency localization, making it suitable for capturing medium-resolution stratigraphic interfaces. To achieve high-fidelity angle gather responses, convolutional modeling was performed independently for each incidence angle, ultimately producing a dataset that matches the characteristics of real seismic records (Fig. 2d–f) and serves as the basis for training and testing the neural network inversion models.

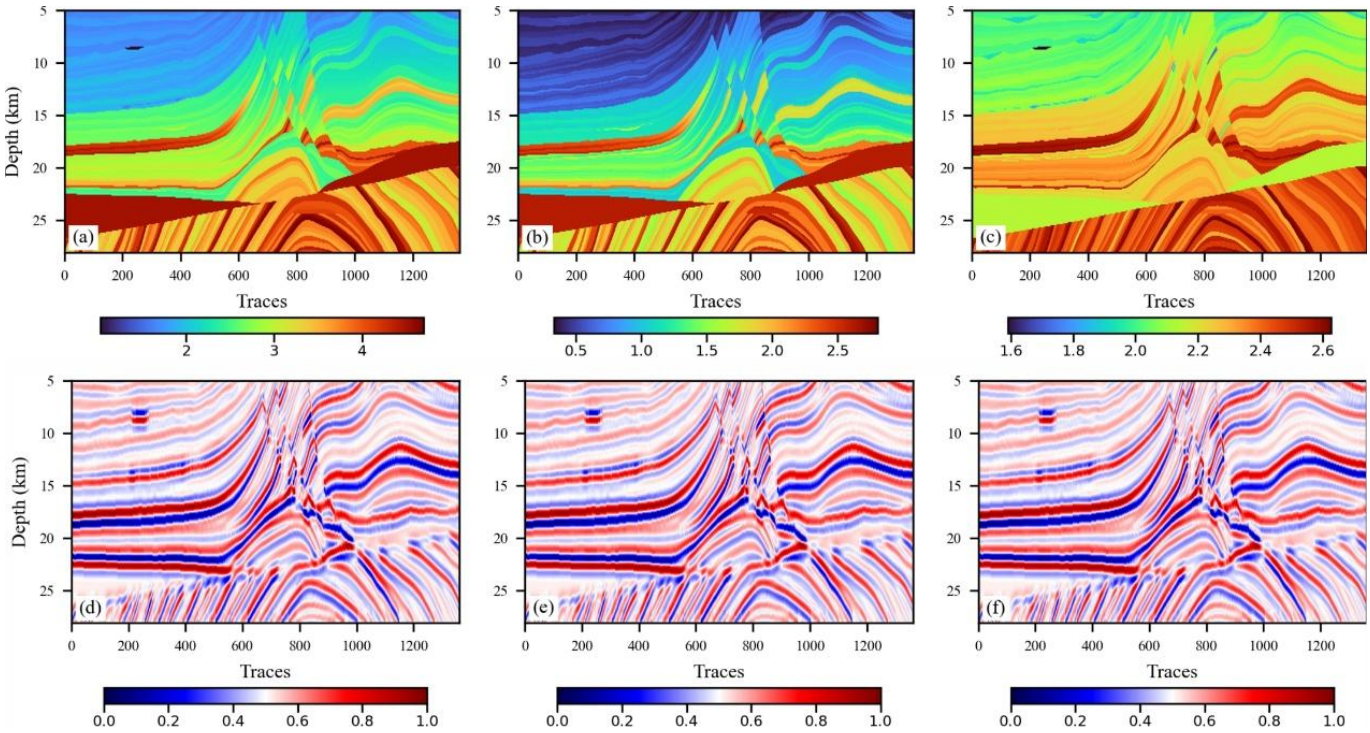


Fig. 2. Marmousi2 model: (a) P-wave velocity, (b) S-wave velocity, (c) density, and synthetic gathers at near, mid, and far angles (d–f).

B.Model Experiments and Results Interpretation

Figure 3a–c presents the loss convergence curves for the three methods during training, enabling comparison of their optimization efficiency and stability. As shown in Figure 3b, BPINN achieves rapid convergence to a low training error within relatively few iterations and maintains a steady downward trend, indicating both efficient and stable optimization. This behavior benefits from the uncertainty modeling introduced by the Bayesian framework, which helps to mitigate gradient fluctuations and overfitting. In contrast, the conventional PINN (Fig. 3a) exhibits a sharp initial loss reduction but undergoes noticeable oscillations in the later training stages, suggesting greater sensitivity to parameter initialization or optimization trajectory. BPI-ViT (Fig. 3c), while displaying a slightly slower initial convergence, shows a consistently decreasing loss that eventually stabilizes at the lowest error level, highlighting superior fitting capacity and optimization robustness. Minor undulations in the BPI-ViT curve are observed mid-training, but the method ultimately converges effectively, reflecting fault tolerance and adaptability to complex network architectures. Overall, both BPINN and BPI-ViT achieve lower loss while maintaining optimization stability, establishing a strong foundation for enhanced inversion accuracy and reliability.

Figure 4 displays the inverted distributions of P-wave velocity (V_P), S-wave velocity (V_S), and density (ρ) for the three methods on the Marmousi2 model, with Figures 2a–c providing the corresponding true model parameters. Comparison of Figures 2a–c with Figure 4 shows that all three methods are able to recover the primary layered structures,

with generally consistent results in the shallow regions. However, in the deeper layers and in areas with strong structural contrasts, notable differences in performance are observed.

For P-wave velocity inversion (Figures 4a, 4d, 4g), PINN (Figure 4a) achieves good alignment with shallow velocity variations but exhibits some blurring at greater depths. BPINN (Figure 4d) preserves the layered structure while providing sharper delineation of mid- and deep interfaces, highlighting the role of Bayesian modeling in suppressing local oscillations under uncertainty constraints. However, this method sometimes produces fragmented and discontinuous boundaries at certain faults. BPI-ViT (Figure 4g) excels in both structural boundary continuity and detailed representation of velocity gradients. Compared with the true model (Figure 2a), BPI-ViT achieves higher fidelity in reconstructing the central fault zones and deep overlapping structures, underscoring the advantage of Transformer-based global dependency modeling.

For S-wave velocity inversion (Figures 4b, 4e, 4h), PINN exhibits local oscillatory artifacts, particularly in regions with weak reflection signals. BPINN demonstrates noticeable suppression of these oscillations, but the interfaces between layers are less sharp. BPI-ViT not only effectively suppresses oscillations, but also provides smooth and well-defined boundaries, resulting in a spatial structure and numerical distribution that more closely matches the true model (Figure 2b), especially in the central dipping structure, where spatial consistency and deep attenuation robustness are superior.

Density inversion, which is the most challenging parameter (Figures 4c, 4f, 4i), reveals even greater variation among the methods. PINN produces a relatively smooth shallow density field, but the deviation increases with depth. BPINN suppresses some outliers, but boundary mixing remains an issue. BPI-ViT maintains good overall consistency with the true density trend, especially in the high-density interleaved region of Figure 2c, showing strong interlayer resolution as well as advantages in boundary continuity and smoothness across transition zones.

In summary, BPI-ViT demonstrates superior structural restoration, lateral consistency, and adaptability to deep regions across all three key parameters, making it well suited for modeling highly complex geological settings. BPINN is advantageous for controlling oscillations and introducing uncertainty modeling, especially under strong data noise. PINN offers computational efficiency and is well suited for rapid modeling of shallow targets, but exhibits limitations in complex structures. Together, these methods represent the state-of-the-art approaches in PINN-based AVO inversion in terms of accuracy, stability, and physical consistency.

Figure 5 shows the standard deviation (std) maps for the three elastic parameters (V_P, V_S, and ρ) derived from the BPINN and BPI-ViT methods, providing a quantitative measure of uncertainty in the inversion results. Figures 5(a–c) present the std maps from BPINN, while Figures 5(d–f) show the corresponding results from BPI-ViT. Overall, BPI-ViT exhibits lower uncertainty across all parameters compared to BPINN, indicating a stronger capability for uncertainty suppression and highlighting the advantage of its architectural design in maintaining inversion stability.

A closer examination of the spatial patterns in the std maps reveals distinct stripe morphologies associated with each method: BPINN produces predominantly horizontal stripes, whereas BPI-ViT more frequently displays vertical stripes. These patterns do not arise from instability in network training, but instead reflect differences in the uncertainty modeling mechanisms of their respective Bayesian structures—offering an interpretable window into model behavior. Specifically, the horizontal stripes in BPINN indicate high sensitivity of shallow layer interfaces to pixel-level perturbations, a hallmark of its response to high-frequency local fluctuations. In contrast, the vertical stripes in BPI-ViT result from patch partitioning and multi-head attention in the Transformer, representing structure-level uncertainty distributed along large-scale spatial features. These two stripe types correspond to different dimensions of uncertainty modeling capability, providing valuable diagnostic insights and structural interpretability for each approach.

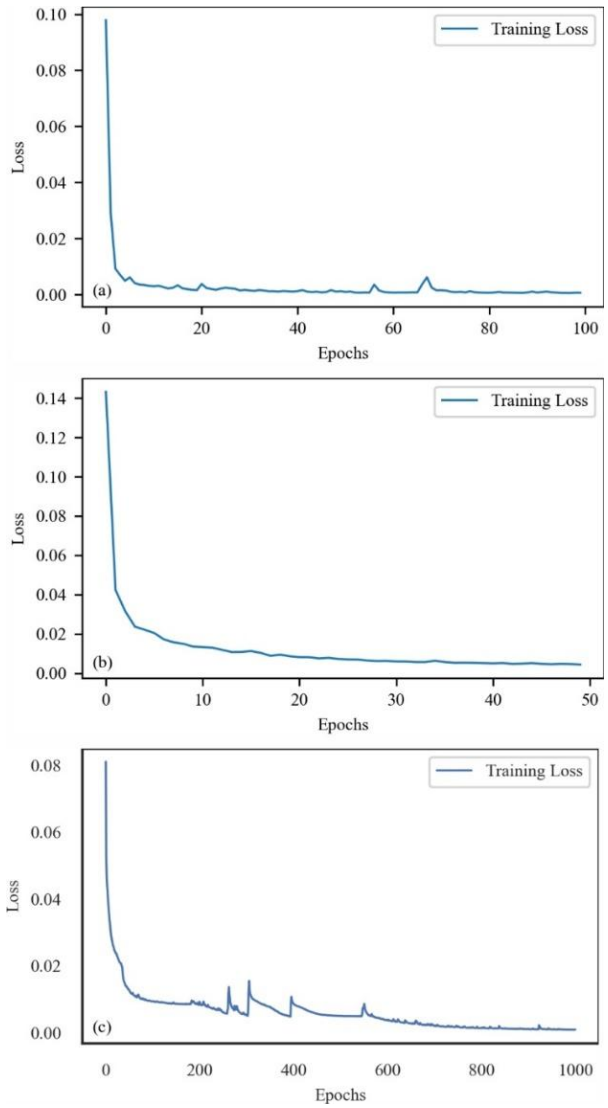


Fig. 3. Convergence curves for three physics-informed neural network architectures. (a) PINN; (b) BPINN; (c) BPI-ViT.

In summary, Figure 5 quantitatively illustrates both similarities and differences in uncertainty modeling between BPINN and BPI-ViT. BPINN demonstrates heightened sensitivity at the pixel level, while BPI-ViT achieves greater structural consistency and robustness. Notably, the ability to visualize and distinguish these uncertainty patterns via model-specific interpretability indicators constitutes a unique contribution of this study, offering a direct bridge between quantitative uncertainty analysis and physical structure recognition.

Figure 6 illustrates the attention distributions of the BPI-ViT model for seismic trace #200 across different Transformer layers and attention heads, providing insight into the spatial response characteristics and hierarchical evolution of the model's attention mechanism. Although the input seismic trace contains 231 sampling points, each attention map in the figure is represented as a 22 × 22 matrix for both the key (horizontal axis) and query (vertical axis) dimensions. This reduction is due to the use of patch embedding: the seismic waveform is first divided into several non-overlapping segments (patches), which are then used as tokens for modeling within the Transformer network. Each token thus represents a local waveform segment, and the attention maps reflect the relationships between tokens, rather than pointwise sampling. This design not only reduces computational complexity but also enhances the model's ability to capture local structural features.

Each subfigure corresponds to a specific attention head, with the horizontal axis representing the key patch position and the vertical axis representing the query patch position. The color encodes the attention strength (from blue/low to red/high). The arrangement of images from top left to bottom right reflects the progression from shallow to deeper network layers.

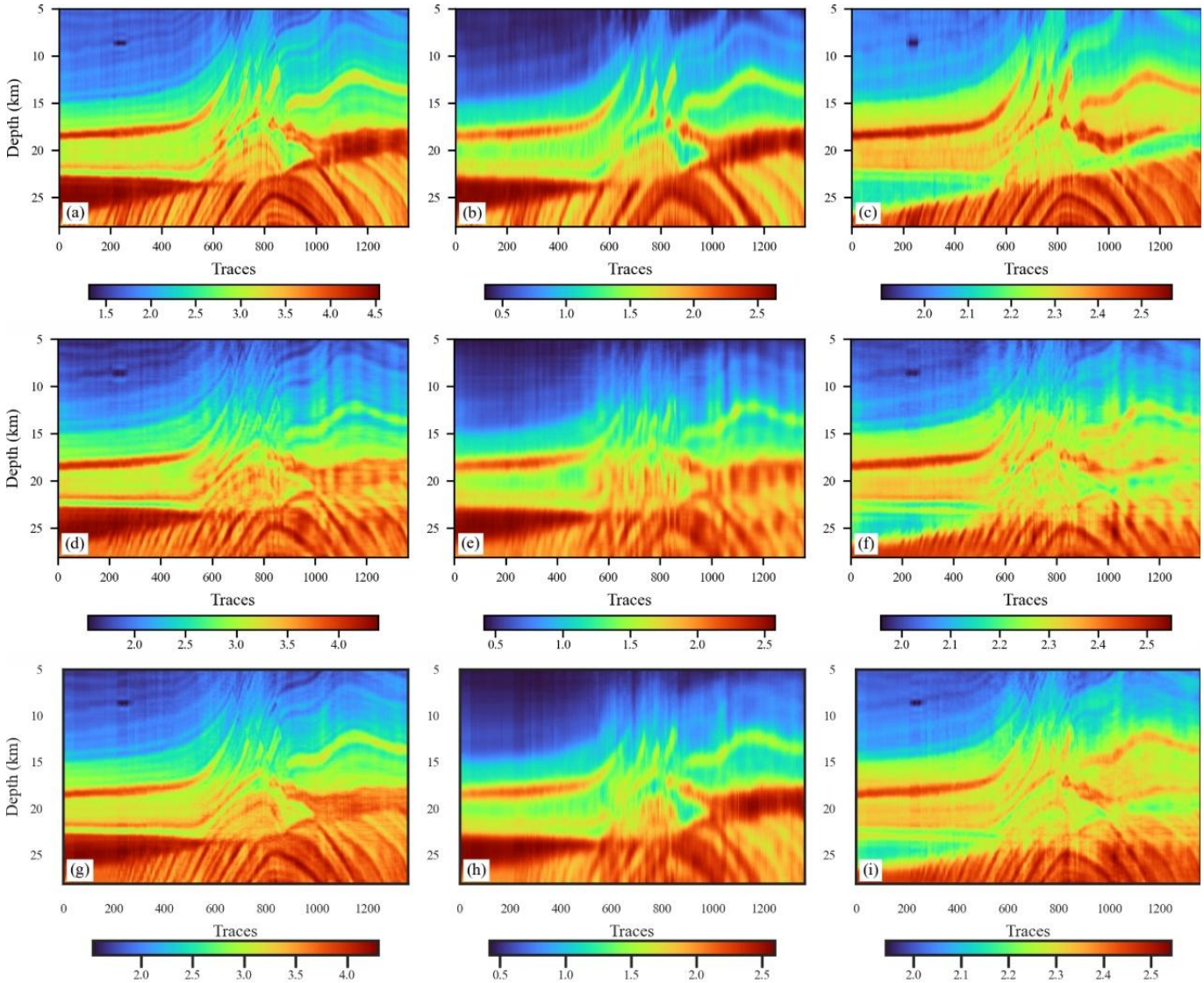


Fig. 4. Comparison of inversion results using three different methods. (a) Inverted P-wave velocity VP using PINN; (b) Inverted S-wave velocity VS using PINN; (c) Inverted density ρ using PINN; (d) Inverted P-wave velocity VP using BPINN; (e) Inverted S-wave velocity VS using BPINN; (f) Inverted density ρ using BPINN; (g) Inverted P-wave velocity VP using BPI ViT; (h) Inverted S-wave velocity VS using BPI ViT; (i) Inverted density ρ using BPI ViT.

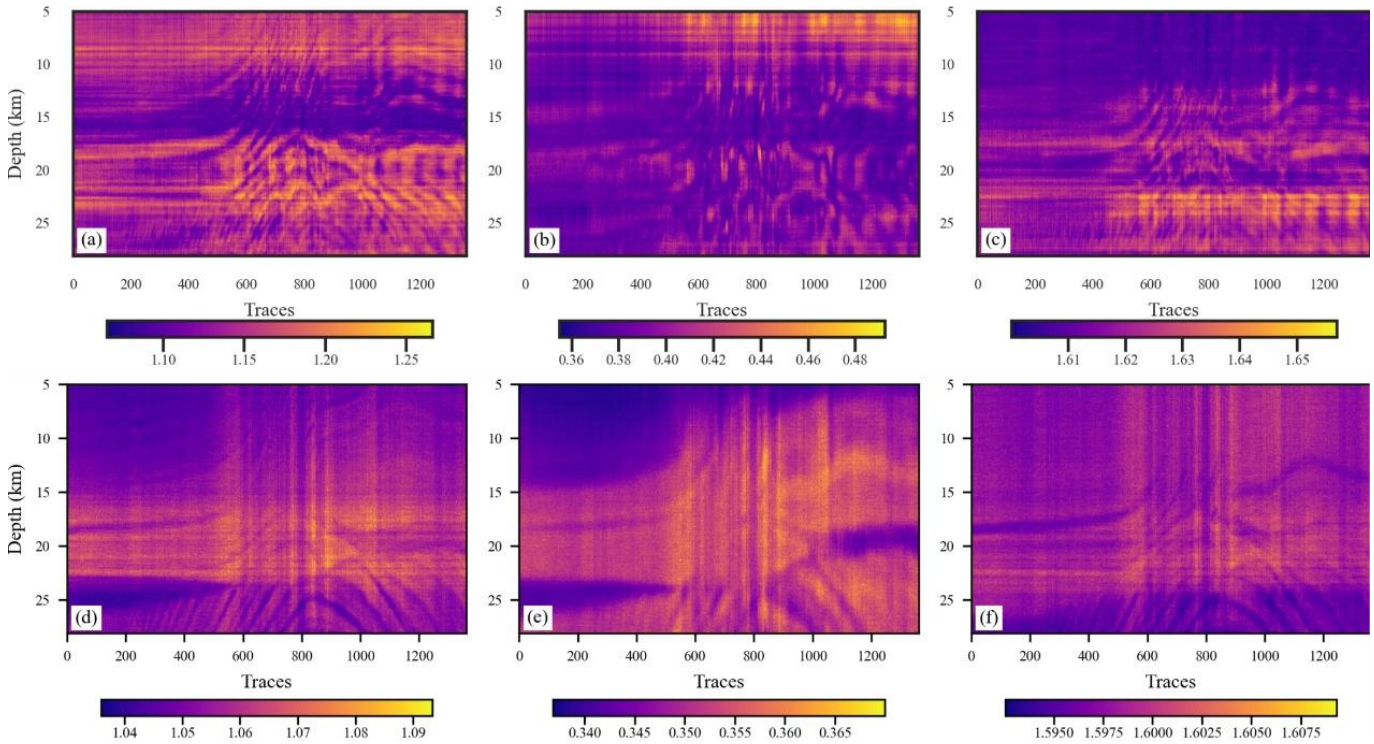


Fig. 5. Comparison of standard deviation (std) maps for BPINN and BPI-ViT in the inversion of elastic parameters. (a)–(c) represent the std of inverted V_P, V_S, and ρ using BPINN, respectively; (d)–(f) represent the corresponding results obtained by BPI-ViT.

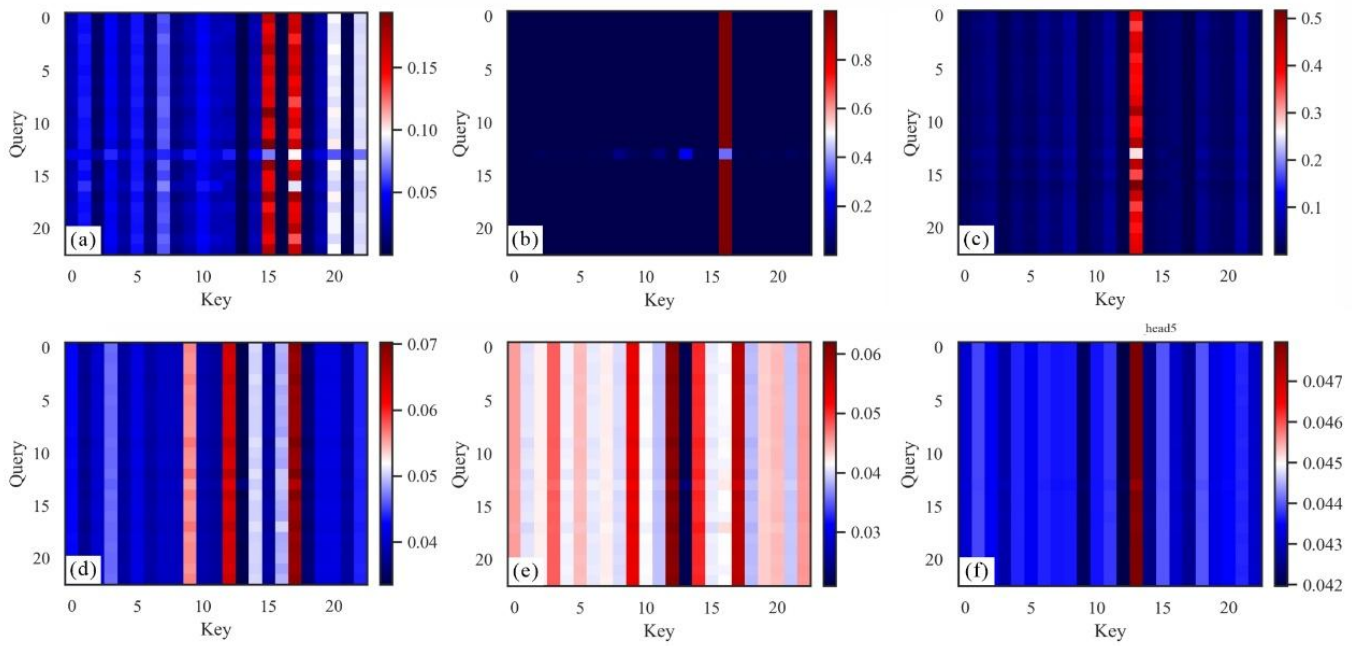


Fig. 6. Attention maps of trace #200 from different heads and layers in the BPI-ViT model. The selected heads illustrate the evolution of attention patterns from shallow to deep layers. (a) Shallow attention focused on a single reflection zone (layer0_head2). (b) Local attention along the diagonal, modeling short-range continuity (layer0_head6). (c) Sparse attention across multiple zones, indicating structural linkage (layer1_head5). (d) Distributed attention pattern capturing medium-scale interactions (layer2_head1). (e) Wide attention spread indicating global integration phase begins (layer2_head6). (f) Nearly uniform attention across all tokens in the final layer (layer3_head5).

A detailed interpretation of each map is as follows: (a) layer0_head2: attention is concentrated near the center of the map, indicating that the shallow layers are highly focused on a few prominent reflection events, which allows the model

to begin extracting abrupt velocity changes or interface anomalies; (b) layer0_head6: attention is distributed along the diagonal, showing that each query token tends to focus on its neighboring tokens, thus establishing local continuity; (c) layer1_head5: attention shifts from a focused pattern to more distributed, non-contiguous patches, suggesting the model is beginning to capture cross-patch structural features; (d) layer2_head1: attention is of moderate strength and dispersed across multiple regions, indicating that the model starts integrating features from different positions for mid-scale structure modeling; (e) layer2_head6: attention is broadly distributed over the entire map but at lower intensity, suggesting the model is entering a phase of global integration; (f) layer3_head5: attention becomes nearly uniform, with each query attending almost equally to all positions, reflecting the deep layer's capacity to integrate global structural information.

Overall, the attention mechanism in the BPI-ViT model exhibits a clear hierarchical evolution from "local focus," through "multi-point integration," to "global awareness," mirroring the cognitive logic of seismic data modeling that progresses from local interface recognition to the interpretation of complex structures. This evolution of attention not only reveals the model's ability for spatial information modeling but also provides an effective foundation for subsequent analyses of structural interpretability and geological credibility.

Figure 7 presents the statistical curves of attention entropy for the BPI-ViT model across all input samples, characterizing how each attention head in different Transformer layers distributes attention over the query tokens. The horizontal axis in each subplot denotes the query token index (corresponding to patches from the original seismic trace), while the vertical axis shows entropy values, reflecting the uncertainty in each query's distribution of attention across all key tokens.

Attention entropy serves as a key metric for quantifying the degree of attention dispersion: lower entropy indicates highly focused attention on a few tokens, whereas higher entropy reflects more uniformly distributed attention with less pronounced focus. This metric can thus quantify the structural selectivity and semantic focusing tendencies of the attention mechanism.

Each subplot corresponds to one Transformer layer (Layers 0–3), with eight entropy curves per layer, each representing an attention head. The entropy values shown are averaged over all input samples (i.e., batch mean), thus reflecting the global statistical behavior of the attention mechanism.

Overall, the evolution of attention entropy across layers exhibits a clear progression: (a) In Layer 0, entropy varies widely (approximately 1.0–2.2), with pronounced differences among heads and significant fluctuations across query tokens, indicating strong selectivity and the ability to focus on specific local reflections or boundaries; (b) In Layer 1, entropy increases slightly overall, but some heads (e.g., Head 0) maintain low-entropy, focused behavior, suggesting that local detail extraction persists even as information begins to integrate; (c) In Layer 2, all heads converge toward similar entropy values (approximately 2.8–3.1), indicating stabilization and an approach toward maximal entropy, as the model transitions into a phase of global semantic integration and structural association; (d) In Layer 3, entropy further converges to the theoretical upper bound, with almost no difference among heads, implying that deep layers distribute attention nearly uniformly to support final global feature extraction and parameter prediction.

In summary, the BPI-ViT model exhibits a clear semantic focusing evolution in its attention mechanism: strong local selectivity in shallow layers, a shift toward global modeling in intermediate layers, and ultimately fully integrated and balanced information in deep layers. The attention entropy curves serve as an interpretable quantitative metric, revealing the organizational dynamics of attention within the Transformer, and further supporting interpretability analysis and the validation of the model's structural awareness.

Figure 8 shows the feature embedding maps output by different Transformer layers in the BPI-ViT model when processing seismic trace #200. Each image represents the feature extraction results for all seismic segments (tokens) at a given layer and is visualized as a two-dimensional matrix. The horizontal axis ("Token") denotes the sequence of waveform segments obtained by equally partitioning the original seismic trace, with each token corresponding to a patch; the vertical axis ("Dimension") indicates the feature vector dimension for each token, i.e., the number of feature types extracted by the network at that layer.

Color encodes the magnitude of each feature value, with brighter colors indicating stronger activations for that segment in a particular dimension. Each column represents the full feature profile of one seismic segment, while each row displays the distribution of a specific feature across all segments. By examining the color patterns, one can assess whether the model has successfully captured structurally meaningful geological features at different layers.

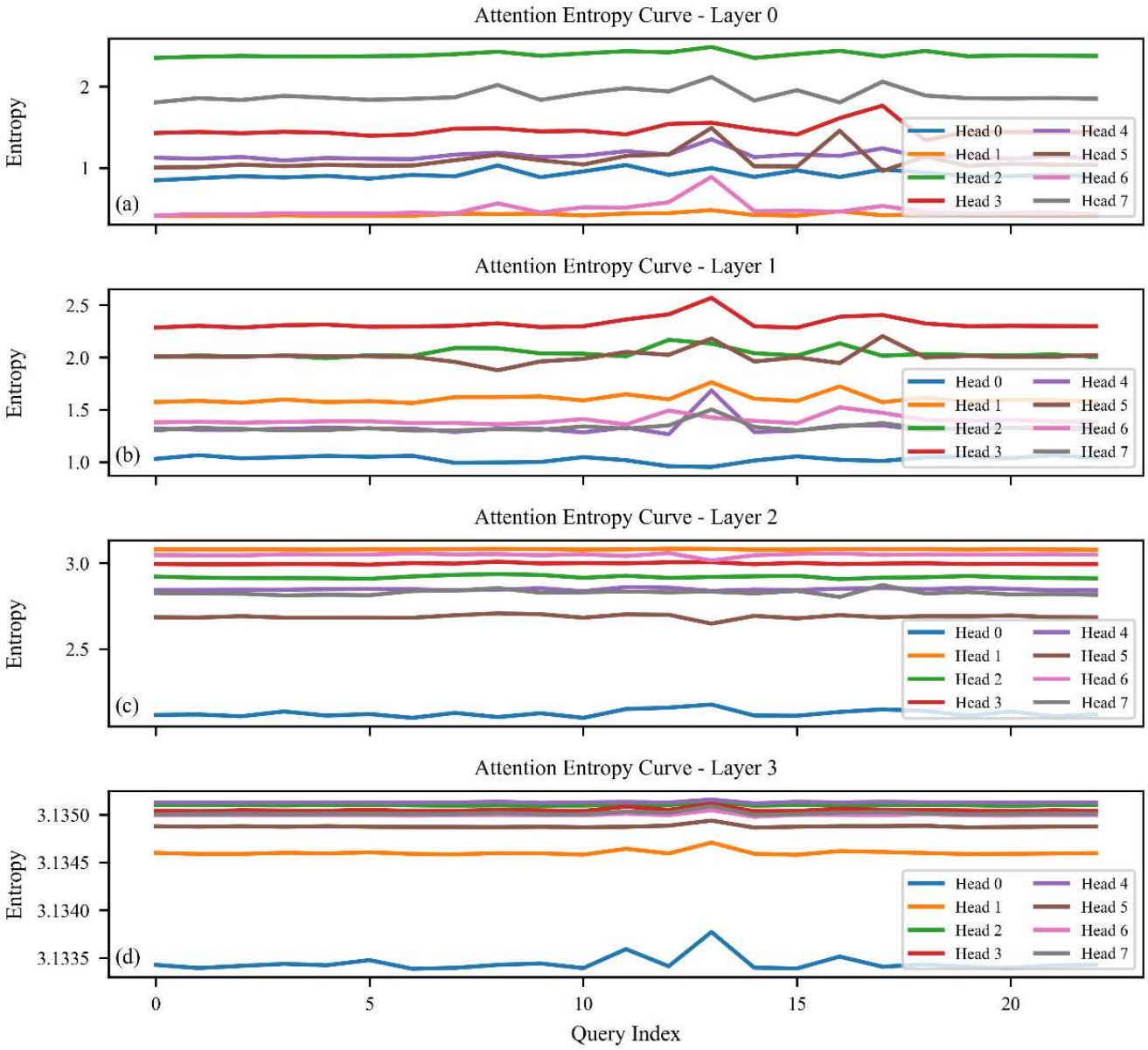


Fig. 7. Attention entropy curves of all heads across different Transformer layers in the BPI-ViT model, showing the evolving distribution of attention focus from shallow to deep layers. (a) Layer 0: Large entropy variation, strong local focus. (b) La yer 1: Slightly increased entropy, partial structure selectivity remains. (c) Layer 2: Entropy grows uniformly across heads, indicating semantic integration. (d) Layer 3: Entropy converges near theoretical maximum, fully distributed attention.

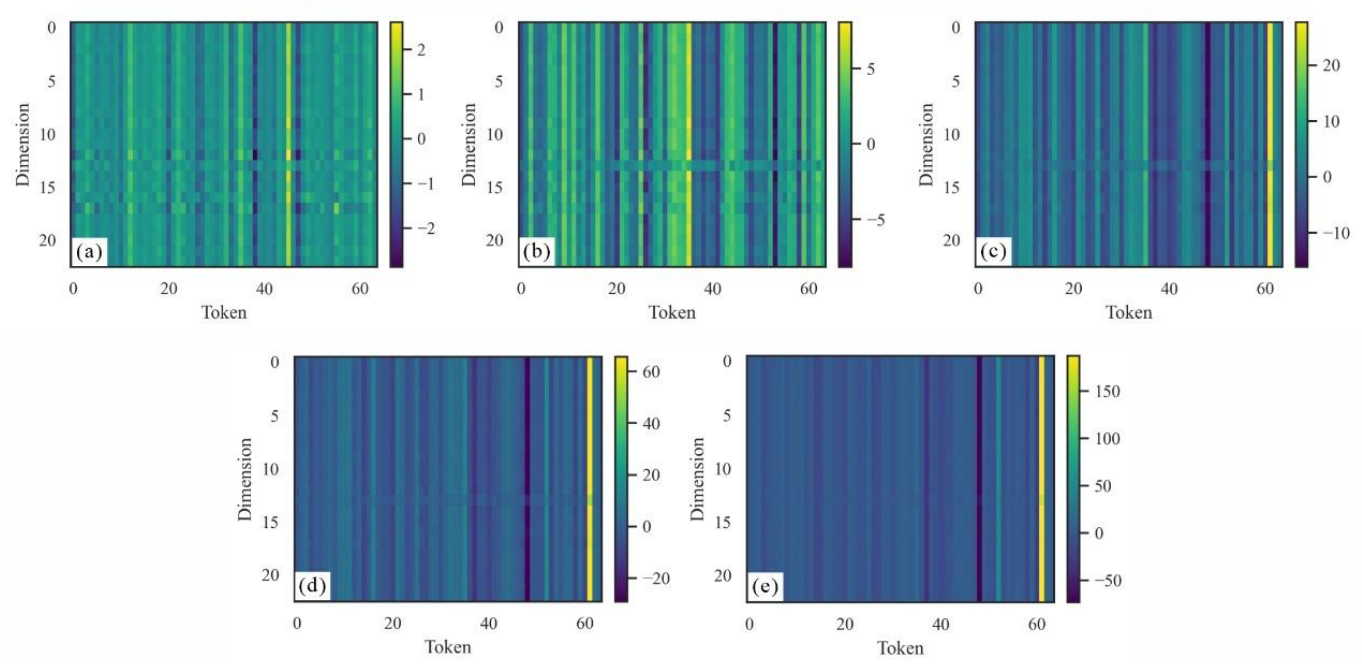


Fig. 8. Token-Dimension embedding maps of sample #200 from different layers of BPI-ViT. (a) Layer 0: disordered shallow features; (b) Layer 1: local grouping emerges; (c) Layer 2: transition stage with partial semantic alignment; (d) Layer 3: clear token-wise structural highlighting; (e) Layer 4: concentrated embedding with maximum contrast for final prediction.

Specifically: (a) Layer 0: The color distribution is scattered and fluctuates frequently, lacking obvious structural patterns. This indicates that the shallowest layer primarily captures local waveform fluctuations without forming a structured understanding of subsurface architecture; (b) Layer 1: Some vertical stripes begin to appear, showing that certain segments are co-activated along specific feature dimensions, and the model is starting to extract local structures or reflection interfaces; (c) Layer 2: The stripes become more pronounced and systematically emerge towards higher token indices, suggesting the model responds strongly to "structural transition zones" such as faults or sharp velocity changes; (d) Layer 3: The activation of a few tokens is substantially enhanced across several dimensions (notably the rightmost tokens), implying that the model is concentrating attention on critical positions and amplifying salient features; (e) Layer 4: The overall color contrast is markedly increased, with some channels displaying extremely high responses, indicating the model has accomplished feature compression and focusing, culminating in high-level semantic representations for final inversion outputs.

This evolutionary process demonstrates that the BPI-ViT model possesses robust hierarchical understanding: extracting raw signal features in the shallow layers, establishing spatial associations in intermediate layers, and focusing on key structures in the deeper layers—gradually compressing information to enable efficient and accurate semantic expression. Such hierarchical modeling not only enhances inversion prediction accuracy, but also provides strong internal support for physical interpretability.

Figures 6–8 collectively reveal the coordinated evolution of attention mechanisms and semantic modeling within the BPI-ViT architecture from three perspectives: spatial distribution, entropy dynamics, and embedding features. First, Figure 6 shows that shallow attention heads exhibit scattered and diffuse focus, while deeper heads progressively concentrate on specific regions, demonstrating a bottom-up transition from low-level perception to high-level structural focus. Figure 7 quantitatively captures this transition: entropy curves across query dimensions indicate the progressive compression of informational redundancy, with shallow heads exhibiting high entropy (dispersed information) and deep heads rapidly converging to low values—especially for certain heads (e.g., Layer 3 Head 5) that become highly focused, signifying the formation of stable and efficient high-level feature extraction.

Concurrently, the feature embedding maps in Figure 8 depict a clear trajectory of semantic representation: shallow features lack structure and display high inter-dimensional interference, while deeper layers exhibit increasingly organized and directional distributions of tokens across embedding dimensions. Notably, Layer 4 displays distinct feature compression, which closely matches the trends of attention concentration (Figure 6) and entropy convergence (Figure 7), indicating that the attention mechanism plays a critical role in guiding semantic abstraction.

The integration of these three types of visualizations provides a robust chain of evidence for the model's structural transparency and interpretability. This multi-faceted visualization approach not only elucidates how the model incrementally extracts key features from seismic gathers layer by layer, but also lays a solid theoretical foundation for future attention-guided interpretable inversion and geological interpretation.

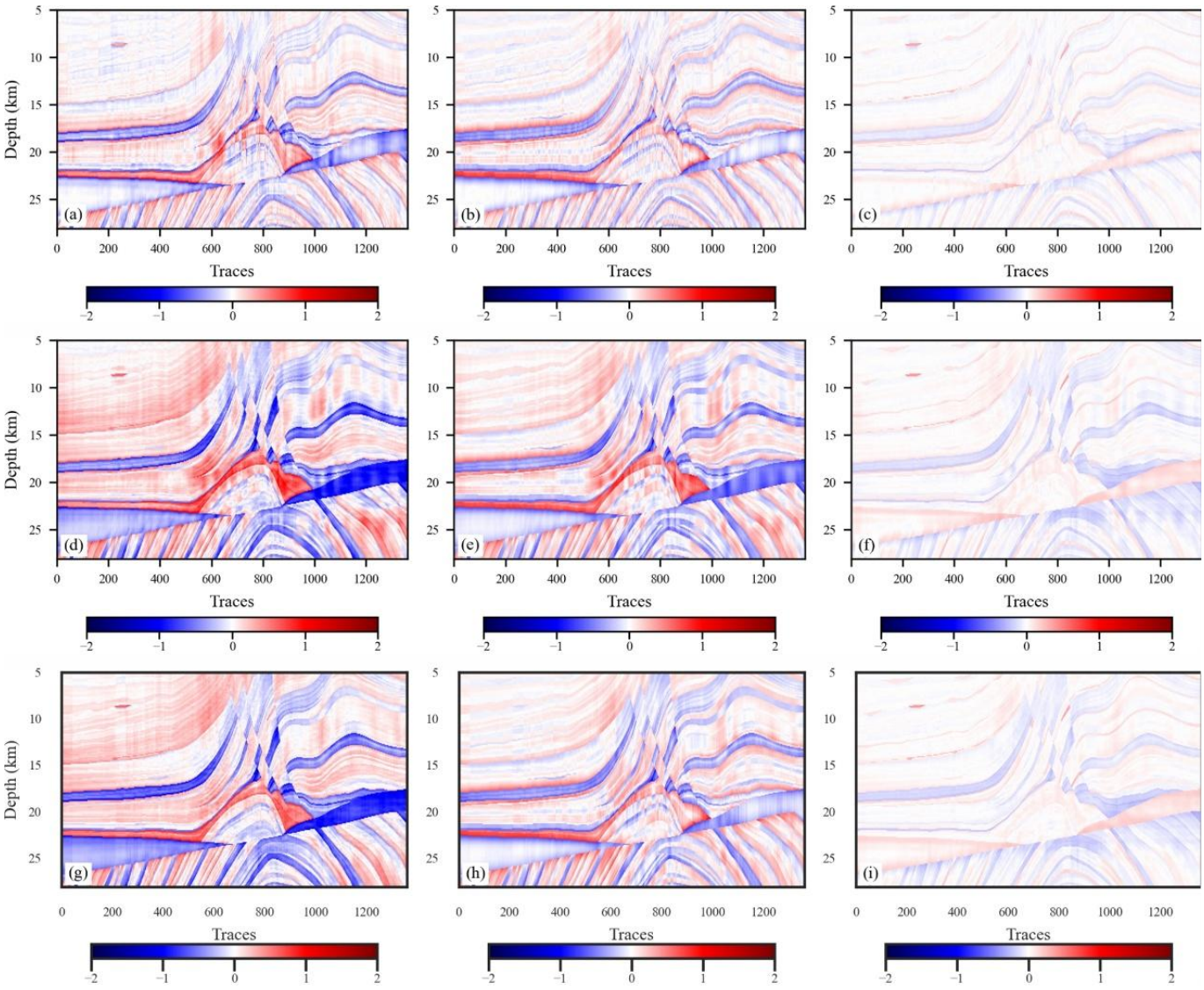


Fig. 9. Residual maps between inversion results and the ground truth using three methods. (a–c) show the differences between the predicted V_P, V_S , and ρ by PINN and the true models, respectively; (d–f) show the residuals for the BPINN inversion results; (g–i) illustrate the residuals for the BPI-ViT inversion results.

Figure 9 presents the residual maps between the inversion results and the ground truth for the three methods, providing a quantitative assessment of error distribution for each approach. Subfigures (a–c), (d–f), and (g–i) correspond to the residuals of V_P, V_S, and ρ, respectively, for PINN, BPINN, and BPI-ViT. The color scale is consistently set to [−2, 2], where blue indicates underestimation, red indicates overestimation, and white denotes perfect agreement with the true values.

From the figure, it can be observed that all three methods exhibit some degree of error in density (ρ) inversion—subfigures (c), (f), and (i)—but BPI-ViT yields the smallest magnitude of residuals, with more regular error distribution and no extensive regions of large residuals, indicating higher stability in the inversion of weakly constrained parameters. In contrast, both PINN and BPINN show significant deviations at greater depths, with BPINN displaying a pronounced band of high error between 15 and 20 km in subfigure (f).

For V_P and V_S residuals (subfigures a–b, d–e, g–h), the BPI-ViT inversion results most closely match the ground truth, with errors primarily localized in geologically complex regions, which is consistent with the inherent challenges of the inversion task. While BPINN provides a generally accurate outline, it produces more pronounced local errors near structural boundaries (as indicated by the blue regions in subfigures d and e), possibly due to sampling instability in high-gradient zones in the Bayesian model. PINN, meanwhile, shows diffusive errors across several shallow regions (subfigures a and b), likely attributable to the lack of uncertainty modeling, making it more susceptible to local oscillations.

In summary, Figure 9 demonstrates that the BPI-ViT method offers superior inversion accuracy, effectively suppressing the spread and accumulation of local errors, and exhibiting greater fitting capacity and stability, especially in

regions with complex structural boundaries and sharp parameter variations. These residual analyses further corroborate the comprehensive performance advantages of BPI-ViT in both inversion accuracy and robustness.

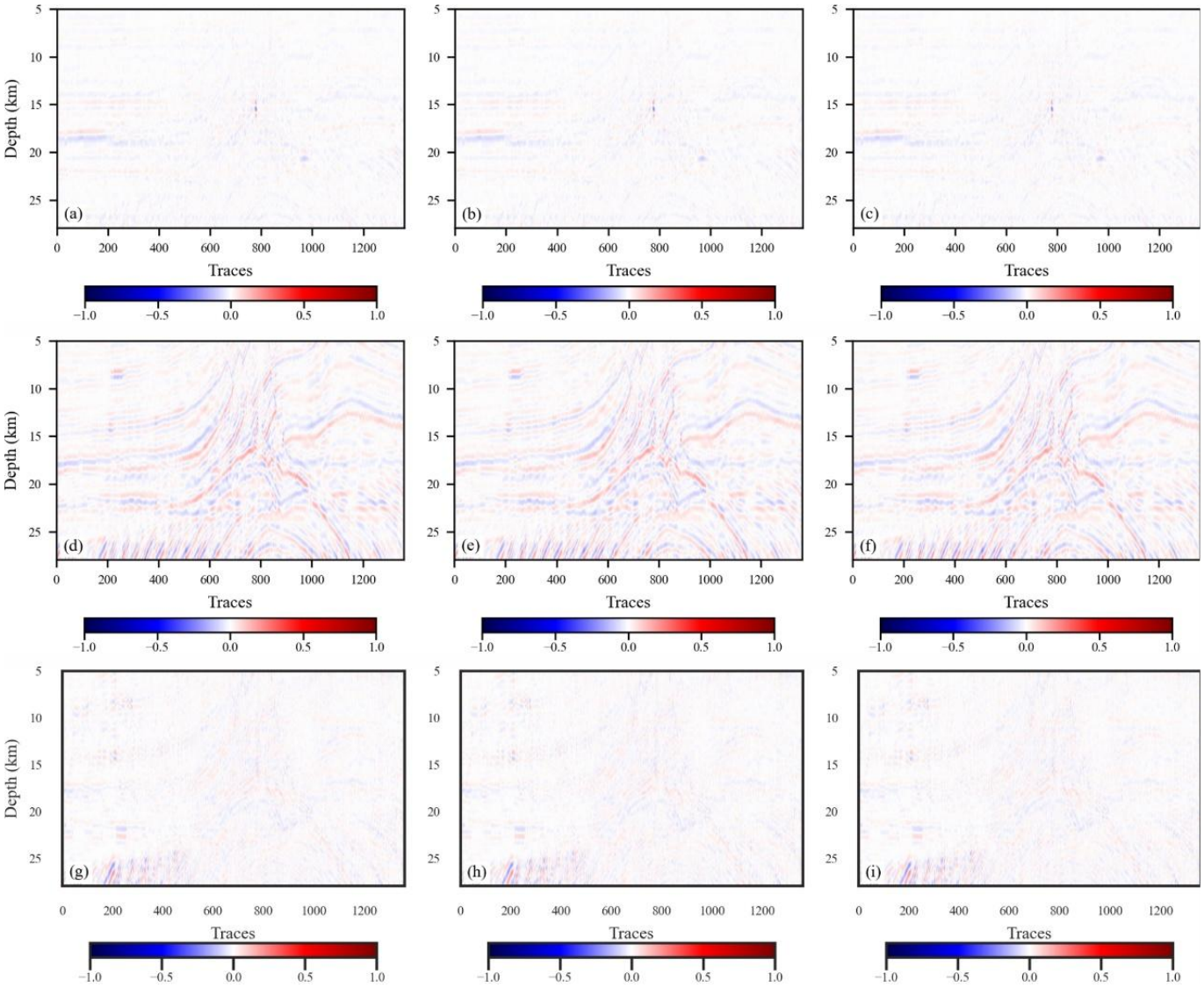


Fig. 10. Residual comparison between synthetic records generated by inversion-based forward modeling and observed seismic data. (a)–(c) show the residuals between the observed data and the synthetic records reconstructed using V_P, V_S, and ρ predicted by the PINN model; (d)–(f) correspond to the residuals obtained from the BPINN model; (g)–(i) represent the residuals produced by the BPI-ViT model.

Figure 10 displays the residual distributions between the synthetic seismic records—generated by forward modeling using the inverted property models from the three methods—and the true observed records, providing a comprehensive assessment of data-fitting performance. Subfigures a–c, d–f, and g–i correspond to the residual maps based on the V_P, V_S, and ρ inversions, respectively, for PINN, BPINN, and BPI-ViT. Overall, the residuals generated by PINN are almost negligible, with Figures a–c showing extremely low amplitudes, indicating that this method achieves exceptionally high precision in fitting the observed data. In contrast, the residuals of BPINN and BPI-ViT are slightly higher, especially near fault boundaries and velocity discontinuities, which may result from the stronger regularization or structural constraints introduced by Bayesian and structural modeling, leading to minor tensions between the inverted models and the observed data.

It is noteworthy that while BPINN and BPI-ViT exhibit certain structured residual features—such as localized shifts or amplitude differences near main faults, dipping interfaces, and high-velocity zones as seen in Figures d–f and g–i—the overall error remains controlled within ±1, consistent with typical source noise levels. Moreover, in terms of spatial residual distribution, BPI-ViT demonstrates better error uniformity and suppression at both shallow and deep interfaces compared to BPINN, corroborating its advantage in structural consistency modeling.

Figure A1 presents the synthetic records generated by

forward modeling from the inversion results of each method, provided as supplementary material. All three approaches are capable of reconstructing the main waveform features of the original seismic records, further supporting the physical interpretability and forward consistency of their inversion results.

In summary, Figure 10 clearly highlights the differences in inversion accuracy and data-fitting ability among the methods. PINN significantly outperforms the others in terms of record fitting, while BPINN and BPI-ViT exhibit additional advantages in structural and uncertainty modeling, offering valuable reference points for method selection and multi-objective inversion strategies.

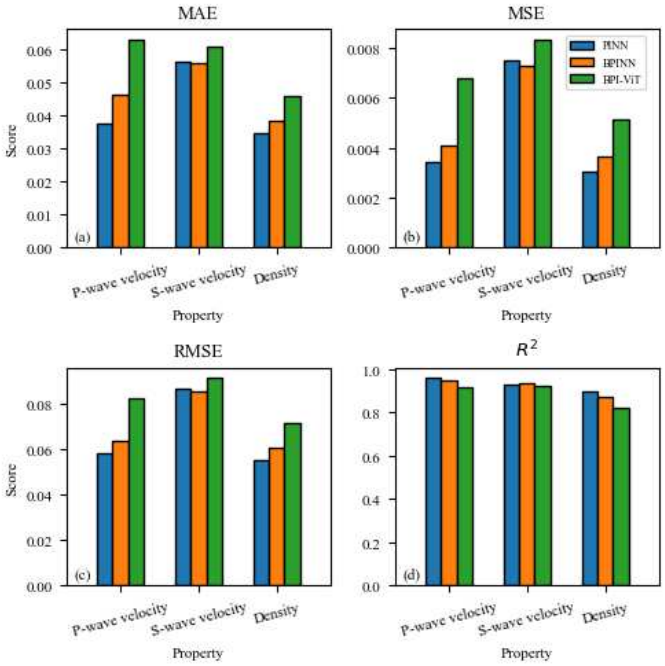


Fig. 11. Quantitative comparison of prediction accuracy for three inversion methods across different geophysical properties. (a) Mean Absolute Error (MAE); (b) Mean Squared Error (MSE); (c) Root Mean Squared Error (RMSE); (d) Coefficient of determination (R^2).

A comparison between the residuals in Figures 9 and 10 reveals the different trade-offs between inversion accuracy and data-fitting capability for the three methods. PINN achieves extremely low residuals in Figure 10, underscoring its strong ability to fit the observed seismic records; however, its property parameter residuals in Figure 9 are higher, indicating that while it can accurately reproduce the observed data, it may deviate from the true subsurface property distribution, exhibiting a form of "inversion overfitting." This phenomenon differs from conventional overfitting in machine learning, as it reflects over-adaptation to the observed data at the expense of model parameter fidelity and physical consistency. In contrast, BPI-ViT attains the smallest parameter residuals in Figure 9, with inverted properties closely matching the true model, thus demonstrating superior structural restoration and inversion robustness. Although its synthetic record residuals in Figure 10 are slightly higher, the overall data consistency remains acceptable, suggesting a greater emphasis on global structural consistency rather than local record fitting. BPINN achieves a balance between the two, maintaining effective Bayesian regularization while delivering satisfactory synthetic record fitting accuracy.

Figure 11 presents a comparative analysis of the predictive accuracy of the three inversion methods for different geophysical properties—P-wave velocity, S-wave velocity, and density. Evaluation metrics include mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), and the coefficient of determination (R^2). The results reveal performance variations across methods and properties, reflecting the dual influences of "attribute response imbalance" and "structural differences among methods" in the inversion task.

Specifically, PINN achieves the lowest error metrics (Figures a–c) and relatively high R^2 values (Figure d) for P-wave velocity and density, indicating strong pointwise fitting capability. However, for S-wave velocity inversion, PINN exhibits higher MAE, MSE, and RMSE than BPINN, suggesting limited robustness when handling medium- to high-frequency perturbations.

BPINN demonstrates more balanced performance across all three properties, notably achieving the lowest errors and highest fitting accuracy for S-wave velocity—attributable to its Bayesian uncertainty modeling, which effectively mitigates local overfitting. BPINN also outperforms BPI-ViT in density prediction, indicating stronger regularization that is advantageous for inversion in high-noise environments.

BPI-ViT, by contrast, displays relatively higher error metrics for all three properties, with the largest discrepancies observed in density, where all four metrics are the highest. This may be due to the Transformer architecture's emphasis on global structural consistency at the expense of pixel-level error suppression. Nevertheless, BPI-ViT maintains favorable fitting trends for P- and S-wave velocity, suggesting its potential in geological settings characterized by strong continuity or clear structure.

In summary, Figure 11 validates the distinct and complementary characteristics of the three approaches in terms of accuracy metrics: PINN tends to minimize residuals but is susceptible to local error amplification; BPINN maintains overall accuracy and suppresses error diffusion through Bayesian regularization; and BPI-ViT prioritizes global structural expression, albeit potentially at the cost of some pixel-level accuracy. These conclusions are further corroborated by the residual analyses in Figures 9 and 10.

Figure 12 shows the correspondence between the inversion results and the ground truth for the three methods, visualized as scatter plots to intuitively reflect the degree of fit between predicted and true values. Panels a–c correspond to the results of PINN, BPINN, and BPI-ViT, respectively. Each point represents a sample's predicted value and its corresponding true value; ideal predictions would fall near the diagonal dashed line ($y = x$), with R^2 values approaching 1 indicating stronger explanatory power of the model.

The figure demonstrates that all three methods achieve

high degrees of fit, with R^2 values of 0.96 (PINN), 0.96 (BPINN), and 0.94 (BPI-ViT), respectively. The fit curves for PINN and BPINN are more tightly clustered around the diagonal, with relatively compact error distributions, indicating stronger pointwise numerical fitting capability. Although BPI-ViT exhibits slight deviations for some predictions, the overall trend remains closely aligned with the diagonal, with a smooth and uniform distribution of data points, reflecting a greater emphasis on correctly capturing global structural patterns.

Combined with the residual analyses in Figures 9 and 10, Figure 12 provides further pointwise evidence of the fitting abilities of the three methods, comprehensively demonstrating that BPI-ViT maintains strong predictive reliability while achieving superior overall structural consistency. This figure highlights that in structurally complex regions, relying solely on pointwise numerical metrics may underestimate a model's generalization capability; thus, performance assessment should consider both residual field distributions and scatter plot analyses in combination.

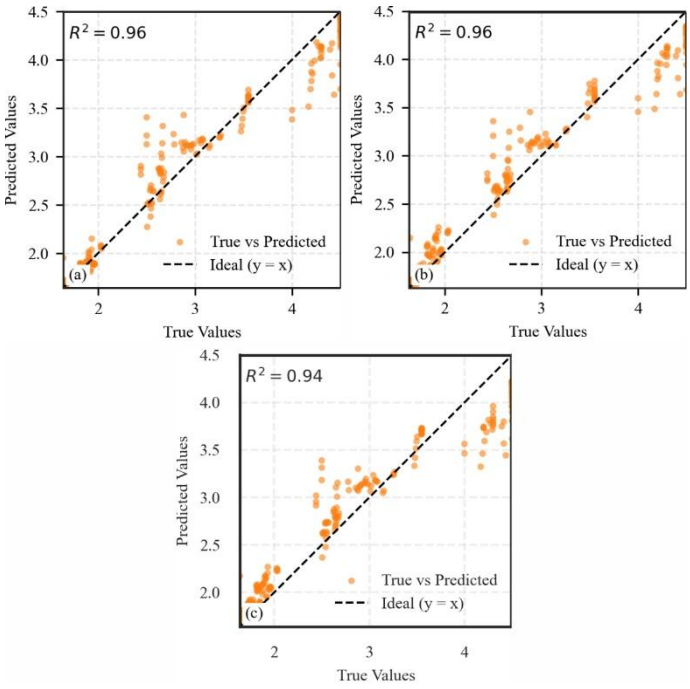


Fig. 12. Crossplots between predicted and true values obtained by (a) PINN, (b) BPINN, and (c) BPI-ViT.

Figure 13 presents violin plots comparing the predicted and true distributions of three physical parameters—P-wave velocity, S-wave velocity, and density—for the PINN, BPINN, and BPI-ViT inversion methods. Panels a–c display the results for PINN, d–f for BPINN, and g–i for BPI-ViT. In each violin plot, the left side shows the distribution of the true parameters, while the right side shows the corresponding predicted distribution; the white horizontal line denotes the median, and the shaded regions represent the main density intervals.

The results show that all methods maintain good overall distributional consistency for P-wave and S-wave velocity inversions, with the predicted distributions closely matching the true distributions in both median and shape, indicating strong stability and accuracy in modeling velocity parameters. However, for the density parameter (ρ), the predicted distributions are generally slightly narrower than the true distributions. Notably, BPI-ViT (panel i), despite some discrepancies, outperforms PINN (panel c) and BPINN (panel f) in terms of peak alignment, symmetry, and tail extension, reflecting a more reasonable modeling of weakly constrained parameters.

Additionally, the violin contours reveal that PINN exhibits certain anomalous fluctuations in the high-frequency perturbation regions (e.g., the jagged edges in panel a), indicating insufficient control over local uncertainty in the absence of Bayesian modeling. BPINN, by incorporating Bayesian inference, yields smoother overall distributions and enhanced stability. BPI-ViT, further integrating structural modeling and multi-scale information fusion, not only maintains distributional stability but also better reconstructs complex distributional features, demonstrating advantages in both structural-level fitting and distributional consistency.

In summary, Figure 13 validates the comprehensive superiority of BPI-ViT in terms of modeling accuracy and stability from a distributional perspective, highlighting its greater potential for recovering complex physical property distributions and controlling uncertainty compared to conventional PINN and BPINN approaches.

Figure 14 shows the 95% confidence intervals (CIs) of inversion results for the three physical property parameters—P-wave velocity (V_P), S-wave velocity (V_S), and density (ρ)—for BPINN (Figures a–c) and BPI-ViT (Figures d–f), providing a quantitative evaluation of the uncertainty modeling and result credibility of the two approaches.

It can be observed that BPI-ViT achieves higher consistency between the predicted mean (blue line) and the true value (black dashed line) across the entire depth range, maintaining small deviations even in regions with strong structural variation (e.g., depths of 15–20 km and 22–25 km). Additionally, the confidence intervals (blue bands) are overall narrower, reflecting superior stability and credible constraint. In contrast, although BPINN fits well in most shallow regions, its confidence intervals widen significantly in the deeper regions (below 20 km), particularly for the ρ parameter, indicating increased uncertainty and decreased result reliability at parameter boundaries and in weakly constrained zones.

Moreover, BPI-ViT produces smoother and more continuous inversion curves for all three parameters, effectively mitigating the severe local fluctuations seen in BPINN and demonstrating the clear advantage of the ViT architecture in modeling global dependencies and feature reconstruction.

In summary, Figure 14 demonstrates that BPI-ViT outperforms BPINN not only in inversion accuracy, but also in uncertainty quantification, providing more reliable predictions for subsequent geological interpretation and reservoir assessment.

In analyzing Figure 14, it should be noted that, although both BPI-ViT and BPINN exhibit some deviations from the true values in certain depth intervals (e.g., 15–17 km and 22–25 km), these differences do not undermine their scientific validity. Rather, from the perspective of uncertainty modeling, they reflect added value. The primary objective of this figure is to assess the models' ability to quantify uncertainty in inversion results, rather than merely minimizing point prediction error. BPI-ViT's mean curve is structurally consistent with the true values over most depth intervals, and its relatively narrow 95% confidence intervals indicate good stability and credibility. Furthermore, in structurally complex zones, even when predictions are slightly offset, the confidence intervals effectively cover the true trend, evidencing robustness and generalization superior to point estimates alone. Given that seismic inversion is inherently an underdetermined problem—particularly in deep regions with low signal-to-noise ratios and strong property variations—some prediction error is inevitable. Thus, BPI-ViT's ability to provide credible uncertainty intervals while maintaining reasonable prediction errors is more suited to real-world applications in complex geological environments than traditional methods lacking uncertainty modeling. In this light, the deviations observed in Figure 14 do not constitute a limitation but rather reinforce the comprehensive advantages of BPI-ViT in credible inversion and deep structural identification.

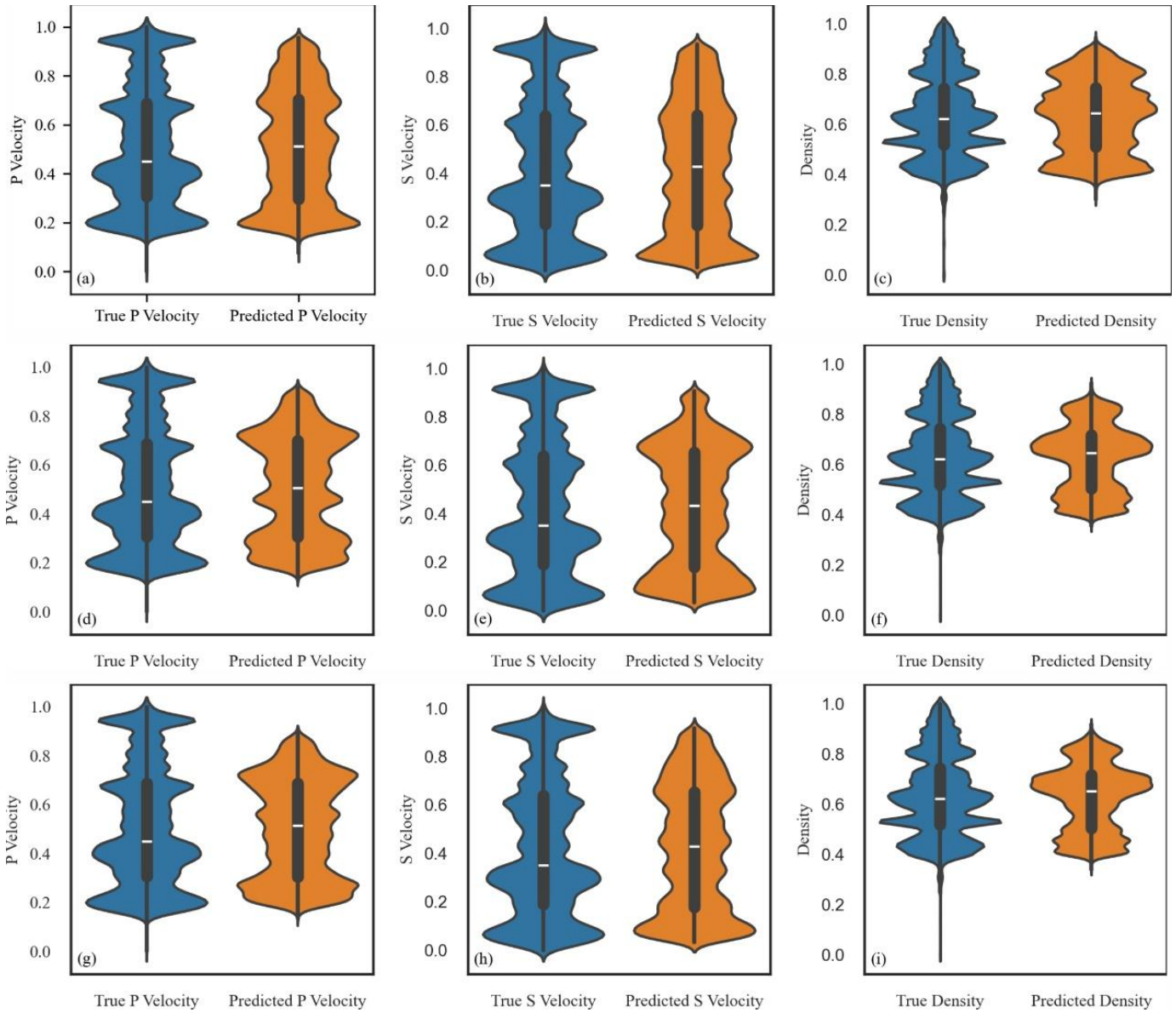


Fig. 13. Violin plots comparing the predicted and true distributions of physical properties for different inversion methods. Figures (a)–(c) show the predicted and true distributions of P-wave velocity, S-wave velocity, and density obtained using the PINN model. Figures (d)–(f) present the corresponding results for the BPINN model. Figures (g)–(i) illustrate the prediction versus ground truth distributions using the BPI-ViT model.

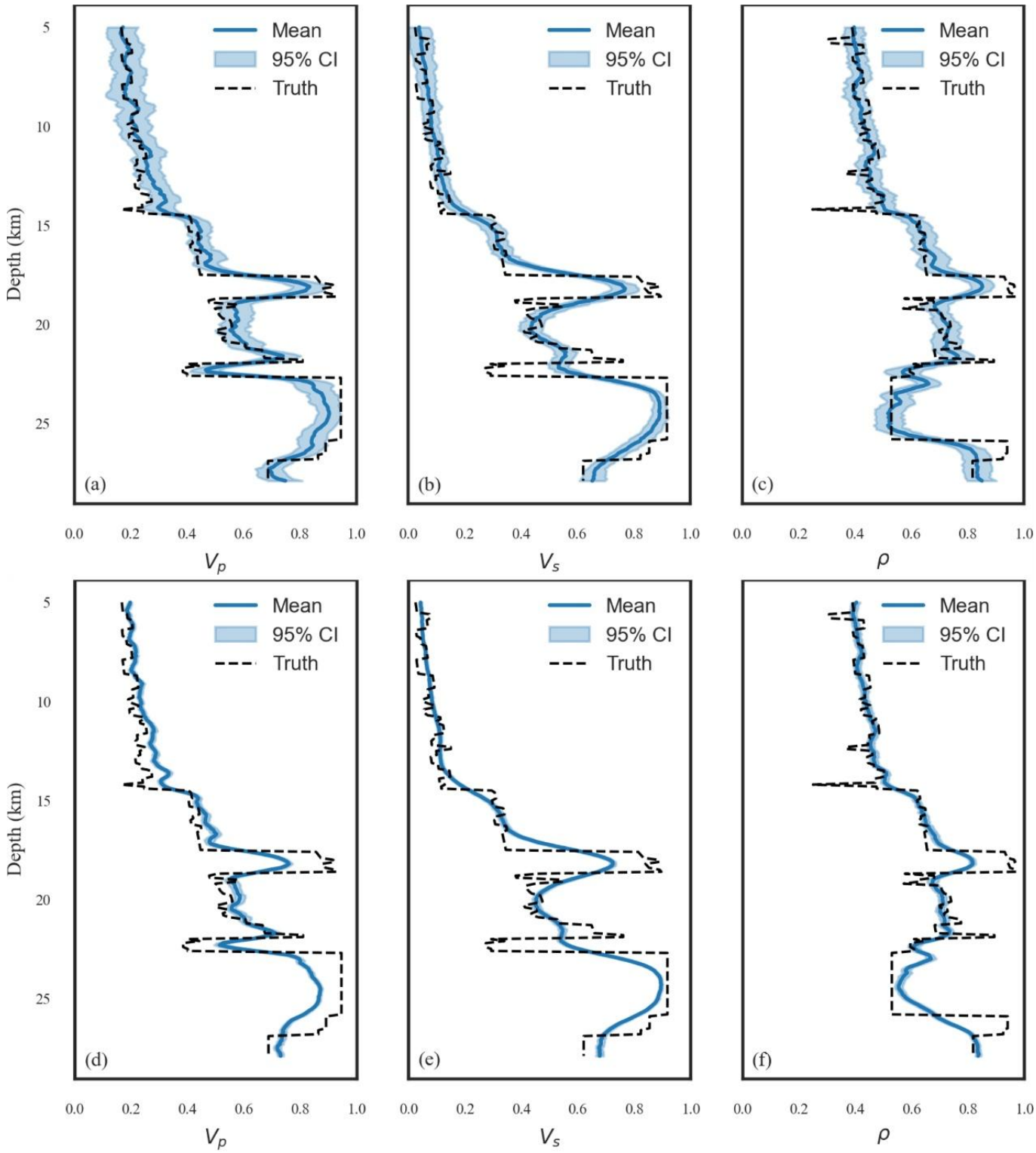


Fig. 14. Comparison of 95% confidence intervals (CI) for inversion results using the BPINN and BPI-ViT methods. (a–c) show the mean and 95% CI of inverted P-wave velocity (V_P), S-wave velocity (V_S), and density (ρ) from the BPINN method; (d–f) display the corresponding results from the BPI-ViT method. The black dashed lines represent the ground truth, blue solid lines indicate the predicted mean values, and the shaded areas denote the 95% confidence intervals.

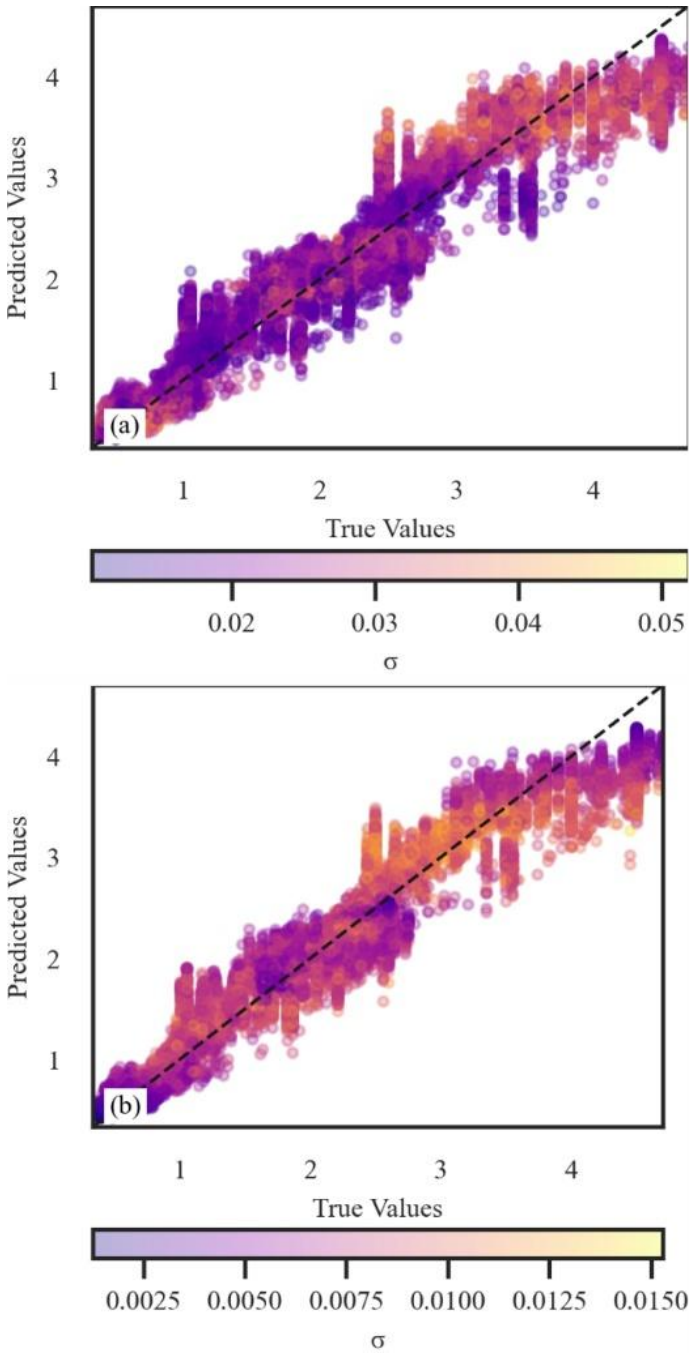


Fig. 15. Scatter plots of inversion results versus ground truth for BPINN (a) and BPI-ViT (b).

Figure 15 presents scatter plots comparing the BPINN and BPI-ViT inversion results with the true values, where the color of each point represents the standard deviation (σ) of the model's uncertainty for each prediction. This visualization captures the relationship between prediction bias and model uncertainty. In Figure 15a, BPINN's predicted points are generally concentrated along the ideal diagonal ($y = x$), demonstrating a favorable fitting trend, and the majority of σ values fall within the range of 0.02–0.05, indicating stable confidence levels for most samples. However, some predictions in the low-value region (True Value < 2.5) exhibit elevated σ, suggesting that BPINN experiences excessive uncertainty in certain weak-signal zones, which may impact model reliability.

By contrast, the BPI-ViT results in Figure 15b are also closely aligned with the ideal diagonal, indicating strong overall fitting. More importantly, the uncertainty range for BPI-ViT is significantly compressed to 0.0025–0.015, with generally higher prediction confidence and a more uniform color distribution. There is almost no noticeable aggregation of "high-σ bands," demonstrating that BPI-ViT effectively reduces uncertainty while maintaining predictive accuracy, thus enhancing stability and robustness on a global scale.

In summary, Figure 15 validates the advantage of BPI-ViT in confidence modeling from the perspective of uncertainty distribution. Its effective characterization of the relationship between prediction bias and uncertainty provides important support for subsequent confidence-based reliability analysis and structure-aware optimization.

The core objective of this work is to demonstrate the effectiveness of structural interpretability and feature modeling mechanisms in AVO inversion, with a particular focus on the theoretical construction and fundamental validation of model structural capacity and structure-aware mechanisms. The present study primarily addresses methodological innovation and mechanistic exploration, without yet incorporating computational efficiency optimization, which will be a major focus in future research. In addition, further investigations into model generalizability, robustness, and adaptability to multiple regions and structural backgrounds will be undertaken in subsequent work and presented systematically in an extended version, in order to build a more comprehensive and practical next-generation structure-aware inversion framework.

IV. Validation with Field Data

A. Overview of the Study Area

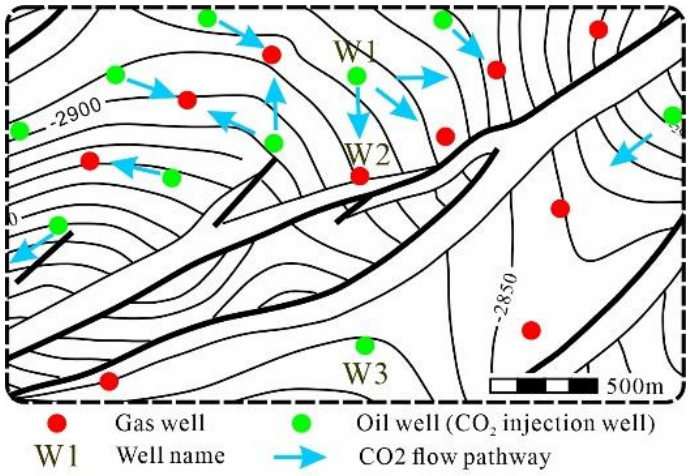


Fig. 16. Map of the study area.

This study builds upon the research framework established in previous work on PINN-based methods, utilizing real monitoring data from the CO_2 demonstration area of Shengli Oilfield to enable direct comparison with earlier

results and systematically assess the potential gains from incorporating structural interpretability and Transformer architectures.

As shown in Figure 16, the X block is located in the central part of the Boxing Sag within the Dongying Depression, belonging to the Jinjia–Zhenglizhuang–Fanjia nose-shaped structural belt. The area is characterized by complex structures and abundant hydrocarbon-bearing strata, providing favorable conditions for oil and gas accumulation and reservoir formation, as well as a suitable geological setting for enhanced oil recovery (EOR) techniques such as CO_2 flooding.

Boxing Sag is an important secondary unit in the western Dongying Depression, bounded to the west by the Gaoqing Fault and Qingcheng Uplift, to the east by the Chunhua Structure and eastern depression, to the south by the Luxi Uplift, and to the north by the Lijin Sag. The sag exhibits an asymmetric half-graben geometry, steep in the northwest and gentle in the southeast, primarily controlled by the activity of the Shicun and Gaoqing faults. During the early Paleogene, the subsidence center was located in the footwall of the Shicun Fault, gradually migrating toward the Gaoqing area in later stages. The Jinjia–Fanjia nose-shaped structure is an inherited belt trending south to north, partitioning the sag into eastern and western domains and playing a critical role in sedimentary system development and hydrocarbon accumulation.

Stratigraphically, the study area comprises, from top to bottom, the Quaternary Pingyuan Formation, Neogene Minghuazhen and Guantao Formations, and the Paleogene Dongying, Shahejie, and Kongdian Formations. The Shahejie Formation is subdivided into four members, with the fourth member (Es4) transitioning upward from red mudstone interbedded with sandstone to grey mudstone, thin-bedded sandstone, marl, and oil shale, reflecting a sedimentary evolution from arid lacustrine to humid shallow-to-semi-deep lacustrine environments. This study focuses on the upper sub-member of Es4, which is further divided into upper and lower intervals based on lithological and well-logging characteristics.

The X block serves as a key pilot area for CO_2 capture, utilization, and storage (CCUS) in Sinopec Shengli Oilfield. The injection and production wells are systematically distributed, with CO_2 injection pathways closely aligned with major structural trends, highlighting the strong geological control. Since the commencement of CO_2 injection in 2008, cumulative injection has exceeded 350,000 cubic meters, with oil production surpassing 330,000 tonnes. Implementation of CO_2 EOR has significantly enhanced recovery efficiency while delivering both economic and environmental benefits, providing an important reference for the development of similar faulted oilfields and carbon reduction practices.

B.Field Data Test

Figure 17 provides a comparative display of five geophysical property profiles inverted by the PINN, BPINN, and BPI-ViT methods. In the figure, P-wave impedance (Z_P) and S-wave impedance (Z_S) are calculated as the products of the inverted P-wave velocity (V_P) and density (ρ), and S-wave velocity (V_S) and density (ρ), respectively, i.e., $Z_P=V_P\times\rho$ and $Z_S=V_S\times\rho$. The profiles are annotated with ellipses marking the target reservoir interval for CO_2 monitoring and arrows indicating the main flow direction of CO_2, providing a clear visualization of the geological setting and fluid migration pathways within the injection-production area.

The results show that all three methods successfully recover the major geological horizons and fault structures. In the inverted P-wave velocity, S-wave velocity, and density profiles (a–i), BPI-ViT delivers more prominent structural details and continuity, particularly at well locations and within the target interval (elliptical areas), with clearer attribute interfaces at faults and thin interlayers. BPINN demonstrates improvements over PINN in imaging interlayer interfaces and local anomalies, though it remains slightly inferior to BPI-ViT in spatial continuity and anomaly delineation. The PINN method exhibits a certain degree of interface blurring and noise in high-gradient zones and weak reflection bands.

For the P-wave and S-wave impedance profiles (j–o), BPI-ViT achieves higher interface clarity, with attribute distributions around the target reservoir and CO_2 migration pathways (arrowed areas) in good agreement with geological expectations. BPINN performs well in resolving interfaces and anomaly blocks, while PINN shows limited capacity for identifying properties in some thin layers and weak reflection zones.

Overall, all three methods effectively invert the geological attributes of the main reservoir and faulted zones. BPI-ViT, in particular, provides enhanced stratigraphic continuity and delineation of anomalies at the target interval and CO_2 migration channel, offering a more detailed geophysical information foundation for subsequent complex reservoir inversion and CO_2 monitoring.

Figure 18 presents the standard deviation profiles of P-wave velocity (V_P), S-wave velocity (V_S), and density as inverted by the BPINN and BPI-ViT methods. Panels (a)–(c) display the standard deviations for V_P, V_S, and density from BPINN, while panels (d)–(f) show the corresponding results from BPI-ViT. Ellipses indicate the target reservoir interval, and arrows mark the main flow direction of CO_2.

The standard deviation distributions reveal that BPINN reflects pixel-level uncertainty. Its standard deviation maps show localized increases at layer interfaces and within local anomaly blocks, particularly near well locations and CO_2 injection–production channels, indicating greater uncertainty in inversion results for these areas. Overall, BPINN achieves high spatial resolution in its uncertainty distribution, but its ability to suppress noise in structurally complex zones is limited, resulting in relatively pronounced local uncertainties in the attribute profiles.

In contrast, BPI-ViT realizes structural-level uncertainty modeling. The standard deviation profiles (d–f) reveal that BPI-ViT maintains lower and smoother uncertainty levels, especially within the main reservoir, fault zones, and CO_2 migration pathways, demonstrating superior spatial consistency and interface continuity. This indicates that BPI-

ViT not only enhances the overall robustness of the inversion, but also effectively suppresses attribute uncertainty in complex geological blocks, reflecting its higher resolving power for structural features and anomalies. Particularly at the target interval and CO_2 injection channels, the structure-level uncertainty model provides stronger support for inversion credibility and anomaly identification.

In summary, the BPI-ViT method offers significant advantages in structure-level uncertainty modeling, markedly improving the reliability of inversion results for main reservoirs and fluid migration zones. This provides a new approach and technical foundation for seismic attribute uncertainty analysis in complex hydrocarbon reservoirs and CO_2 monitoring fields.

Figure 19 displays the multi-head attention weight distributions for representative seismic traces near well locations, revealing the BPI-ViT model's capacity to model spatial relationships in seismic data across different layers and attention heads. Specifically, panels (a)–(c) show the attention distributions of head 2 in the 1st, 2nd, and 3rd layers, while panels (d)–(f) illustrate the distributions for head 1 in layer 0, head 3 in layer 0, and head 6 in layer 1, respectively.

The overall patterns reveal substantial differences in attention weight distributions across layers and heads, highlighting the BPI-ViT model's multi-scale capability for capturing spatial structural information in seismic trace sequences. In panels (a)–(c), as network depth increases, the attention weights become more evenly distributed, with fewer localized high-weight regions, indicating that deeper layers are more effective at integrating global information. In panels (d) and (e), the attention is highly concentrated, with most regions exhibiting low weights except for specific key positions. This suggests that certain heads in early layers focus intensely on particular spatial relationships, which aids in capturing critical geological structures. Panel (f) exhibits a more complex and dispersed weight distribution, with some regions displaying higher weights than their surroundings, indicating the model's ability to flexibly model complex structures or anomalies at intermediate layers and specific attention heads.

Overall, the multi-head attention mechanism endows the model with a rich capacity for spatial feature representation, enabling BPI-ViT to effectively capture complex relationships between seismic traces at multiple levels and from various perspectives. This forms a solid foundation for improving the precision of reservoir structure identification and inversion performance.

Figure 20 shows the feature representations of seismic traces near well locations across different Transformer layers, aiming to elucidate the hierarchical spatial modeling capability of the BPI-ViT model for seismic signals. Specifically, panels (a)–(e) correspond to feature representations from layers 0 to 4, with the horizontal axis indicating token indices, the vertical axis representing feature dimensions, and color denoting feature value magnitude.

The distribution patterns reveal that in the shallow layers (a, b), features display significant variability across multiple dimensions and tokens, with dispersed distributions and localized highlights or low values, indicating that the model is highly sensitive to signal details and local structures at early stages. As the layer depth increases, panels (c)–(e) exhibit increasingly concentrated highlight regions, with feature values gradually focusing on specific tokens and more pronounced extreme values along certain dimensions. The overall features demonstrate greater spatial consistency and integrity. This progression reflects the ability of deep Transformer layers to effectively integrate detailed information extracted by shallow layers, enabling high-level abstraction and fusion of spatial structures and thereby strengthening the model's capability to characterize key stratigraphic horizons and important geological structures.

In summary, the BPI-ViT model achieves a progressive spatial feature extraction process across different layers, transitioning from detailed depiction to global abstraction. This provides a rich informational foundation for efficient identification of complex reservoir structures and property inversion.

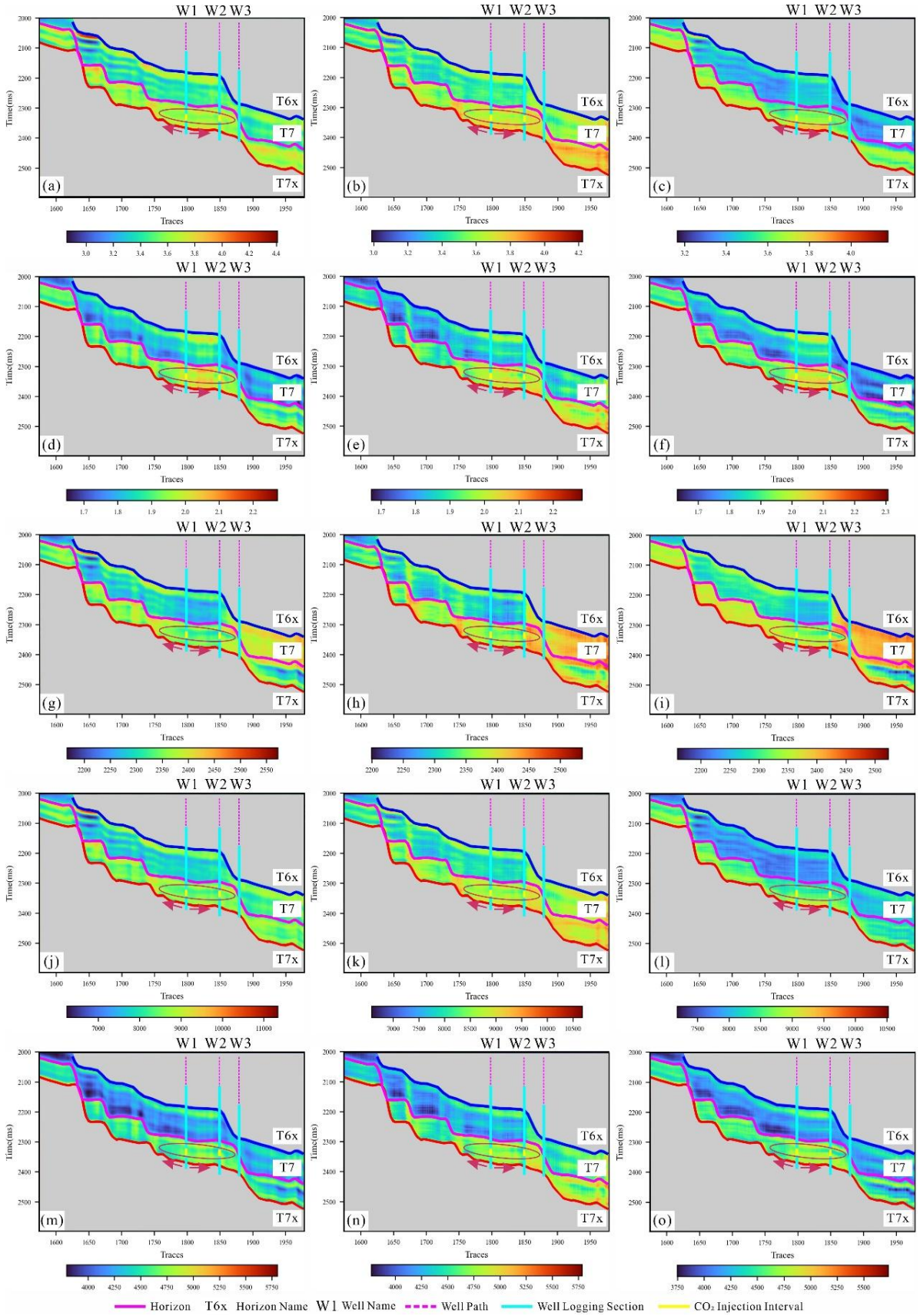


Fig. 17. Comparison of inversion sections for five geophysical properties using PINN, BPINN, and BPI-ViT. (a)–(c) show the inverted P-wave velocity (V_P), (d)–(f) show the inverted S-wave velocity (V_S), (g)–(i) show the inverted Density, (j)–(l) show the inverted P-wave impedance (Z_P), and (m)–(o) show the inverted S-wave impedance (Z_S). Each row corresponds to a specific property, while each column represents the inversion results from PINN, BPINN, and BPI-ViT, respectively.

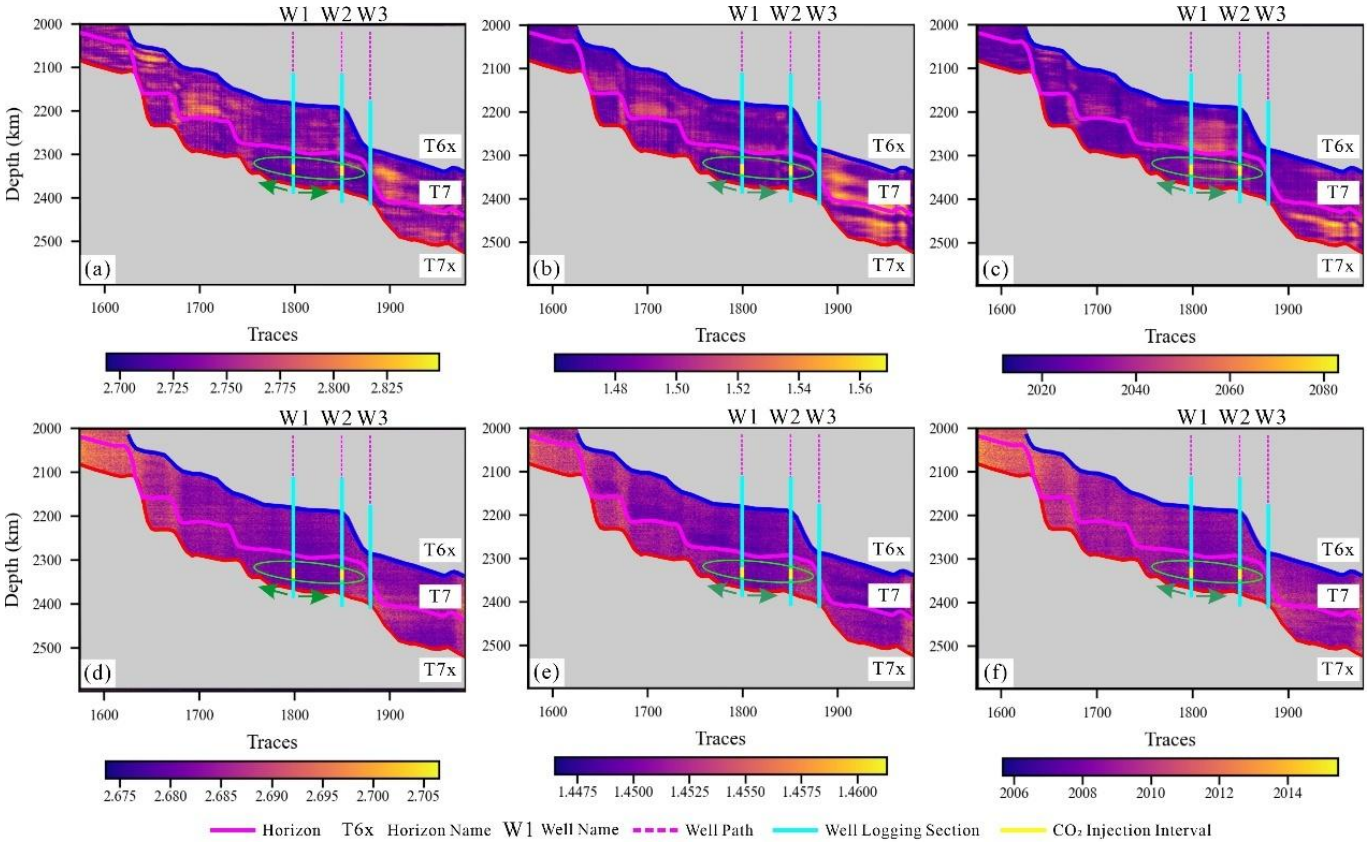


Fig. 18. Comparison of standard deviations for the inversion results of V_P, V_S, and density using BPINN and BPI-ViT. (a)–(c) Standard deviations of inverted P-wave velocity (V_P), S-wave velocity (V_S), and density by BPINN; (d)–(f) Standard deviations of inverted P-wave velocity (V_P), S-wave velocity (V_S), and density by BPI-ViT.

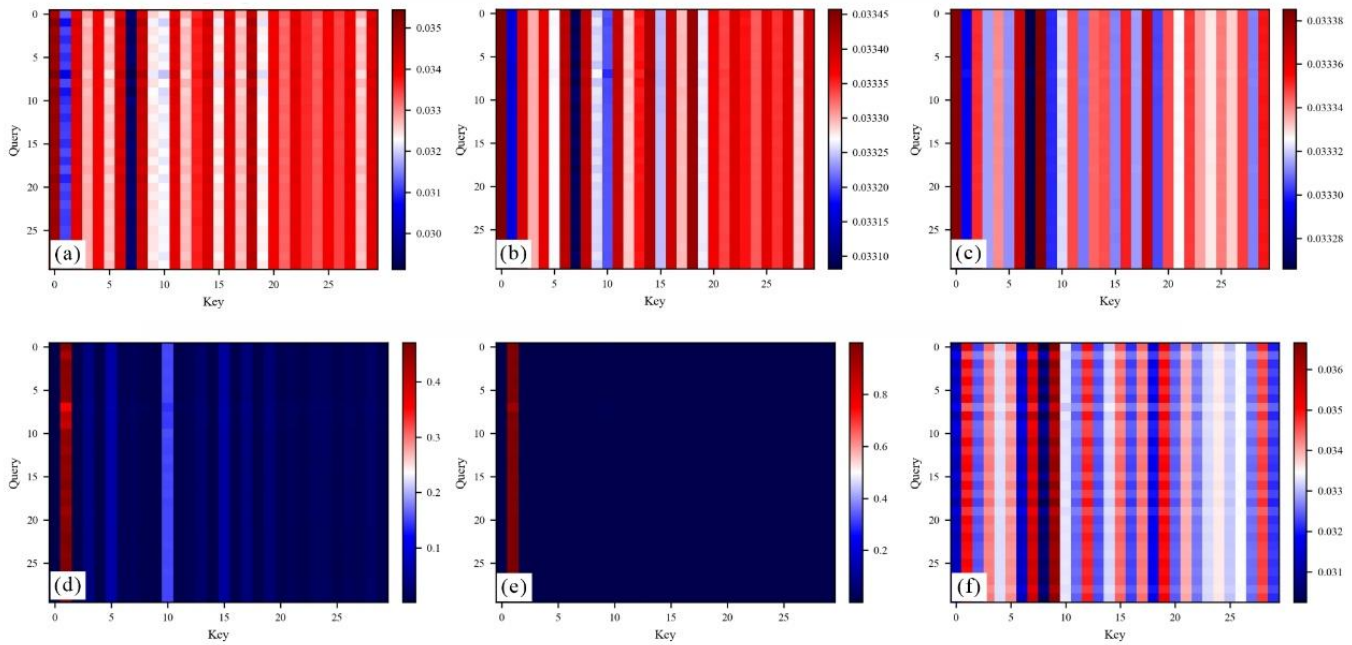


Fig. 19. Representative multi-head attention weight distributions at the seismic trace near Well A. (a) layer1_head2; (b) layer2_head2; (c) layer3_head2; (d) layer0_head1; (e) layer0_head3; (f) layer1_head6.

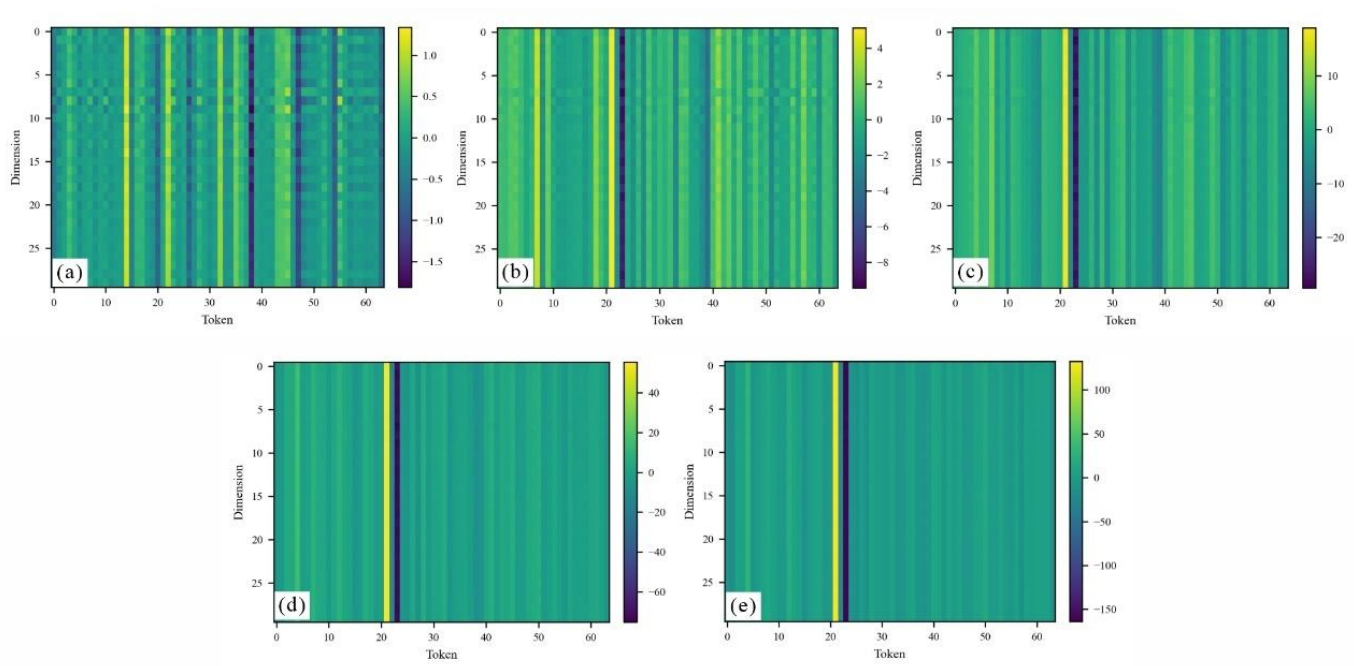


Fig. 20. Feature representations of the seismic trace near Well A at different Transformer layers. (a) layer0; (b) layer1; (c) layer2; (d) layer3; (e) layer4.

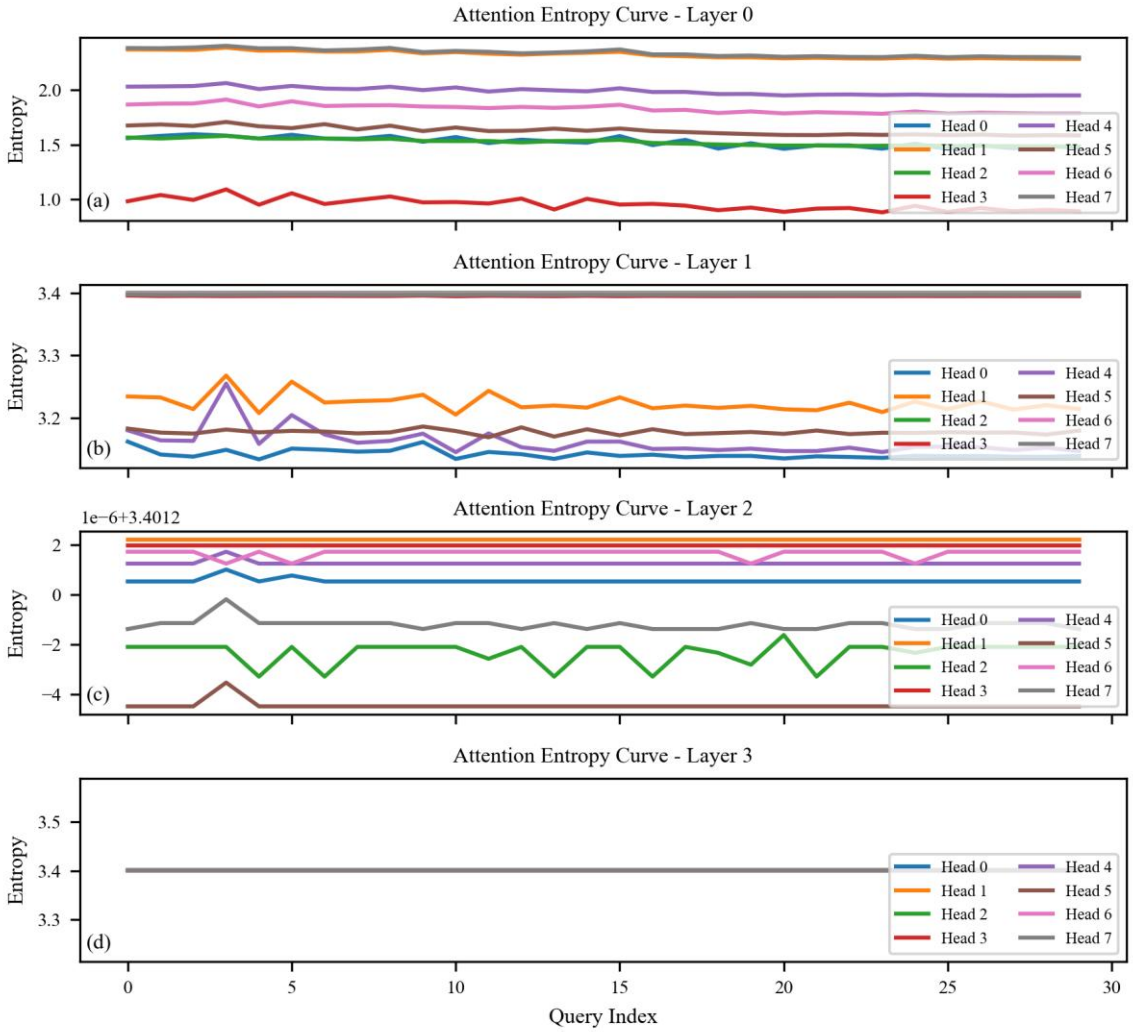


Fig. 21. Globally averaged multi-head attention entropy curves at the seismic trace near Well A for different Transformer layers.

Figure 21 illustrates the global mean attention entropy curves for multiple heads across different Transfomer layers in the BPI-ViT model, applied to real seismic field data. In the shallow layers (layers 0 and 1), the attention entropy for each head shows considerable fluctuations with respect to the query index, and significant differences are observed among the heads. This indicates that the model exhibits strong local feature focusing and spatial selectivity in the shallow layers. As the network depth increases (layer 2), the overall level of attention entropy rises and the inter-head differences are notably reduced, reflecting the model's progressive integration of larger-scale spatial information and feature fusion. In the deepest layer (layer 3), the entropy curves for all heads nearly overlap and form a flat, constant line, signifying highly uniform and convergent attention distributions, where the contributions from all spatial locations become nearly equal. This demonstrates that the model has achieved comprehensive integration of the global spatial structure in the seismic data, providing stable and unified feature representations for inversion.

In contrast, Figure 7, which is based on Marmousi2 model data, displays the same hierarchical evolution of attention entropy within the BPI-ViT model. The shallow layers (layers 0 and 1) also show pronounced inter-head differences and spatial fluctuations, indicating that the attention mechanism in shallow layers effectively captures spatial details in the context of the physical synthetic model. However, in the deeper layers (layers 2 and 3), although the attention entropy for all heads increases and approaches the theoretical maximum—signaling the fusion of spatial semantics—subtle differences and slight fluctuations persist among the head curves, and complete convergence as seen with real data is not fully achieved. This suggests that, for the Marmousi2 synthetic data, the model's global information representation in deep layers is enhanced but still retains some structural selectivity, with spatial consistency slightly lower than that observed in real seismic data.

Comparative analysis indicates that the BPI-ViT model exhibits a consistent trend of attention evolution from shallow local focus to deep global integration in both real seismic field data and Marmousi2 synthetic model data. However, with real field data, the deep-layer attention distributions are more consistent and convergent, indicating that, in complex real geological environments, BPI-ViT can more fully integrate spatial information and achieve highly unified feature representations. In contrast, with the Marmousi2 data, minor structural differences persist in the deep layers, reflecting the model's adaptability and sensitivity to spatial structures. This comparison not only reveals differences in the spatial modeling mechanisms of the model for different data types, but also provides a theoretical basis for future structural perception and inversion reliability analysis in complex seismic datasets.

Figure 22 presents a comparative analysis of the residual distributions between re-synthesized seismic records (obtained via inversion by PINN, BPINN, and BPI-ViT) and field seismic data. Ellipses denote the target CO_2 reservoir for monitoring, while arrows indicate the principal migration pathways of injected CO_2, highlighting critical geological bodies and fluid transport pathways in the injection–production region.

Panels (a–c) illustrate the residuals for the near-, middle-, and far-angle gathers generated by PINN; panels (d–f) correspond to BPINN; and panels (g–i) to BPI-ViT. Across all three methods, the main faults, stratigraphic interfaces, and target reservoir (as indicated by ellipses) display low residuals, evidencing a high degree of consistency between inversion results and field observations in these critical zones. However, the distribution of residuals near the injection–production channels and within structurally complex bodies reveals marked differences among the methods.

Specifically, PINN exhibits more scattered residuals in the injection–production region and adjacent anomalous zones, with some areas displaying pronounced mismatches. This indicates limited capacity in fitting fluid-driven perturbations and structural anomalies. In contrast, BPINN, through Bayesian uncertainty modeling, achieves a noticeable reduction in residual amplitude within principal fault zones and target intervals, reflecting enhanced model stability and robustness to local anomalies. Most notably, BPI-ViT yields the most uniform and minimized residual distribution within both the target reservoir and the main CO_2 migration channel, with pronounced spatial continuity and anomaly suppression. This demonstrates superior adaptability and fidelity in reconstructing complex structures and fluid migration anomalies.

Cross-referencing with Supplementary Figure A2, which compares the forward-modeled gathers for each method, further supports the rationality of these residual characteristics. Notably, the synthetic gathers from BPI-ViT best match the field data within the target and fluid channel regions, followed by BPINN, while PINN exhibits minor mismatches in certain thin layers and anomalous blocks. These results collectively indicate that BPI-ViT outperforms in both inversion accuracy and anomaly identification, and that its structure-aware feature modeling delivers seismic responses most consistent with the observed data.

In summary, all three neural network frameworks achieve effective recovery of the seismic response of the principal geological bodies. However, BPI-ViT exhibits superior performance in terms of inversion accuracy, characterization of anomalous structures, and residual convergence within both the target interval and CO_2 migration pathways. This highlights its potential for high-resolution seismic inversion and reservoir management in CO_2 monitoring projects.

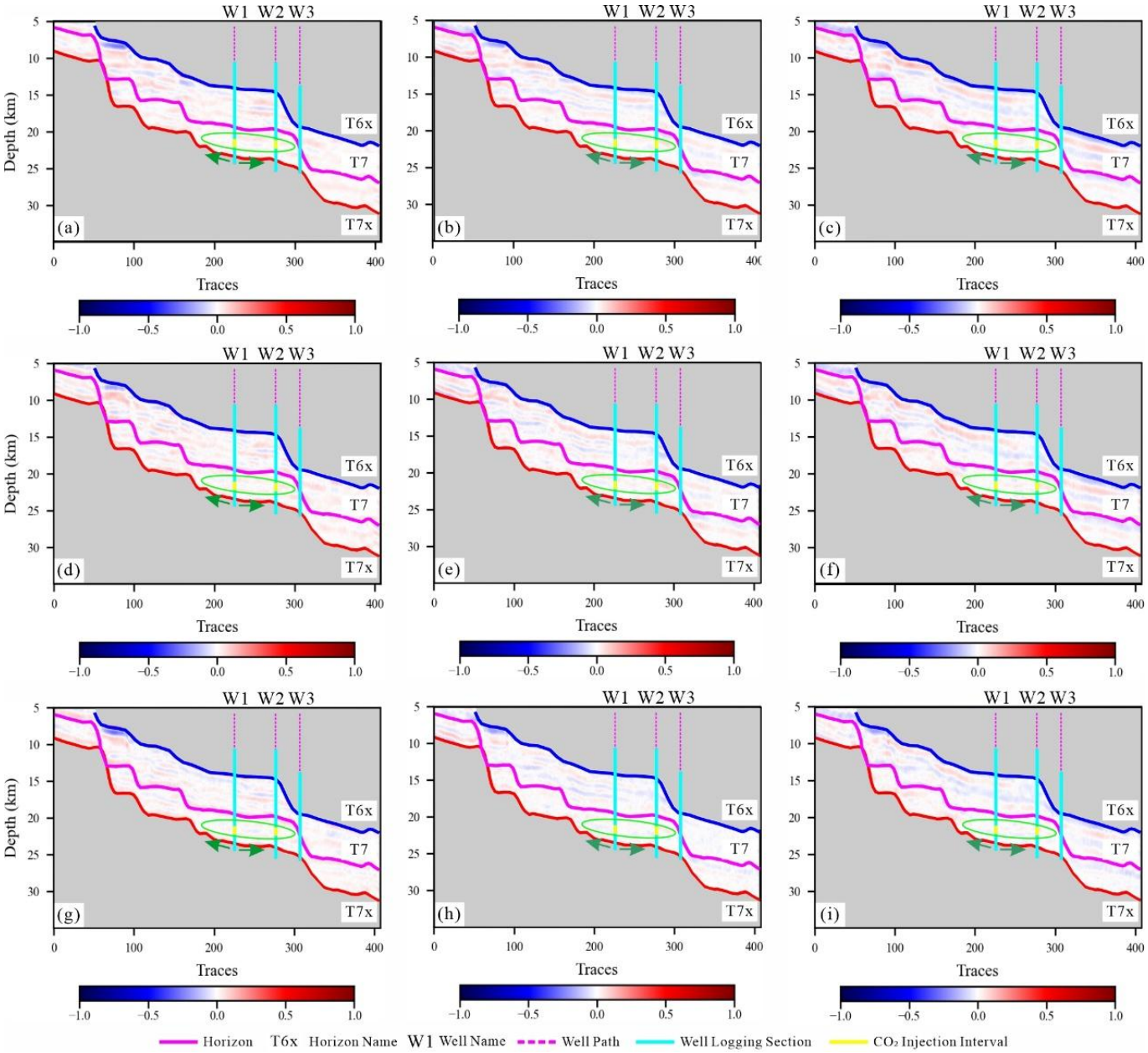


Fig. 22. Comparison of residuals between re-synthesized records by three neural network methods and field seismic data. (a)–(c) Residuals for the near-, middle-, and far-angle gathers by PINN; (d)–(f) residuals by BPINN; (g)–(i) residuals by BPI-ViT.

V. Discussion

A. Physics-Driven Generalization: A Cognitive Leap

Conventional seismic inversion typically focuses on minimizing pixel-level errors (e.g., RMSE, MAE) and often struggles to maintain spatial structural consistency and inversion stability in complex geology. To address this, the proposed BPI-ViT introduces structure-level modeling via attention maps, standard-deviation (std) maps, and multi-scale feature representations, enabling hierarchical modeling of geological structures, spatial coherence, and contextual information. This pixel–structure synergy substantially improves generalization and interpretability, laying a foundation for structure-aware, collaborative inversion.

Unlike data-hungry deep networks, the generalization of Physics-Informed Neural Networks (PINNs) fundamentally depends on the embedded physics. Training "data" in PINNs are samples from the domain of the governing equations used to evaluate physical residuals—not labeled targets. By minimizing these residuals, the network actively fits the underlying physics and achieves end-to-end physical inference, enabling efficient inversion in unsupervised—even unlabeled—settings. This reduces reliance on high-quality observations and allows continual, self-consistent re-optimization when facing new scenes or updated constraints, yielding strong transferability and sustainable optimization.

Crucially, PINN generalization is determined by the applicability and universality of the chosen physical laws (e.g., Zoeppritz equations) rather than by the diversity of labeled

data. As long as the embedded physics remains valid, the model can migrate across scenarios, realizing a genuinely physics-driven, data-agnostic generalization that is difficult for traditional approaches to match in complex or previously unseen settings.

PINNs also exhibit a training–inference duality: each parameter update is simultaneously training and approximate reasoning toward the physical solution. In ideal conditions, a trained PINN functions as a high-quality neural surrogate of the governing equations, efficiently producing solutions under specified constraints. Practically, this is further accelerated in two ways: (i) automatic differentiation and GPU parallelism substantially reduce per-iteration costs for residual computation and optimization; and (ii) thanks to the physics constraints, approximately 30% of physical points often suffice to reach target accuracy, lowering total computation and wall-time. Together, these properties yield speed and resource advantages for large-scale seismic inversion.

In summary, by coupling physics-driven generalization with pixel–structure integration, the PINN/BPI-ViT family provides a robust theoretical and methodological basis for seismic inversion and multi-physics coupling in complex geological environments.

B. Analysis of Methodological Scalability

Beyond strong performance in pre-stack AVO inversion, BPI-ViT exhibits notable structural scalability. Its modular Transformer design adapts to varying data dimensions, resolutions, and multi-physics inputs, facilitating the integration of diverse physical priors (e.g., Zoeppritz, elastic wave equations) and supporting multi-task, multi-source fusion. The Bayesian mechanism and structure-aware attention enhance uncertainty quantification and reliability expression, and synergize with tasks such as active learning and structural identification. Collectively, BPI-ViT offers theoretical and architectural support for large-scale, multi-physics, and cross-task applications.

C. Evolution of Model Intelligence and a Proto "Embodied Cognitive Agent"

Building on advances in accuracy, structural consistency, and uncertainty quantification, we outline a task-oriented hierarchy of model "awareness" for seismic inversion: zero-level (reactive), first-level (self-perception), and second-level (reasoning and behavioral self-evaluation). Here "awareness" is not human-like consciousness; it denotes a model's ability to explicitly represent its behavior, confidence, and structural focus.

Traditional CNN-based PINNs embed physical consistency but have limited structural expression and lack explicit modeling of predictive confidence, placing them between zero- and first-level awareness. In contrast, BPI-ViT augments the PINN backbone with Bayesian uncertainty and multi-scale attention. Std maps and attention maps express uncertainty distributions and structural focus; token features can be aligned to geological structures, yielding traceable paths for structural reasoning. BPI-ViT thus attains an incipient form of self-perception of its internal state—a proto first-level awareness for inversion.

Essentially, PINN and BPI-ViT form consecutive stages of one evolutionary sequence: PINN provides a physics-neurostructural prototype; BPI-ViT strengthens structural perception and confidence expression, laying groundwork for higher-order, embodied intelligent inversion systems.

D. Paradigm Evolution: From Capability Building to Structure-Level Trustworthy Collaboration

We summarize four progressive, complementary paradigms spanning capability, evaluation, structure, and ontology:

1.Intelligent Structural Capability Paradigm: From statistical black-box DNNs to logic-driven PINNs and then to logic + reasoning (Bayesian and structure-aware modeling). Through structural modeling (e.g., attention and white-box mechanisms), capability, trustworthiness, and interpretability are unified—shifting inversion from black-box prediction to structure-aware, trustworthy modeling.

2.Trustworthiness-Driven Multidimensional Evaluation Paradigm: Single-point accuracy (e.g., RMSE) is insufficient. Composite evaluation incorporating structural consistency, uncertainty distributions, and spatial trends balances error minimization with structural awareness and reliability. BPI-ViT exemplifies this shift.

3.Structure-Level Trustworthy Collaboration Paradigm: Within BPI-ViT, accuracy, structural identification, uncertainty assessment, and interpretability are deeply integrated. Attention, Bayesian uncertainty, and multi-scale features improve decision consistency and reliability in complex geology, emphasizing the unity of structure modeling and error control.

4.Neural Awareness Paradigm—Physics-as-Structure Modeling: Physical laws are embedded in the network rather than added as external losses, achieving deep coupling between neural computation and physical mechanisms. End-to-end training on physical residuals enables active recovery of physical structure while unifying representation, physical consistency, and interpretability. Physical operators, priors, and regularizers share the same autodiff graph as the network, allowing flexible incorporation of customized physics and providing a basis for structure-level Bayesian UQ—toward seismic inversion systems with features of embodied intelligent agents.

Together, these paradigms form a coherent framework for next-generation seismic intelligence characterized by accuracy, stability, and interpretability..

E. Challenges and Future Directions in Interpretability

Although BPI-ViT can output attention maps and multi-layer features that visualize internal mechanisms, Transformer-based models still face limits in interpretability. Attention/feature maps describe focus patterns and correlations; they do not imply physical causality or uniquely correspond to geophysically meaningful structures or properties. Consequently, most current interpretability remains

at the visualization level and has yet to achieve "strong interpretability" through causal inference and tight domain-knowledge integration.

Future work should emphasize physically grounded causal modeling and expert knowledge integration to improve transparency and credibility, advancing AI from "black-box correlation" toward "mechanistic transparency." We advocate for interpretability research that goes beyond visual explanations and moves toward trustworthy AI models that integrate physics constraints with causal reasoning.

VI. Conclusion

Starting from a CNN-based PINN baseline, this work successively developed a Bayesian PINN (BPINN) with uncertainty modeling and a structure-level Bayesian physics-informed Vision Transformer (BPI-ViT) for pre-stack AVO inversion. Together, these components form a coherent methodology for "next-generation seismic AVO inversion," advancing from pixel-level optimization to structure-level synergy, from black-box correlation to structural interpretability, and from physics embedding to an agent-like inference prototype.

Extensive experiments on the Marmousi2 benchmark and on real field data from the Shengli Oilfield CO_2 demonstration area show that BPI-ViT improves structural recovery, anomaly identification, uncertainty quantification, and spatial consistency over PINN and BPINN. Beyond accurately delineating principal horizons, faults, and fluid-related anomalies, the combination of multi-layer self-attention and Bayesian modeling enables structure-level uncertainty analysis and multi-dimensional interpretability, enhancing reliability in complex geological settings.

Conceptually, we propose a structure-level intelligence paradigm that unifies capability, evaluation, and system-level collaboration. By tightly coupling structural perception with physical constraints, BPI-ViT mitigates limitations of pixel-wise interpretability and offers traceable decision pathways linked to governing physics.

Overall, this work presents a prototype of a structure-level, trustworthy multi-parameter inversion system tailored to complex geology and provides a path toward physics-grounded, structure-aware intelligent inversion. Future efforts will extend the framework to multi-physics joint inversion, more capable structure-aware agents, and large-scale applications in real reservoirs.

Limitations and future work: This study emphasizes methodological innovation and structural mechanism validation. A systematic assessment of computational efficiency at ultra-large 3D scales, broader ablation on physics priors, and deeper causal interpretability analyses remain open and are the focus of ongoing work.

APPENDIX A

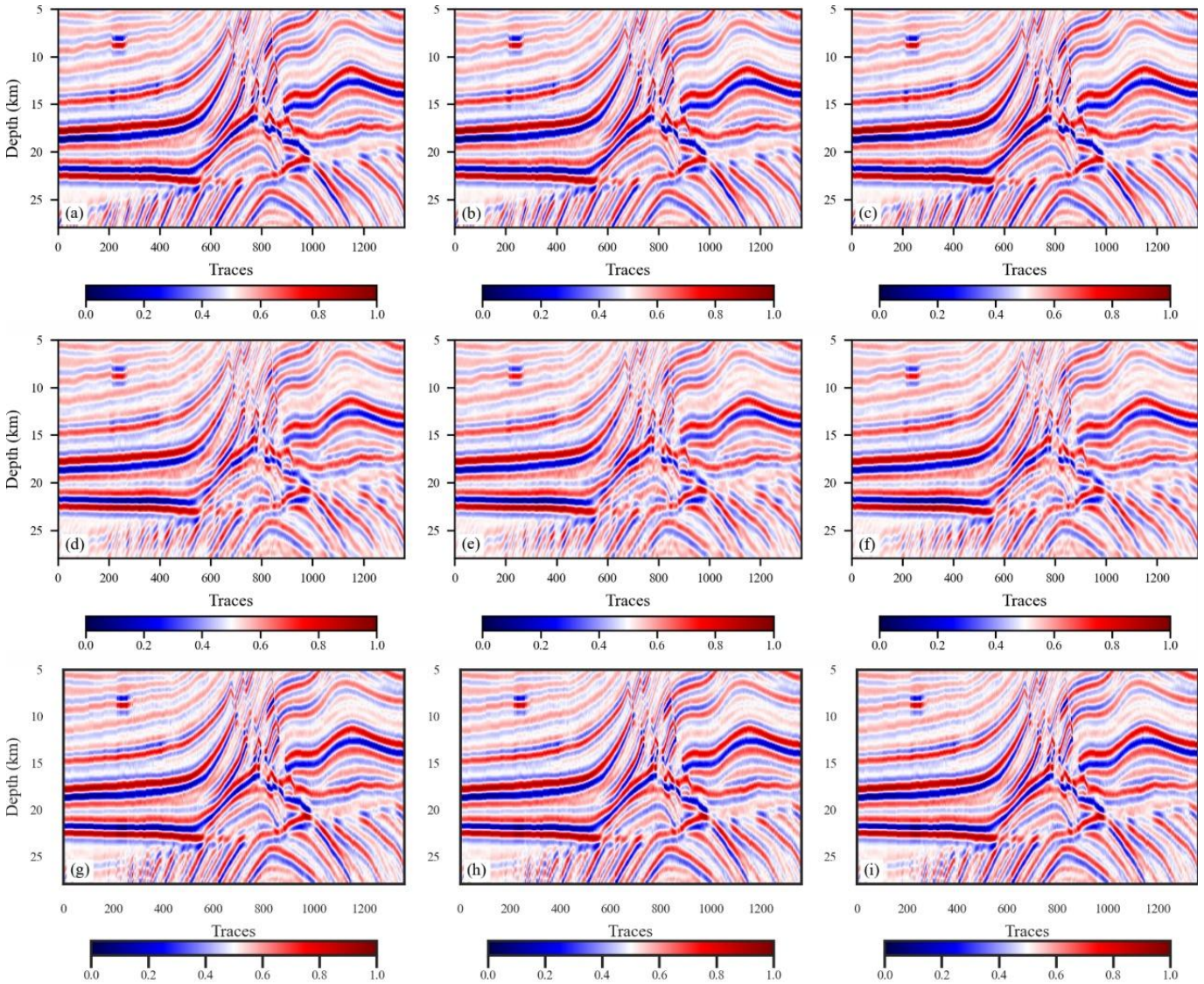


Fig. A1. Synthetic seismic records generated by forward modeling based on inversion results from three methods. (a)–(c) are synthetic records based on the inversion results of P-wave velocity, S-wave velocity, and density from the PINN method, respectively. (d)–(f) correspond to the BPINN-based inversion results. (g)–(i) are derived from the BPI-ViT-based inversion results.

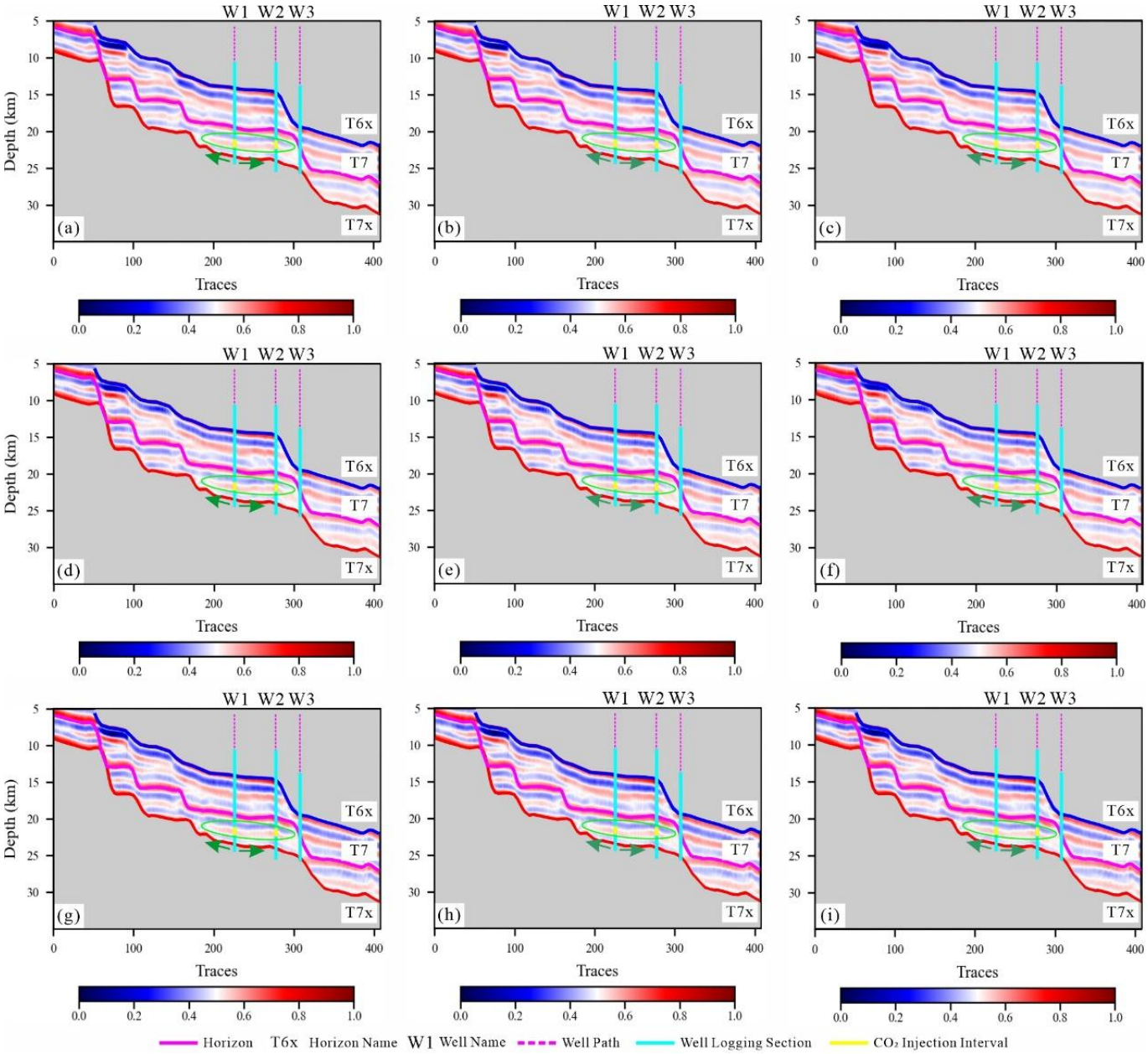


Fig. A2. Comparison of forward-modeled synthetic records from field data inversion results using three neural network methods. (a)–(c) Synthetic records for near-, middle-, and far-angle gathers by PINN; (d)–(f) by BPINN; (g)–(i) by BPI-ViT.

REFERENCES

[1] Aki K, Richards P G. Quantitative Seismology: Theory and Methods[M]. San Francisco: W. H. Freeman, 1980.

[2] Shuey R T. A simplification of the Zoeppritz equations[J]. Geophysics, 1985, 50(4): 609–614.

[3] Buland A, Omre H. Bayesian linearized AVO inversion[J]. Geophysics, 2003, 68(1): 185–198.

[4] Raissi M, Perdikaris P, Karniadakis G E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving PDEs[J]. Journal of Computational Physics, 2019, 378: 686–707.

[5] Karniadakis G E, Kevrekidis I G, Lu L, et al. Physics-informed machine learning[J]. Nature Reviews Physics, 2021, 3(6): 422–440.

[6] Rasht-Behesht, M., Huber, C., Shukla, K., & Karniadakis, G. E. (2022). Physics-informed neural networks (PINNs) for wave propagation and full waveform inversions. Journal of Geophysical Research: Solid Earth, 127, e2021JB023120. https://doi.org/10.1029/2021JB023120.

[7] Bin Waheed, Umair & Haghighat, Ehsan & Alkhalifah, Tariq & Song, Chao & Hao, Qi. (2021). PINNeik: Eikonal solution using physics-informed neural networks. Computers & Geosciences. 155. 104833. 10.1016/j.cageo.2021.104833.

[8] Li P, Grana D, Liu M. Bayesian neural network and Bayesian physics-informed neural network via variational inference for seismic petrophysical inversion[J]. Geophysics, 2024, 89(6): M185–M196.

[9] Chen, Y., de Ridder, S. A. L., Rost, S., Guo, Z., Wu, X., & Chen, Y. (2022). Eikonal tomography with physics-informed neural networks: Rayleigh wave phase velocity in the northeastern margin of the Tibetan Plateau. Geophysical Research Letters, 49, e2022GL099053. https://doi.org/10.1029/2022GL099053

[10] Chen, Y., de Ridder, S. A. L., Rost, S., Guo, Z., Wu, X., Li, S., & Chen, Y. (2023). Physics-informed neural networks for elliptical-anisotropy eikonal tomography: Application to data from the northeastern Tibetan Plateau. Journal of Geophysical Research: Solid Earth, 128, e2023JB027378. https://doi.org/10.1029/2023JB027378

[11] Reetam Biswas, Mrinal K. Sen, Vishal Das, and Tapan Mukerji, (2019), "Prestack and poststack inversion using a physics-guided

convolutional neural network," Interpretation 7: SE161-SE174.https://doi.org/10.1190/INT-2018-0236.1

[12] .Wang, S., Liu, C., Song, C. et al. Prestack AVO Inversion Based On Physics-constrained Deep Learning Method. Appl. Geophys. (2025). https://doi.org/10.1007/s11770-025-1179-y

[13] Wen, Yeming, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. 2018. "Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches." Proceedings of the International Conference on Learning Representations.

[14] Yang, Liu, Xuhui Meng, and George Em Karniadakis. 2021. "B-PINNs: Bayesian Physics-Informed Neural Networks for Forward and Inverse PDE Problems with Noisy Data." Journal of Computational Physics 425 (January): Article 109913. https://doi.org/10.1016/j.jcp.2020.109913.

[15] Peng Li, Dario Grana, Mingliang Liu; Bayesian neural network and Bayesian physics-informed neural network via variational inference for seismic petrophysical inversion. Geophysics 2024;; 89 (6): M185–M196. doi: https://doi.org/10.1190/geo2023-0737.1

[16] R. Gou, Y. Zhang, X. Zhu and J. Gao, "Bayesian Physics-Informed Neural Networks for the Subsurface Tomography Based on the Eikonal Equation," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-12, 2023, Art no. 4503012, doi: 10.1109/TGRS.2023.3286438.

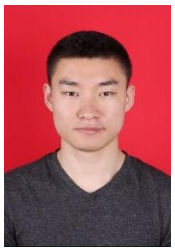
[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017.

[18] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. 2021.

[19] Hongzhou Wang, Jun Lin, Xintong Dong, Shaoping Lu, Yue Li, and Baojun Yang, (2023), "Seismic velocity inversion transformer," GEOPHYSICS 88: R513-R533. https://doi.org/10.1190/geo2022-0283.1

[20] Dou, Yimin, and Kewen Li. 2024. "3D Seismic Mask Auto Encoder: Seismic Inversion Using Transformer-Based Reconstruction Representation Learning." Computers and Geotechnics 169 (March): Article 106194. https://doi.org/10.1016/j.compgeo.2024.106194.

[21] Gerard T. Schuster, Yuqing Chen, and Shihang Feng, (2024), "Review of physics-informed machine-learning inversion of geophysical data," GEOPHYSICS 89: T337-T356. https://doi.org/10.1190/geo2023-0615.1



Z. Liu was born in 1995. He received his Bachelor's degree in Exploration Technology and Engineering from Southwest Petroleum University, China, in 2018. He is pursuing a Ph.D. in Geological Resources and Geological Engineering at China University of Petroleum (East China), Qingdao, since September 2020. From April 2023 to January 2024, he was funded by the European Union (EU) Erasmus+ program for academic exchange at Transilvania University of Brașov, Romania, and in 2025 was awarded the SEG Travel Grant by the Society of Exploration Geophysicists.

Mr. Liu's research focuses on the theoretical development and practical application of artificial intelligence and Physics-Informed Neural Networks (PINNs), as well as quantum computing and quantum intelligence technologies, in oil and gas exploration and geophysical reservoir interpretation. His main research interests include geophysical modeling methods that integrate data-driven approaches with physical constraints, intelligent inversion techniques for reservoir prediction, trustworthy artificial intelligence systems with reasoning capabilities and uncertainty quantification, and pioneering international applications of quantum intelligence in geophysical inversion and hydrological forecasting. In complex geological environments, Mr. Liu is committed to developing multi-source integrated inversion frameworks that balance accuracy, interpretability, and robustness, aiming to advance intelligent sensing, cognitive decision-making, and frontier quantum technologies in geophysical exploration.