PHYSICS-SUPERVISED CONVERGENT BORN NEURAL OPERATORS FOR FULL WAVEFORM MODELING AND INVERSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Full Waveform Inversion (FWI) is vital for subsurface resource exploration and geological hazard monitoring, but remains computationally prohibitive. Deep learning-based acceleration faces three core obstacles: scarce supervised labels, unstable physics-constrained training, and difficulty in enforcing accurate open-boundary modeling under broadband settings. We propose Physics-supervised Convergent Born Neural Operators (CBNO), a framework whose central mechanism learns wavefield mappings without paired wavefield labels by minimizing a Convergent Born Series (CBS) fixed-point discrepancy. CBS updates with Perfectly Matched Layers (PML) provide the physical training signal, avoiding the need for precomputed target wavefields. To support broadband prediction, frequency-partitioned Axial-AFNO experts combine shared low-rank projections with axis- and wavenumber-aware gates. The trained surrogate is then frozen and used in differentiable inversion, where upgoing-wave separation emphasizes weak reflection signals. Empirically, CBNO achieves highly accurate forward modeling on OpenFWI with a $10\times$ speedup over the CBS reference solver, enables zero-shot transfer of the frozen forward surrogate to inversion on 2D Marmousi and Overthrust benchmarks, and demonstrates extensibility through separately trained 3D forward models.

1 INTRODUCTION

Full-waveform inversion (FWI) is a central tool for high-resolution subsurface imaging, with broad applications in exploration geophysics, hazard monitoring, and global-scale Earth structure studies (Yuan & Romanowicz, 2017; Tape et al., 2009; French & Romanowicz, 2015; Zhu & Tromp, 2013; Guo et al., 2022; Cobden et al., 2024; Virieux & Operto, 2009). FWI estimates subsurface parameters by iteratively minimizing the misfit between observed and simulated seismic data, which requires repeated forward and adjoint wave-equation solves over many sources and frequencies (Wang et al., 2021b; Hoffmann et al., 2024; Tarantola, 1984b;a; Fichtner et al., 2006; Métivier et al., 2014). Despite sustained progress in algorithms and high-performance computing, this repeated PDE-solving loop remains a major computational bottleneck in practical multi-scale workflows (Hoffmann et al., 2024).

Neural surrogates and operator learning (Mousavi & Beroza, 2022; Guibas et al., 2022; Li et al., 2023; 2024; Lu et al., 2022; Mousavi & Beroza, 2023; Kovachki et al., 2023; Lu et al., 2021; Wang & Wang, 2024; Li et al., 2025) offer a promising route to accelerating FWI by learning maps from medium properties to wavefields or seismic observations and reusing them inside differentiable inversion. Incorporating geological and geophysical priors into such models can further improve robustness, interpretability, and physical consistency (Wu et al., 2023). However, three obstacles limit their deployment in realistic broadband and open-boundary settings.

First, paired supervision is expensive. Generating large-scale training data requires many high-fidelity simulations, and the cost becomes prohibitive for 3D or large-domain problems (Li et al., 2025; Schuster et al., 2024; Sun et al., 2023). Second, residual-based physics constraints, such as PINN-style objectives, are often brittle for high-frequency wave propagation because local differential residuals can suffer from optimization pathologies and imbalanced loss terms (Krishnapriyan et al., 2021; Wang et al., 2021a; 2022). Third, accurate open-boundary modeling is essential. Perfectly matched

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

layers (PML) suppress artificial boundary reflections in wave simulations (Berenger, 1994; Liu & Tao, 1997; Hesthaven, 1998; Bécache et al., 2006; Bermúdez et al., 2007; Zhang et al., 2025). Their effectiveness depends on appropriate formulation and discretization (Johnson, 2021).

To address these challenges, we propose **Physics-supervised Convergent Born Neural Operators (CBNO)**, a self-supervised neural-operator framework for full-waveform modeling and inversion (Fig. 1). CBS fixed-point consistency provides the supervision signal for operator training. During training, we unroll a fixed number of CBS updates from the predicted wavefield and minimize the discrepancy between the prediction and the final CBS-updated wavefield. Here, label-free refers specifically to the absence of paired wavefield targets: training still uses velocity models and source configurations, and inversion requires observed data.

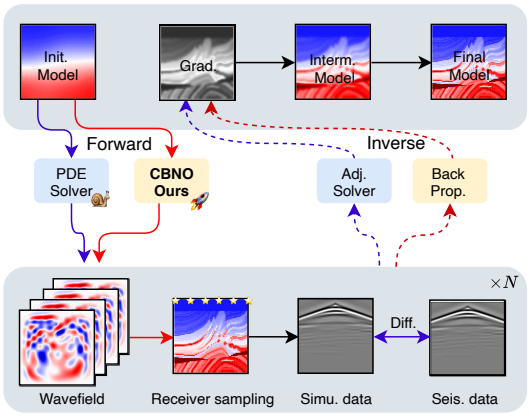

Figure 1: Overview of the proposed AI-accelerated FWI framework versus traditional FWI. **Red paths** denote forward modeling, while **blue paths** denote inversion/model update. **Dashed lines** represent traditional FWI bottlenecked by slow numerical PDE solvers *(slow)*, whereas **solid lines** represent our CBNO-based framework with a fast neural-operator surrogate *(fast)*. Here, *data* denotes observations.

To support this training framework, CBNO uses an **Axial-factorized Adaptive Fourier Neural Operator (Axial-AFNO)** combining shared low-rank projections, directional spectral gates, and source-frequency experts for compact broadband prediction. For downstream inversion, we keep the trained surrogate frozen and separate upgoing wavefields during the online inversion loop. This reflection-aware objective emphasizes weak reflection energy that can be overwhelmed by direct arrivals. Thus, CBS supplies the training signal, the neural architecture parameterizes the forward map, and wavefield separation defines the downstream inversion objective without changing the offline training loss.

Our contributions are summarized as follows:

- **CBS-based label-free operator training.** Our primary methodological contribution is a wavefield-training framework based on CBS fixed-point consistency with PML-aware boundary treatment, learning the forward operator without paired wavefield labels.
- **Broadband neural parameterization.** We combine cross-axis shared low-rank projections, axis- and wavenumber-dependent gates, and source-frequency experts to build a compact representation tailored to broadband seismic wavefields.
- **Reflection-aware surrogate-based inversion.** During online inversion, upgoing-wave decomposition focuses the misfit on reflection-dominated components to enhance subsurface updates without changing the offline training objective.

Experiments assess these designs through forward modeling on OpenFWI, zero-shot inversion on 2D Marmousi and Overthrust benchmarks, and separate 3D pre-training, including a 10$\times$ forward-inference speedup over the CBS reference solver. Related approaches use wavefield-supervised pretraining with physics-constrained inversion (Huang et al., 2025) or CBS-inspired architectures trained on simulated wavefields (Zeng et al., 2025). CBNO instead uses CBS consistency to train the forward operator without paired wavefield targets. Appendix A details these distinctions.

2 METHOD

2.1 PROBLEM STATEMENT

We consider frequency-domain acoustic wave propagation on $\Omega \subset \mathbb{R}^d$ ($d \in \{2, 3\}$). Let $c(\mathbf{x})$, $\hat{u}(\mathbf{x}, \omega)$, and $s(\mathbf{x}, \omega)$ denote the velocity, wavefield, and source, respectively. The forward physics is governed by the Helmholtz equation with Perfectly Matched Layers (PML):

$$\begin{cases} -\omega^2 \hat{u} - c(\mathbf{x})^2 \nabla^2 \hat{u} = s(\mathbf{x}, \omega), \\ \text{s.t. } \mathcal{B}_{\mathrm{PML}}(\hat{u}) = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{1}$$

The inversion is formulated as a two-stage optimization. First, defining the Helmholtz operator $\mathcal{H}(c, \omega) := -\omega^2 - c^2 \nabla^2$, we learn a neural surrogate $\mathcal{G}_\theta : (c, s, \omega) \mapsto \hat{u}$ by minimizing a label-free physics loss $\mathcal{L}_{\mathrm{phy}}$ derived from the Convergent Born Series (CBS):

$$\theta^* = \arg\min_\theta \mathbb{E}_{c,s,\omega} \Big[\mathcal{L}_{\mathrm{phy}} \big(\mathcal{G}_\theta(c, s, \omega); \mathcal{H}(c, \omega) \big) \Big]. \tag{2}$$

Second, with the surrogate $\mathcal{G}_{\theta^*}$ frozen, we reconstruct $\hat{c}$ by minimizing the total correlation-based objective $\mathcal{J}_{\mathrm{corr}}$ between predicted and observed receiver records, where $\mathcal{P}_i$ samples receivers. The constituent loss $\mathcal{L}_{\mathrm{corr}}$ is defined in Appendix D.2:

$$\hat{c} = \arg\min_c \mathcal{J}_{\mathrm{corr}} \big(\{\mathcal{P}_i \mathcal{G}_{\theta^*}(c, s_i, \omega)\}_{i,\omega}, \{d_i^{\mathrm{obs}}(\omega)\}_{i,\omega} \big) . \tag{3}$$

2.2 MAIN STRUCTURE OF CBNO

As illustrated in Fig. 2, the CBNO framework bridges rigorous physical principles with efficient neural parameterization through a unified closed-loop workflow. The pipeline begins with a *Nondimensionalized Input Formulation* (Sec. 2.3), which combines Helmholtz-based nondimensionalization with online input sampling. This is followed by a *Parameter-Efficient Neural Prediction* module (Sec. 2.4) featuring an Axial-Factorized AFNO backbone with cross-axis shared low-rank bases and axis/wavenumber gates for compact spectral mixing, complemented by source-frequency experts for broadband modeling. Training is anchored by *CBS-Guided Physics Supervision* (Sec. 2.5) via a label-free integral loss that penalizes inconsistency with the source-driven CBS update. Finally, *Reflection-Aware Inversion* (Sec. 2.6) isolates upgoing reflection energy to improve gradient sensitivity to deep structures.

2.3 NONDIMENSIONALIZED INPUT FORMULATION

To decouple physical scales (c, ω, s), we employ a nondimensionalization process. Letting h be the grid spacing, we define dimensionless velocity $c^* = c/(\omega h)$ and wavefield $u^* = \hat{u}\omega^2 / s_{\mathrm{scale}}$. Substituting these into Eq. (1) yields the unified scale-invariant equation:

$$u^* + (c^*)^2 \tilde{\nabla}^2 u^* + f^* = 0, \tag{4}$$

where $\tilde{\nabla}^2$ is the normalized Laplacian and f^* is the source mask. This transformation removes explicit dependence on ω and amplitude, simplifying the learning task to $(c^*, f^*) \mapsto u^*$. We use *online sampling and augmentation* of the inputs constituting c^* and f^* without requiring precomputed wavefield targets:

- **Continuous Frequency Sampling:** We sample $\omega \sim \mathcal{U}[\omega_{\min}, \omega_{\max}]$. Since $c^* \propto 1/\omega$, this implicitly augments the effective velocity distribution, exposing the model to varying wavelength-to-grid ratios.
- **Spatial Perturbations:** We apply random shifts to the source mask f^* and structural transforms (flip, blur, swap) to the velocity c to diversify the training inputs.

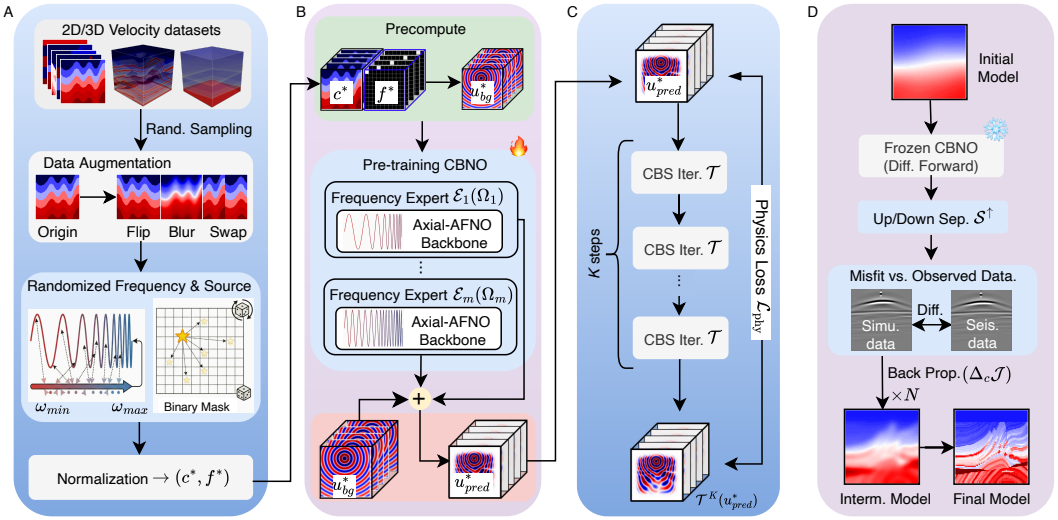

Figure 2: **Architecture of the CBNO framework. (a) Nondimensionalized Input Formulation:** Velocity models undergo sampling and augmentation, then are paired with stochastic frequency ω and source location x_s to produce normalized dimensionless inputs (c^*, f^*). **(b) Parameter-Efficient Neural Prediction:** After precomputing the background u^*_{bg}, frequency-partitioned experts perform active pre-training *(active)* to predict the scattered wavefield. At each frequency, one expert is selected, and its predicted scattered field is added to the background field. **(c) CBS-Guided Physics Supervision:** The label-free physics loss $\mathcal{L}_{\text{phy}}$ is computed by unrolling the CBS iteration for K steps. **(d) Reflection-Aware Inversion:** With the surrogate frozen, velocity is updated through the reflection-aware operator $\mathcal{Q}^\uparrow$.

2.4 Parameter-Efficient Neural Prediction

To enable high-fidelity broadband modeling, we integrate a physics-informed residual scheme with a frequency-partitioned expert ensemble. First, to stabilize learning near source singularities, we analytically compute a dimensionless background field u^*_{bg} (see Appendix B.4). The network inputs $X = \text{Concat}(c^*, f^*, u^*_{\text{bg}}/\|u^*_{\text{bg}}\|_\infty)$ are mapped to a scattered residual, recovering the full field via $u^*_{\text{pred}} = \mathcal{N}(X) + u^*_{\text{bg}}$.

Frequency-Partitioned Experts. To mitigate spectral bias, the target spectrum is divided into M bands $\{\Omega_m\}$ that overlap only at shared endpoints. A single-valued router $r(\omega)$ selects one residual expert whose band contains ω, giving $u^*_{\text{pred}}(\omega) = \mathcal{E}_{r(\omega)}(X) + u^*_{\text{bg}}$. This single-expert routing also applies at shared endpoints.

Physics-Inspired Axial-AFNO Backbone. Each expert uses the Axial-AFNO backbone in Fig. 3.

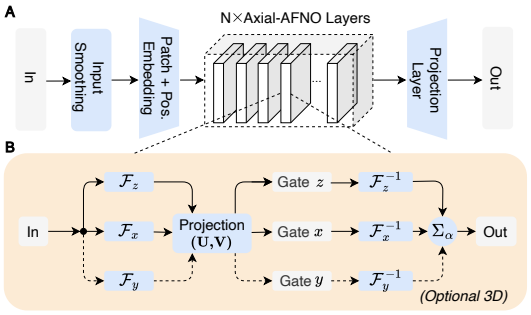

Figure 3: **Detailed architecture of the CBNO expert.** (a) Expert Architecture: The model sequentially processes inputs via Input Smoothing, Patch and Positional Embedding, a stack of N Axial-AFNO layers, and a final Layer Projection. (b) Axial-AFNO Backbone: Spectral processing is factorized into parallel branches. Fourier-transformed inputs ($\mathcal{F}$) are modulated by shared low-rank projections ($\mathbf{U}, \mathbf{V}$) and axis/wavenumber-band gates. Inverse Fourier transforms ($\mathcal{F}^{-1}$) reconstruct the branch features before fusion.

To encode directional and scale-dependent wavefield structure, we introduce a *Shared-Basis Low-Rank Factorization* mechanism. DPOT uses grouped Fourier MLPs with weights shared across modes (Hao et al., 2024). Our mixer uses a low-rank channel subspace shared across spatial axes, with axis- and wavenumber-band-specific gates. This provides a compact directional parameterization within FFT-based operator learning while retaining the asymptotic FFT order. Source-frequency experts specialize broadband prediction; CBS supervision encourages physical consistency.

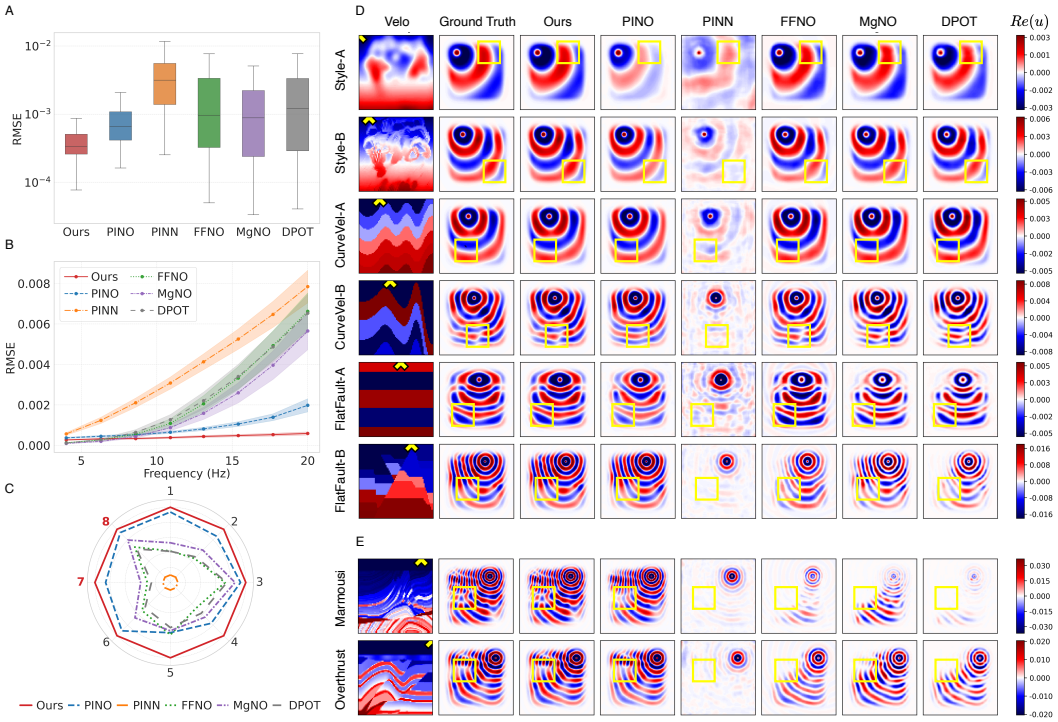

Figure 4: **Inference results of different baselines. (a)** Global RMSE distributions (log scale). **(b)** Mean RMSE vs. source frequency (4–20 Hz); shading indicates ± 1.96 standard errors of the mean over pooled velocity-model–source pairs. **(c)** Radar chart of min-max normalized scores across 8 datasets (Indices 7–8 in red are OOD). **(d–e)** Wavefield snapshots for ID (d) and OOD (e) datasets. Each row shows a unique velocity model, source position ($x \in [0, 690]$ m), and frequency ($f \in [4, 20]$ Hz). Yellow boxes highlight magnified regions. Wavefields depict the real part $\mathrm{Re}(u^*)$ using corner-cut visualization.

2.5 CBS-Guided Physics Supervision

To enable label-free supervision, we use the damped scattering potential $V^* = (c^*)^{-2} - (c_0^*)^{-2} - i\epsilon^*$, where $\epsilon^* > 0$ is the dimensionless damping. The preconditioned CBS update is:

$$\mathcal{T}(u^*) = u^* + \gamma \left[G_0^* \left(V^* u^* + (c^*)^{-2} f^* \right) - u^* \right], \tag{5}$$

Here G_0^* is the Green's operator for the constant damped background $(c_0^*)^{-2} + i\epsilon^*$, with PML as in Appendix B.2. The pointwise preconditioner $\gamma(\mathbf{x}) = iV^*(\mathbf{x})/\epsilon^*$ is fixed by the medium and damping, not learned; the damping bound is given in Appendix B.1. A physically valid wavefield satisfies $\mathcal{T}(u^*) = u^*$.

The network predicts an approximate wavefield u^*_{pred}. We unroll $\mathcal{T}$ for K steps (e.g., $K = 5$) and compare only the final iterate with the prediction. Before spatial and precomputed-field normalization, the objective is:

$$\mathcal{L}_{\text{phy}} = \frac{1}{K} \left\| u^*_{\text{pred}} - \mathcal{T}^K(u^*_{\text{pred}}) \right\|_2^2. \tag{6}$$

The factor $1/K$ scales the final-iterate discrepancy, not intermediate losses. Appendix B.3 establishes the physical wavefield as its unique zero under stated discrete-operator assumptions and provides a numerical check. Training requires no paired wavefield labels; inference uses a single network pass without CBS updates.

2.6 Reflection-Aware Velocity Inversion

To mitigate the dominance of direct waves that mask weak deep reflections, we deploy the frozen network as a differentiable surrogate equipped with a reflection-enhancing separation operator. Given a wavefield u, we apply the Fourier transform $\mathcal{F}_z$ along the depth axis and retain the negative

vertical-wavenumber components:

$$u^{\uparrow} = \mathcal{S}^{\uparrow}(u) = \mathcal{F}_z^{-1}\left(H(-k_z) \cdot \mathcal{F}_z(u)\right), \tag{7}$$

where $H(-k_z)$ is the Heaviside step function selecting upgoing components. We then define the reflection-enhanced measurement operator as

$$\mathcal{Q}^{\uparrow} = \mathcal{M}\mathcal{S}^{\uparrow}, \tag{8}$$

where $\mathcal{M}$ samples receivers. We optimize c using the total correlation-based objective applied to reflection-enhanced receiver records:

$$\mathcal{J}(c) = \mathcal{J}_{\text{corr}}\left(\left\{\mathcal{Q}^{\uparrow}\left(u^{*}_{\text{pred}}(c; s, \omega)\right)\right\}_{s,\omega}, \left\{d^{\uparrow}_{\text{obs}}(s, \omega)\right\}_{s,\omega}\right). \tag{9}$$

Here, $d^{\uparrow}_{\text{obs}}$ denotes the observed data after reflection-enhancing preprocessing. The preprocessing convention is kept identical for the two inputs to the correlation loss, ensuring that the optimization is performed in a consistent reflection-dominated subspace. Both predicted and observed data are mapped to this common receiver-data domain before evaluating the correlation loss.

By backpropagating through $\mathcal{Q}^{\uparrow}$ and the frozen surrogate via automatic differentiation (Liu et al., 2025), the inversion suppresses direct-wave-dominated components and emphasizes weak upgoing reflection energy. This separation is designed for surface-acquisition settings where vertical propagation direction provides a useful proxy for reflection energy; extensions to cross-well geometries, complex topography, or strongly dipping structures may require more general multidimensional wavefield-decomposition operators. For field observations where full-wavefield decomposition is unavailable, receiver-domain direct-arrival suppression, such as direct-wave muting or subtraction, could provide a practical alternative for emphasizing reflection energy.

2.7 Experimental Setup

This section systematically evaluates CBNO, starting with the experimental setup and forward modeling accuracy in both frequency and time domains. We then demonstrate its downstream utility for Full Waveform Inversion (FWI) on in- and out-of-distribution benchmarks, concluding with a 3D pre-training proof-of-concept and ablation studies.

Datasets and Sampling Strategy. We evaluate our framework on OpenFWI (Deng et al., 2022) (70×70, $dh = 10$ m). To probe data efficiency, we restrict the in-distribution (ID) training set (*Vel* family: CurveVel, Style, FlatFault) to 150 samples per category. The 2D model is trained only on these OpenFWI subsets; neither Marmousi nor 2D Overthrust is used for training. Zero-shot OOD evaluation on these external benchmarks uses the frozen forward operator without fine-tuning and denotes dataset-level transfer, not disjoint geological features. Inversion still optimizes the velocity model against observed data. Simulations cover 4–20 Hz with surface sources ($z = 0$). While supervised baselines use fixed discrete sampling ($df = 0.167$ Hz, 5 fixed sources), our self-supervised approach combines online velocity augmentation with uniform sampling of $f \in [4, 20]$ Hz and $x_s \in [0, 690]$ m, broadening the sampled training inputs.

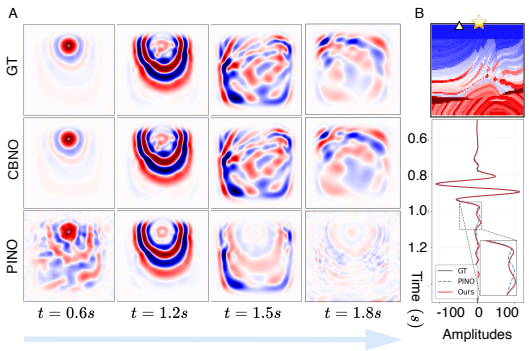

Figure 5: **Comparison on Marmousi. (a)** Wavefield snapshots at $t = 0.6, 1.2, 1.5, 1.8$ s for Ground Truth (top), CBNO (middle), and PINO (bottom). **(b)** Top: Velocity model showing source (⋆) and receiver (△) locations. Bottom: Seismogram traces at the receiver with a zoomed inset (0.9-1.1s).

Baselines and Metrics. We validate CBNO against a comprehensive set of methods detailed in Appendix D.1. These include: (1) Numerical Solvers: The Convergent Born Series (CBS) (Osnabrugge et al., 2016) serves as the high-precision reference solver. (2) Unsupervised Physics-Driven Models: We compare against PINN (Raissi et al., 2019) (with simplified absorbing boundary conditions) and a physics-enhanced PINO (Li et al., 2024) integrated with a one-step CBS loss. (3) Supervised

Operator Learners: We benchmark against FFNO (Tran et al., 2023), MgNO (He et al., 2024), and DPOT (Hao et al., 2024), where DPOT provides a Fourier-attention-based operator baseline alongside our axial low-rank spectral design. For downstream FWI, we compare against both physics-based AD-FWI with CBS solver and data-driven baselines including InversionNet (Wu & Lin, 2020) and VeloGAN (Zhang & Lin, 2020).

To comprehensively assess both forward modeling (wavefield reconstruction) and inversion (velocity reconstruction) performance, we employ five quantitative metrics. Pixel-wise accuracy is measured by Relative Error (RelError), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), while structural fidelity and waveform phase alignment are evaluated using Pearson Correlation (Corr) and Structural Similarity Index (SSIM). Detailed mathematical formulations and physical interpretations for each metric are provided in Appendix D.3.

Training Details. All models are implemented in PyTorch and trained on a single NVIDIA A100 GPU. CBNO is trained using AdamW with a batch size of 50, an initial learning rate of 1×10^{-3}, and cosine annealing. For our frequency-partitioned strategy, the 4–20 Hz spectrum is decomposed into four equal-width sub-bands ($[4, 8]$, $[8, 12]$, $[12, 16]$, and $[16, 20]$ Hz), overlapping only at 8, 12, and 16 Hz. At shared endpoints, prediction retains a single routed output as described in Sec. 2.4. Each band uses a specialized *lightweight* expert. Parameter efficiency is considered separately from the training objective; both the expert count and the per-expert channel configuration contribute to the overall model size. Comprehensive hyperparameter settings are provided in Appendix D.2.

3 Experiments

3.1 Forward Modeling Performance

We evaluate the forward modeling capability of our surrogate model across distinct dimensions: spectral fidelity, generalization, time-domain reconstruction quality, and computational efficiency.

Frequency Effect. Neural networks can exhibit spectral bias toward learning low-frequency functions (Rahaman et al., 2019). Fig. 4(b) demonstrates that CBNO remains robust across the 4–20 Hz range, maintaining a consistently low RMSE profile, whereas baselines like PINN and MgNO degrade. Fig. 4(a) shows a low median RMSE and a compact interquartile range for CBNO across the pooled evaluation records. Qualitatively, this spectral fidelity translates into sharper wavefield reconstruction. In both in-distribution (ID) and out-of-distribution (OOD) scenarios (Fig. 4(d)–(e)), CBNO preserves high-frequency scattering details around curved interfaces, resolving complex interference patterns that the evaluated baselines fail to capture.

ID and OOD Generalization. Evaluations across eight datasets (Table 1 and Appendix Table 7) assess ID accuracy and transfer to external 2D benchmarks. In ID benchmarks, our model achieves strong numerical fidelity, especially in RMSE, MAE, and RelError. Specifically on the CurveVel-A dataset, CBNO reduces RMSE by *40.00%* and MAE by *33.33%* compared to the strongest baseline. On CurveVel-A, CBNO achieves a lower relative error than MgNO (0.1573 vs. 0.1799), while MgNO achieves a higher SSIM (0.957 vs. 0.924).

Table 1: **Comparison on out-of-distribution (OOD) datasets.** Rel. Change is against strongest baseline; positive values denote improvement.

Data	Model	RMSE↓	MAE↓	RelError↓	Corr.↑	SSIM↑
Marm.	PINN	0.0070	0.0043	0.9697	0.218	0.395
	FFNO	0.0056	0.0033	0.6035	0.758	0.700
	MgNO	0.0050	0.0028	0.5059	0.822	0.791
	DPOT	0.0060	0.0035	0.6597	0.766	0.688
	PINO	0.0018	0.0011	0.2707	0.965	0.896
	Ours	**0.0009**	**0.0005**	**0.1464**	**0.986**	**0.940**
	Rel. Change	50.0%	54.5%	45.9%	2.2%	4.9%
Over.	PINN	0.0032	0.0020	0.8764	0.477	0.412
	FFNO	0.0015	0.0008	0.3118	0.949	0.876
	MgNO	0.0011	0.0005	0.2194	0.968	0.917
	DPOT	0.0016	0.0008	0.3235	0.931	0.890
	PINO	0.0006	0.0004	0.2422	0.970	0.891
	Ours	**0.0004**	**0.0002**	**0.1608**	**0.981**	**0.935**
	Rel. Change	33.3%	50.0%	26.7%	1.1%	2.0%

On the external 2D benchmarks, CBNO also achieves strong forward-modeling accuracy. On the complex *Marmousi* benchmark, CBNO outperforms the physics-informed PINO by *50.00%* in RMSE, achieving a correlation coefficient of *0.986* and an SSIM of *0.940*. Since our PINO baseline also uses one-step CBS supervision, this comparison measures differences between complete configurations, not the isolated benefit of CBS supervision.

Time-Domain Reconstruction and Efficiency. We evaluate time-domain reconstruction quality in Fig. 5, benchmarking against PINO (the strongest evaluated baseline on Marmousi in Table 1). Since CBNO operates in the frequency domain, the time-domain wavefield snapshots and seismograms are synthesized by evaluating the frequency-domain surrogate over uniformly sampled frequencies in the 4–20 Hz range with $df = 0.167$ Hz and applying an inverse Fourier transform with a Ricker source wavelet. In the displayed snapshot at $t = 1.8\,\mathrm{s}$, PINO shows weaker amplitudes in parts of the wavefield (Fig. 5(a)), with missing weak-reflection details also visible in Appendix Fig. 9. CBNO preserves wavefield structure across the displayed snapshots, and the receiver trace in Fig. 5(b) shows agreement with the ground truth.
Forward timings compare CBNO and the evaluated PyTorch CBS implementation on the same NVIDIA A100.

Table 2: **Inversion performance on ID/OOD benchmarks. Bold** indicates the best result and underline indicates the second best.

Method	Data	RMSE↓	MAE↓	RelError↓	Corr.↑	SSIM↑	Time (s)
I. Data-driven							
InversionNet	ID	127.43	95.64	0.0384	0.986	0.733	**<0.1**
	OOD	890.17	717.68	0.2667	0.730	0.310	
VeloGAN	ID	**100.09**	**74.93**	**0.0301**	**0.992**	**0.817**	<0.2
	OOD	984.97	791.57	0.2910	0.713	0.304	
II. Physics-consistent							
CBS	ID	181.49	128.42	0.0560	0.966	0.614	∼720
	OOD	**254.80**	**156.47**	**0.0818**	**0.962**	**0.737**	
CBNO	ID	158.05	115.59	0.0469	0.967	0.561	∼60
	OOD	410.74	288.66	0.1293	0.899	0.515	

CBNO achieves a 10× wall-clock speedup per forward simulation batch while maintaining an average correlation > 0.98. A separate RTX 4090 benchmark compares learned surrogates (Appendix C.3): CBNO achieves lower Marmousi error but higher latency per sample than FFNO, PINO, and DPOT.

3.2 Full Waveform Inversion

We deploy the frozen CBNO as a differentiable forward engine for inversion, benchmarking against AD-FWI with the CBS solver and data-driven baselines. Table 2 reveals a clear trade-off. Data-driven methods, such as InversionNet and VeloGAN, achieve the lowest errors on ID benchmarks, but their performance degrades on the external 2D benchmarks. CBNO achieves better inversion accuracy than these data-driven baselines on the evaluated OOD cases, ranking second only to the numerical CBS reference while being approximately 12× faster than CBS-based AD-FWI. These results support dataset-level transfer of the frozen forward operator within the tested inversion setup. Appendix Fig. 10 visualizes the upgoing and downgoing wavefield components. Appendix D.2 details inversion settings. Additional tests show gradual degradation under receiver dropout and sensitivity to severe observation noise (Appendix E.1).

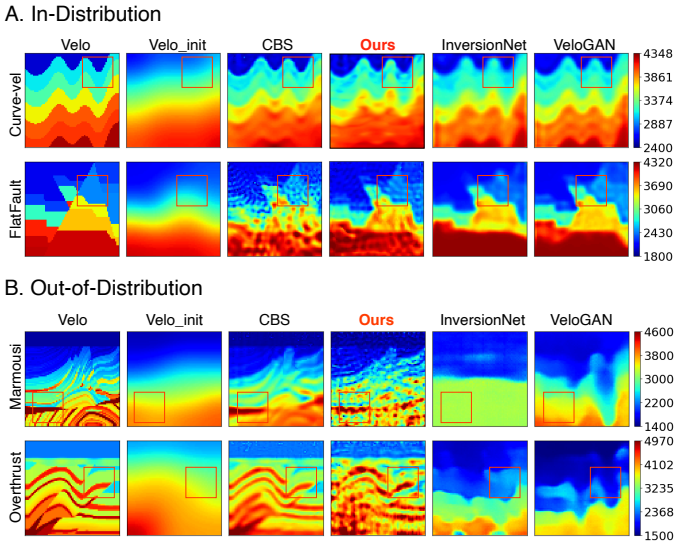

Figure 6: **Comparison of inversion results.** In-distribution (ID): *CurveVel*, *FlatFault*; zero-shot 2D OOD: *Marmousi*, *Overthrust*. Columns (left to right): ground truth (*Velo*), initial model (*Velo_init*), CBS, **CBNO (Ours)**, InversionNet, and VeloGAN.

3.3 Scalability to 3D Modeling

We train a separate CBNO model on a 128^3 grid using 50 OpenFWI 3D volumes and structurally augmented variants of *Overthrust*. Figure 7 evaluates augmented *Overthrust* cases, assessing 3D scalability on variants of a volume included in training; the zero-shot OOD tests use external 2D benchmarks. The Gaussian smoothing standard deviation $\sigma \in [0, 5]$ controls structural complexity, with smaller values giving sharper structures. CBNO achieves relative errors of $0.15 \sim 0.19$ and Pearson correlation > 0.98, versus errors > 0.86 and correlation < 0.52 for PINO in this configuration. CBNO preserves spherical wavefronts and fine-scale scattering, while PINO exhibits phase distortions and degraded interference patterns.

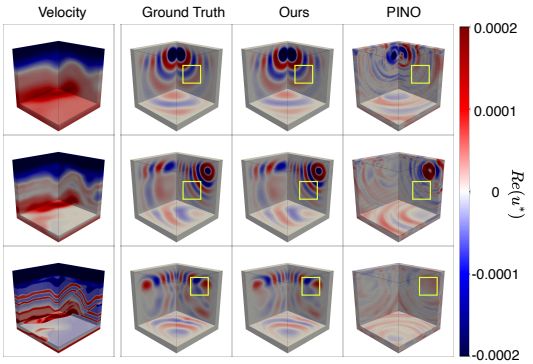

Figure 7: **3D forward modeling** (128^3)**.** Rows: $\sigma = 5 \to 0$ (increasing complexity). Columns: velocity, ground truth, CBNO, PINO.

3.4 Ablation Study

Table 3(a) evaluates forward-modeling components on Marmousi and Overthrust across 4–20 Hz with fixed sources, including input formulation, backbone, frequency experts, supervision, and CBS depth. Each metric is an equal-weight mean of the two dataset-level results; the CBNO reference corresponds to the two-dataset average in Table 1. Removing nondimensionalization or the background field degrades performance, while more frequency experts and deeper CBS supervision generally improve accuracy. Appendix E.2 provides additional controls: under matched one-step CBS supervision, Axial-AFNO achieves 0.2180 RelError versus 0.2707 for FNO, a 19.5% reduction. Three training seeds yield 0.1535 ± 0.0002 (mean $\pm$ sample standard deviation; Appendix E.3). For inversion, Table 3(b) compares full-wavefield and upgoing objectives. Their RMSE values are similar (412.76 versus 410.74), while the upgoing objective gives lower MAE and RelError and higher correlation and SSIM. Metrics are averaged within each dataset and then equally across both datasets.

Table 3: **Component ablations.** (a) Equal-weight means over Marmousi and Overthrust; reference: CBNO (4 experts, $K = 5$). (b) Inversion objectives; Upgoing matches Table 2 (CBNO OOD).

(a) Forward modeling Variant	RMSE↓	MAE↓	RelErr.↓	Corr.↑	SSIM↑
CBNO (avg.)	0.0007	0.0004	0.1536	0.9835	0.9375
Input formulation					
No nondim.	0.0012	0.0009	0.1835	0.9510	0.8850
No bg. field	0.0057	0.0049	0.2472	0.9522	0.4108
Backbone					
DPOT backbone	0.0009	0.0006	0.1544	0.9732	0.9259
Frequency experts					
1 expert	0.0014	0.0012	0.2470	0.9689	0.9094
2 experts	0.0010	0.0007	0.1950	0.9750	0.9210
Supervision					
PINN loss	0.0072	0.0049	0.9685	0.2495	0.3287
CBS depth					
$K = 1$	0.0011	0.0009	0.2180	0.9426	0.9186
$K = 3$	0.0008	0.0006	0.1784	0.9632	0.9319
$K = 10$	0.0006	0.0004	0.1528	0.9830	0.9394
(b) Velocity inversion Objective	RMSE↓	MAE↓	RelErr.↓	Corr.↑	SSIM↑
Full wavefield	412.76	396.53	0.1475	0.7957	0.2893
Upgoing	410.74	288.66	0.1293	0.899	0.515

4 Conclusion

We presented CBNO, a physics-supervised neural-operator framework for acoustic wavefield modeling and inversion. CBS fixed-point consistency enables training without paired wavefield targets, while frequency-partitioned Axial-AFNO experts support broadband prediction. The frozen surrogate is integrated with a reflection-aware inversion objective. Experiments demonstrate forward-modeling accuracy, transfer to external 2D benchmarks, and scalability through separately trained 3D models. Timing comparisons show acceleration over CBS and an accuracy–latency trade-off against learned surrogates. Noise and receiver-dropout tests characterize robustness within the evaluated synthetic settings and reveal sensitivity to severe noise. Future work will address field observations, elastic wave propagation, and uncertainty quantification.

AI USE STATEMENT

Generative AI tools were used to assist with bibliography verification and LaTeX format conversion. All AI-assisted changes were reviewed by the authors, who take responsibility for the final manuscript, its claims, and its artifacts.

ETHICS STATEMENT

This work studies computational seismic modeling and inversion using synthetic or public benchmark data. It does not involve human subjects or personal data. Any downstream use for subsurface decision-making should include domain-expert validation and follow the applicable safety, environmental, and regulatory requirements.

REPRODUCIBILITY STATEMENT

The appendices document the mathematical formulation, model architecture, baseline methods, implementation settings, evaluation metrics, and additional results used in this study. These details are provided to support independent reproduction of the reported experiments.

REFERENCES

Eliane Bécache, A. S. Bonnet-Ben Dhia, and Guillaume Legendre. Perfectly matched layers for time-harmonic acoustics in the presence of a uniform flow. *SIAM Journal on Numerical Analysis*, 44(3):1191–1217, 2006.

Jean-Pierre Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of computational physics*, 114(2):185–200, 1994.

Alfredo Bermúdez, Luis Hervella-Nieto, Andrés Prieto, and Rodolfo Rodríguez. An optimal perfectly matched layer with unbounded absorbing function for time-harmonic acoustic scattering problems. *Journal of Computational Physics*, 223(2):469–488, 2007.

Jon F. Claerbout. Toward a unified theory of reflector mapping. *Geophysics*, 36(3):467–481, Jun 1971. doi: 10.1190/1.1440185.

Laura Cobden, Jingyi Zhuang, Wenjie Lei, Renata Wentzcovitch, Jeannot Trampert, and Jeroen Tromp. Full-waveform tomography reveals iron spin crossover in Earth's lower mantle. *Nature Communications*, 15(1):1961, 2024.

Chengyuan Deng, Shihang Feng, Hanchen Wang, Xitong Zhang, Peng Jin, Yinan Feng, Qili Zeng, Yinpeng Chen, and Youzuo Lin. OpenFWI: large-scale multi-structural benchmark datasets for full waveform inversion. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Andreas Fichtner, Hans-Peter Bunge, and Heiner Igel. The adjoint method in seismology: I. theory. *Physics of the Earth and Planetary Interiors*, 157(1–2):86–104, 2006. doi: 10.1016/J.PEPI.2006.03.016.

Scott W French and Barbara Romanowicz. Broad plumes rooted at the base of the Earth's mantle beneath major hotspots. *Nature*, 525(7567):95–99, 2015.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive Fourier neural operators. In *International Conference on Learning Representations*, 2022.

Peng Guo, Satish C Singh, Venkata A Vaddineni, Ingo Grevemeyer, and Erdinc Saygin. Lower oceanic crust formed by in situ melt crystallization revealed by seismic layering. *Nature Geoscience*, 15 (7):591–596, 2022.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. DPOT: auto-regressive denoising operator transformer for large-scale PDE pre-training. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024.

Juncai He, Xinliang Liu, and Jinchao Xu. MgNO: Efficient parameterization of linear operators via multigrid. In *International Conference on Learning Representations*, 2024.

Jan S Hesthaven. On the analysis and construction of perfectly matched layers for the linearized Euler equations. *Journal of computational Physics*, 142(1):129–147, 1998.

Alexandre Hoffmann, Romain Brossier, Ludovic Métivier, and Alizia Tarayoun. Local uncertainty quantification for 3-D time-domain full-waveform inversion with ensemble Kalman filters: application to a North Sea OBC data set. *Geophysical Journal International*, 237(3):1353–1383, 2024. doi: 10.1093/gji/ggae114.

Xinquan Huang, Fu Wang, and Tariq Alkhalifah. Physics-informed waveform inversion using pretrained wavefield neural operators. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–14, 2025. doi: 10.1109/TGRS.2025.3624025.

Peng Jin, Xitong Zhang, Yinpeng Chen, Sharon X. Huang, Zicheng Liu, and Youzuo Lin. Unsupervised learning of full-waveform inversion: Connecting CNN and partial differential equation in a loop. In *ICLR 2022*, 2022.

Steven G Johnson. Notes on perfectly matched layers (PMLs), 2021.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Shiqian Li, Zhi Li, Zhancun Mu, Shiji Xin, Zhixiang Dai, Kuangdai Leng, Rita Zhang, Xiaodong Song, and Yixin Zhu. GlobalTomo: A global dataset for physics-ML seismic wavefield modeling and FWI. In *NeurIPS*, 2025.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for PDEs on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM / IMS J. Data Sci.*, 1(3), May 2024. doi: 10.1145/3648506.

Feng Liu, Haipeng Li, Guangyuan Zou, and Junlun Li. Automatic differentiation-based full waveform inversion with flexible workflows. *Journal of Geophysical Research: Machine Learning and Computation*, 2(1):e2024JH000542, 2025.

Qing-Huo Liu and Jianping Tao. The perfectly matched layer for acoustic waves in absorptive media. *The Journal of the Acoustical Society of America*, 102(4):2072–2082, 1997.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Lu Lu, Xuhui Meng, Shengze Cai, Zhiping Mao, Somdatta Goswami, Zhongqiang Zhang, and George Em Karniadakis. A comprehensive and fair comparison of two neural operators (with practical extensions) based on fair data. *Computer Methods in Applied Mechanics and Engineering*, 393:114778, 2022.

Ludovic Métivier, François Bretaudeau, Romain Brossier, Stéphane Operto, and Jean Virieux. Full waveform inversion and the truncated newton method: quantitative imaging of complex subsurface structures. *Geophysical Prospecting*, 62(6):1353–1375, 2014.

S Mostafa Mousavi and Gregory C Beroza. Deep-learning seismology. *Science*, 377(6607):eabm4470, 2022.

S Mostafa Mousavi and Gregory C Beroza. Machine learning in earthquake seismology. *Annual review of earth and planetary sciences*, 51:105–129, 2023.

Gerwin Osnabrugge, Saroch Leedumrongwatthanakun, and Ivo M. Vellekoop. A convergent Born series for solving the inhomogeneous Helmholtz equation in arbitrarily large media. *Journal of Computational Physics*, 322:113–124, 2016. doi: 10.1016/j.jcp.2016.06.034.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Gerard T. Schuster, Yuqing Chen, and Shihang Feng. Review of physics-informed machine-learning inversion of geophysical data. *Geophysics*, 89(6):T337–T356, 2024. doi: 10.1190/geo2023-0615.1.

Jian Sun, Zhan Niu, Kristopher A Innanen, Junxiao Li, and Daniel O Trad. A theory-guided deep-learning formulation and optimization of seismic waveform inversion. *Geophysics*, 85(2):R87–R99, 2020.

Jian Sun, Kristopher Innanen, Tianze Zhang, and Daniel Trad. Implicit seismic full waveform inversion with deep neural representation. *Journal of Geophysical Research: Solid Earth*, 128(3): e2022JB025964, 2023.

Carl Tape, Qinya Liu, Alessia Maggi, and Jeroen Tromp. Adjoint tomography of the Southern California crust. *Science*, 325(5943):988–992, 2009.

Albert Tarantola. Linearized inversion of seismic reflection data. *Geophysical Prospecting*, 32(6): 998–1015, 1984a. doi: 10.1111/j.1365-2478.1984.tb00751.x.

Albert Tarantola. Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49 (8):1259–1266, 1984b. doi: 10.1190/1.1441754.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized Fourier neural operators. In *International Conference on Learning Representations (ICLR)*, 2023.

Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009. doi: 10.1190/1.3238367.

Fang Wang, Daniela Donno, Hervé Chauris, Henri Calandra, and François Audebert. Waveform inversion based on wavefield decomposition. *Geophysics*, 81(6):R457–R470, 2016.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021a.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *J. Comput. Phys.*, 449:110768, 2022.

Tengfei Wang, Jiubing Cheng, and Jianhua Geng. Reflection full waveform inversion with second-order optimization using the adjoint-state method. *Journal of Geophysical Research: Solid Earth*, 126(8):e2021JB022135, 2021b.

Tian Wang and Chuang Wang. Latent neural operator for solving forward and inverse PDE problems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33085–33107, 2024.

Xinming Wu, Jianwei Ma, Xu Si, Zhengfa Bi, Jiarun Yang, Hui Gao, Dongzi Xie, Zhixiang Guo, and Jie Zhang. Sensing prior constraints in deep neural networks for solving exploration geophysical problems. *Proceedings of the National Academy of Sciences*, 120(23):e2219573120, 2023. doi: 10.1073/pnas.2219573120.

Yue Wu and Youzuo Lin. InversionNet: An efficient and accurate data-driven full waveform inversion. *IEEE Transactions on Computational Imaging*, 6:419–433, 2020. doi: 10.1109/TCI.2019.2956866.

Fangshu Yang and Jianwei Ma. Deep-learning inversion: A next-generation seismic velocity model building method. *Geophysics*, 84(4):R583–R599, 2019.

Yan Yang, Angela F. Gao, Kamyar Azizzadenesheli, Robert W. Clayton, and Zachary E. Ross. Rapid seismic waveform modeling and inversion with neural operators. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. doi: 10.1109/TGRS.2023.3264210.

Kaiqing Yuan and Barbara Romanowicz. Seismic evidence for partial melting at the root of major hot spot plumes. *Science*, 357(6349):393–397, 2017.

Qili Zeng, Shihang Feng, Brendt Wohlberg, and Youzuo Lin. InversionNet3D: Efficient and scalable learning for 3-D full-waveform inversion. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. doi: 10.1109/TGRS.2021.3135354.

Zhijun Zeng, Youjia Zheng, Chang Su, Qianhang Wu, Hao Hu, Zeyuan Dong, Yang Lv, Ligang Cui, Zhiyong Hou, Weijun Lin, Zuoqiang Shi, Yubing Li, and He Sun. Ultrasound tomography of musculoskeletal tissues with generative neural physics. *arXiv preprint arXiv:2508.12226*, 2025. URL `https://arxiv.org/abs/2508.12226v4`. Version 4, revised August 6, 2026.

Wei Zhang, Tong Li, Yang Shen, and Xiaofei Chen. Efficient perfectly matched layer implementation for seismic-wave modeling in 3d spherical coordinate sections. *Seismological Research Letters*, 96 (4):2368–2381, 2025.

Zhongping Zhang and Youzuo Lin. Data-driven seismic waveform inversion: A study on the robustness and generalization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10): 6900–6913, 2020.

Wei Zhou, Romain Brossier, Stéphane Operto, and Jean Virieux. Full waveform inversion of diving & reflected waves for velocity model building with impedance inversion based on scale separation. *Geophysical Journal International*, 202(3):1535–1554, 2015.

Hejun Zhu and Jeroen Tromp. Mapping tectonic deformation in the crust and upper mantle beneath Europe and the North Atlantic Ocean. *Science*, 341(6148):871–875, 2013.

A Related Work

A.1 Learned inversion networks for FWI

Learned inversion networks formulate FWI as a direct regression/translation problem from seismic records to subsurface parameters, typically via encoder–decoder architectures that approximate an inverse operator (Yang & Ma, 2019). Representative examples include Wu & Lin (2020), which introduces an efficient data-driven inversion network, and Zeng et al. (2022), which extends the idea to 3D with memory- and computation-aware designs. This line of work offers extremely fast inference without iterative PDE solves and can achieve strong reconstructions under fixed acquisition geometries and matched training distributions. However, it often relies on large paired datasets (especially challenging in 3D) and may suffer from limited generalization and weaker physical consistency when acquisition, frequency band, or geology shifts. We adopt a *forward-first* route: a label-free CBS-supervised forward surrogate (with frequency-band MoE and reflection-aware inversion) is trained once for AD-FWI, preserving the physical forward–inverse loop with support for 3D modeling.

A.2 Unsupervised / physics-in-the-loop FWI

Unsupervised or physics-in-the-loop approaches incorporate differentiable forward modeling into the training/optimization loop, using data reconstruction loss rather than velocity labels to drive learning or parameter updates (Sun et al., 2020). A representative example is Jin et al. (2022), which couples a CNN-based inverse model with a differentiable finite-difference wave solver in a closed loop; Deng et al. (2022) further provides a large-scale benchmark that highlights both opportunities and difficulties at scale. Such inverse-learning approaches reduce reliance on target velocity labels, but repeated forward/adjoint PDE computations remain costly in multi-source, multi-frequency, and 3D settings. CBNO addresses a different label requirement: its forward-operator training uses velocity models and sources as inputs but requires no paired wavefield targets. CBS supplies this training signal offline, while the learned surrogate supports repeated differentiable forward evaluations during inversion. Frequency-band experts support broadband prediction, and the reflection-aware objective emphasizes weak reflections in the observed data; these observations are still required for inversion.

A.3 Wavefield surrogates / operator learning / 3D scalable PDE surrogates

Wavefield surrogates and operator-learning methods aim to approximate the PDE solution operator that maps medium parameters and sources to wavefields/records, replacing classical solvers with neural surrogates at inference time. Canonical operator-learning paradigms include Fourier Neural Operators (Li et al., 2021) and DeepONet (Lu et al., 2021), and physics-regularized training has been explored in Physics-Informed Neural Operators (Li et al., 2024). Relative to Fourier-attention operators such as DPOT (Hao et al., 2024), our backbone emphasizes cross-axis low-rank sharing and explicit axis/wavenumber gating, complemented by source-frequency experts. In the seismic context, neural-operator-based surrogates have also been explored for fast waveform modeling and differentiable inversion (Yang et al., 2023). The main advantage is that forward computation can be amortized into offline training, yielding substantial speedups and offering native differentiability for integration with inversion. However, broadband and 3D settings remain difficult: supervised surrogate training can demand massive high-fidelity wavefield/record data (costly to generate and store in 3D), and distribution shifts across frequency bands, acquisition, and model families can induce out-of-distribution errors or instability.

Closely related approaches incorporate physics at different stages. Huang et al. (2025) train an FNO on finite-difference scattered-wavefield targets, then freeze the operator and introduce a wave-equation residual during inversion. For ultrasound tomography, the Strong Scattering Neural Operator (S^2NO) of Zeng et al. (2025) uses CBS-inspired latent iterations and trains frequency-specific networks on CBS-generated wavefields with a relative L_2 loss. In contrast, CBNO uses numerical CBS updates to define a fixed-point consistency loss for training an Axial-AFNO surrogate without paired wavefield targets, and subsequently uses the frozen surrogate for reflection-aware inversion.

A.4 Reflection-enhanced inversion and wavefield decomposition

Reflection-enhanced formulations and upgoing/downgoing decomposition aim to mitigate the dominance of direct arrivals and diving waves by separating wavefield components to emphasize weak reflections. Classic perspectives on using separated wavefields for imaging trace back to Claerbout (1971), with later works exploring decomposition-based inversion objectives (Wang et al., 2016) or strategies that jointly exploit diving and reflected waves to improve velocity building (Zhou et al., 2015). While these approaches inject valuable physical priors that enhance interface identifiability, they often incur substantial computational overhead because the decomposition typically requires repeated PDE solves, especially in 3D.

B Mathematical Formulations and Physics Priors

B.1 CBS Solver & Scattering Formulation

We derive the solver framework starting from the forward problem definition in Eq. (1). The acoustic wave propagation is governed by:

$$-\omega^2 \hat{u} - c(\mathbf{x})^2 \nabla^2 \hat{u} = s(\mathbf{x}, \omega). \tag{10}$$

To transition to a scattering formulation, we normalize the equation by dividing by $-c(\mathbf{x})^2$. Letting $k(\mathbf{x}) = \omega / c(\mathbf{x})$ be the heterogeneous wavenumber and $S(\mathbf{x}, \omega) = s(\mathbf{x}, \omega)/c(\mathbf{x})^2$ be the scaled source, we obtain the canonical Helmholtz form:

$$\nabla^2 \hat{u} + k(\mathbf{x})^2 \hat{u} + S = 0. \tag{11}$$

With damping $\epsilon > 0$, we decompose the squared wavenumber into a complex background and a compensating scattering potential,

$$k(\mathbf{x})^2 = k_0^2 + i\epsilon + V(\mathbf{x}), \qquad V(\mathbf{x}) = k(\mathbf{x})^2 - k_0^2 - i\epsilon. \tag{12}$$

The corresponding damped background Green's operator is

$$G_0 = -\left(\nabla^2 + k_0^2 + i\epsilon\right)^{-1}. \tag{13}$$

The corresponding Lippmann–Schwinger equation is then written as:

$$\hat{u} = G_0 \left(V \hat{u} + S\right), \tag{14}$$

The $-i\epsilon$ term in V exactly compensates the background damping, preserving the original Helmholtz equation.

Direct iteration of Eq. (14), i.e., the classical Born series, may diverge in high-contrast media. Following Eqs. (10)–(11) of Osnabrugge et al. (2016), CBS uses the pointwise preconditioner $\gamma(\mathbf{x}) = iV(\mathbf{x})/\epsilon$ with $\epsilon \geq \max_{\mathbf{x}} |k(\mathbf{x})^2 - k_0^2|$ and $\epsilon > 0$. Thus γ is spatially varying in heterogeneous media and fixed by V and ϵ. The CBS update is

$$u_{n+1} = u_n + \gamma \left[G_0 \left(V u_n + S\right) - u_n\right], \tag{15}$$

equivalent to their Eq. (13), with iteration operator $M = I - \gamma + \gamma G_0 V$. Products with V and γ are pointwise. Their Appendix A establishes $\rho(M) \leq 1$ for passive media (no gain), and convergence with absorbing boundaries or nonzero absorption in a finite volume. This result concerns their Green's operator and boundary treatment; it does not by itself prove convergence of a PML-discretized variant. In Eq. (5), $V^* = h^2 V$ and $\epsilon^* = h^2 \epsilon$, leaving $\gamma = iV^*/\epsilon^*$ unchanged.

Stopping Criterion. For standalone solves, we monitor the relative change in the wavefield. The iteration terminates when the relative iterate change falls below a user-defined tolerance τ (e.g., $\tau = 10^{-5}$):

$$\frac{\|u_{n+1} - u_n\|_2}{\|u_{n+1}\|_2} < \tau. \tag{16}$$

This stopping statistic measures successive-iterate agreement, not the Helmholtz residual or solution error. Training instead uses the fixed K steps specified in Section 2.5.

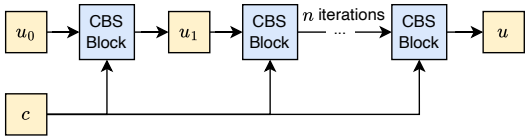

Figure 8: Schematic of the iterative Convergent Born Series (CBS) computation. For standalone solves, the wavefield u is refined until the relative iterate change falls below the stopping tolerance.

B.2 PML Boundary Conditions

To simulate semi-infinite domains on a finite grid, the background Green's function G_0 must satisfy radiation conditions. We implement Perfectly Matched Layers (PML) via **Complex Coordinate Stretching** directly in the frequency domain.

For each spatial direction j, we define the complex stretching function $s_j(x_j) = 1 + i\sigma_j(x_j)/\omega$, with $j \in \{x, z\}$ in 2D and $j \in \{x, y, z\}$ in 3D. Here, σ_j denotes the conductivity profile within the PML region. The differential operators are modified as $\partial_j \to \frac{1}{s_j}\partial_j$. Consequently, the Laplacian operator within G_0 becomes:

$$\nabla^2_{\mathrm{PML}} = \sum_{j=1}^{d} \frac{1}{s_j} \frac{\partial}{\partial x_j} \left(\frac{1}{s_j} \frac{\partial}{\partial x_j} \right). \tag{17}$$

By inverting this PML-augmented operator to compute G_0, waves leaving the computational domain are exponentially attenuated, thereby suppressing artificial boundary reflections and improving the physical accuracy of the solver.

B.3 Consistency of the Final-Iterate Physics Loss

We establish a conditional statement for the final-iterate objective in Eq. (6). Fix the medium, source, frequency, and discretization. Let $u \in \mathbb{C}^N$ collect the dimensionless wavefield degrees of freedom, including the PML, and write the discrete Helmholtz system as $Au = q$, with $q = (c^*)^{-2} f^*$. Let $G = G_0^*$ and $V = V^*$ denote the discrete background Green matrix and diagonal scattering potential, and let $\Gamma = \mathrm{diag}(\gamma)$.

Proposition B.1 (Unique zero of the final-iterate loss)**.** *Assume that G and Γ are nonsingular, that $A = G^{-1} - V$ for the same discrete boundary treatment, and that the CBS iteration matrix $M = I - \Gamma + \Gamma G V$ satisfies $\rho(M) < 1$. For any integer $K \geq 1$, the loss*

$$\mathcal{L}_K(u) = \frac{1}{K} \|u - \mathcal{T}^K(u)\|_2^2, \qquad \mathcal{T}(u) = Mu + \Gamma G q, \tag{18}$$

has a unique zero and global minimizer over $\mathbb{C}^N$, namely the solution $u^\dagger$ of $Au = q$.

Proof. The consistent splitting gives

$$I - M = \Gamma G A, \qquad \mathcal{T}(u) - u = \Gamma G(q - Au). \tag{19}$$

Since $\rho(M) < 1$, $I - M$ is invertible. Nonsingularity of ΓG then implies that A is invertible and that $u^\dagger = A^{-1}q$ is the unique fixed point of $\mathcal{T}$. Subtracting this fixed point from each update and iterating yields

$$\mathcal{T}^K(u) - u = (M^K - I)(u - u^\dagger). \tag{20}$$

Moreover, $\rho(M^K) = \rho(M)^K < 1$, so $I - M^K$ is invertible. Consequently, $\mathcal{L}_K(u) = 0$ if and only if $u = u^\dagger$. Nonnegativity of the loss establishes the unique global minimizer. □

Scope. The norm covers all discrete degrees of freedom. Fixed positive scalar normalization or positive-definite weighting preserves the zero set; discarding residual components requires a separate argument. The convergence analysis of Osnabrugge et al. (2016) does not establish the assumptions for arbitrary PML discretizations (Appendix B.1). This wavefield-space result guarantees neither

neural-network optimization convergence nor a discretization-independent bound from small loss to small solution error.

Numerical verification. We directly optimize the discrete wavefield, without a neural network, on a Marmousi case at 10 Hz with one source and $K = 5$ CBS updates per loss evaluation. The run yields a recorded physics loss of 2.365×10^{-4} and a relative L_2 wavefield error of 5.042×10^{-4} against a converged CBS reference. This supports wavefield recovery in this case, without establishing convergence for other discretizations or neural-network training.

B.4 Analytical Background Wavefield Calculation

In the main text, we use a background wavefield u^*_{bg} to guide residual learning. This corresponds to the `precompute` function in our implementation. The background field provides a stable analytical prior that captures the dominant source-centered oscillation and the dimension-dependent amplitude decay. The network then learns the scattered residual on top of this background field.

1. General formulation. Given the dimensionless velocity c^* and source mask f^*, we compute the spatial mean velocity $\bar{c}^*$ and the Euclidean distance field r from the source centroid. We represent the oscillatory phase using two real-valued channels corresponding to the real and imaginary parts:

$$\Phi(r) = \alpha \left[\cos(\beta r + \gamma), -\sin(\beta r + \gamma)\right]^T, \qquad \Phi(r) \in \mathbb{R}^2, \tag{21}$$

where α, β, and γ are fixed hyperparameters used to stabilize the phase and amplitude of the background initialization.

2. 2D case. For 2D wave propagation, the asymptotic cylindrical-wave amplitude decays approximately as $1/\sqrt{r}$. We therefore use the following stabilized background prior:

$$u^*_{\mathrm{bg,2D}}(r) = \frac{\Phi(r/\bar{c}^*)}{\bar{c}^* \sqrt{\bar{c}^*\, r + \epsilon_r}}, \tag{22}$$

where $\epsilon_r > 0$ prevents a singular value at the source location. In practice, the source-grid value is replaced by a finite clipped value to avoid dominating the dynamic range.

3. 3D case. For 3D wave propagation, the spherical-wave amplitude decays approximately as $1/r$. Accordingly, we use

$$u^*_{\mathrm{bg,3D}}(r) = \frac{\Phi(r/\bar{c}^*)}{(\bar{c}^*)^2 \max(r, r_0) + \epsilon_r}, \tag{23}$$

where $r_0 > 0$ and $\epsilon_r > 0$ stabilize the near-field response.

The background field serves as an input feature and an additive reference for residual prediction. The final prediction is still trained and evaluated through the CBS-based physics supervision, so the analytical prior does not impose an exact closed-form solution in heterogeneous media.

C Neural Architecture Details

C.1 Frequency-Partitioned Experts with Axial-AFNO

Our architecture is designed to balance the global receptive field required for wavefield modeling with the parameter and memory efficiency demanded by high-dimensional seismic data. We use two related but distinct notions of frequency partitioning. First, at the model-routing level, the source-frequency range is divided into M_ω temporal-frequency bands that overlap only at shared endpoints. The single-valued router $r(\omega)$ selects one lightweight expert per frequency, including at the endpoints, as described in Sec. 2.4. Second, inside each expert, the Fourier modes along spatial axes are divided into M_k spatial-wavenumber bands for axis-aware spectral gating. This distinction avoids conflating temporal source frequency with spatial Fourier wavenumber.

For spatial-wavenumber gating, consider a spatial axis of length L_d. For an rFFT mode index $k \in \{0, 1, \ldots, \lfloor L_d/2 \rfloor\}$, the corresponding band index is computed as

$$b_d(k) = \min\left(\left\lfloor \frac{k}{\lfloor L_d/2 \rfloor} \times M_k \right\rfloor, M_k - 1\right). \tag{24}$$

This linear mapping allows the network to apply distinct gating parameters to low- and high-wavenumber components, enabling adaptive filtering of smooth background trends and high-wavenumber scattering patterns within the spectral domain.

Each expert integrates this mechanism into a hierarchical U-Net-like structure adapted for volumetric data. The pipeline begins with an Input Smooth Projection module, consisting of stacked 3D convolutions with kernels 5^3, 7^3, and 9^3 to suppress grid artifacts and high-frequency noise. A strided convolution then projects the input to a latent grid H_0. The backbone consists of an encoder with residual blocks, each integrating Group Normalization, the Axial-AFNO token mixer, and a channel-mixing MLP. The decoder employs trilinear upsampling followed by depthwise-separable convolutions, avoiding checkerboard artifacts caused by transposed convolutions and improving the smoothness of reconstructed wavefields.

The Axial-AFNO module processes the hidden tensor H via parallel 1D spectral branches along each axis $d \in \mathcal{D}$, where $\mathcal{D} = \{H, W\}$ in 2D and $\mathcal{D} = \{D, H, W\}$ in 3D. Each branch mixes features globally along its axis, while stacked layers propagate information across axes. The branches share a low-rank channel subspace and use axis-dependent gates, providing a structured alternative to DPOT's grouped Fourier MLPs. Specifically, we first transform the input into the frequency domain along axis d:

$$Z_d = \mathrm{rFFT}(H, \mathrm{dim} = d). \tag{25}$$

For a specific mode k, we use two low-rank basis matrices U and V, shared across all spatial axes, to compress and reconstruct the channel dimension. A band-specific gate $G_{d,b_d(k)}$ is then applied, where the gate is unique to the current axis and spatial-wavenumber band. For notational simplicity, $Z_d(k)$ in the following equation denotes the real-valued real–imaginary representation of the complex spectral coefficient. The spectral mixing operation is formally defined as

$$Z_{\mathrm{mix},d}(k) = \sigma\left(\left[(Z_d(k)U) \odot G_{d,b_d(k)}\right] V^T + \mathrm{bias}\right). \tag{26}$$

Here, the projection bases U and V compress the channel dimension to a low-rank latent space with rank $r \ll C$, while $G_{d,b_d(k)}$ performs axis-aware and wavenumber-aware modulation. In implementation, complex spectral coefficients are represented by concatenating their real and imaginary components as real-valued channels, and the nonlinearity σ is applied component-wise to this real-valued representation. Thus, GeLU is not applied directly to complex-valued scalars. Low rank and cross-axis sharing constrain the spectral parameterization; they do not change the asymptotic complexity of the FFT operation.

After spectral mixing, the processed features are transformed back to the spatial domain through inverse FFT along the corresponding axis. The multi-axis outputs are aggregated by a learnable weighted residual fusion:

$$H_{\mathrm{final}} = H + \sum_{d \in \mathcal{D}} \mathrm{Softmax}(\alpha)_d \mathcal{A}_d(H), \tag{27}$$

where $\mathcal{A}_d$ denotes the axial spectral branch along axis d, and the learned scalar weights α_d set the relative contributions of different directions, such as vertical depth variations or horizontal continuity. Together with the wavenumber-band gates, this fusion encodes directional and scale-dependent wavefield structure within a shared low-rank representation.

C.2 Efficiency and Scalability Analysis

We analyze *parameter efficiency* relative to DPOT and *token-mixing scalability* relative to standard self-attention. CBNO and DPOT both use FFT-based mixing and share its token-scaling properties.

Parameter Efficiency. DPOT uses channel-grouped Fourier MLPs with weights shared across modes (Hao et al., 2024); CBNO uses cross-axis shared low-rank bases with axis/wavenumber-band gates. For the spectral mixer alone, the leading parameter counts scale as $\mathcal{O}(C^2/h)$ for DPOT and $\mathcal{O}(Cr + aM_k r)$ for Axial-AFNO, where h is the group count, a is the number of spatial axes, and r is the latent rank; biases and real–imaginary constant factors are omitted. A sufficiently small rank can reduce mixer parameters, but also constrains the representation. Whole-model counts additionally depend on channel MLPs, depth, input/output layers, and expert count. Table 4 reports an 18.4% parameter reduction at $C = 2048$ (219.3M vs. 268.8M) for the compared configurations. Runtime and peak-memory comparisons require empirical measurements (Appendix C.3).

Table 4: **Parameter Efficiency Analysis.** Counts include all registered parameters of a single model at each channel width C. This comparison uses 12 layers, 4 channel groups, a 64×64 physical grid, 10-point PML, patch size 4, and MLP ratio 2. Axial-AFNO uses rank 8, 4 wavenumber bands, and multi-axis fusion.

Channel Width	DPOT Params ↓	CBNO Params ↓	Reduction ↑
$C = 256$	4.8 M	**4.2 M**	12.5%
$C = 512$	17.8 M	**14.9 M**	16.3%
$C = 1024$	68.4 M	**56.3 M**	17.7%
$C = 2048$	268.8 M	**219.3 M**	**18.4%**

Table 5: **Leading token-mixer costs.** L: mixer-input tokens; C: channels; h: DPOT groups; r: rank; a: spatial axes. Memory counts intermediate activations, excluding parameters and separate channel MLPs. The self-attention memory bound assumes a materialized attention matrix, not a memory-efficient implementation.

Mechanism	Mixing Cost	Memory	Mixer
Self-Attention	$\mathcal{O}(LC(L + C))$	$\mathcal{O}(L^2 + LC)$	Token pairs
DPOT	$\mathcal{O}(LC(\log L + C/h))$	$\mathcal{O}(LC)$	Grouped Fourier MLP
Axial-AFNO	$\mathcal{O}(LC(\log L + ar))$	$\mathcal{O}(aLC)$	Axial low-rank gates

Scalability to High-Dimensional Data. Let L denote the number of spatial tokens at the mixer input, after patching or downsampling. Table 5 gives leading mixer costs with fixed MLP expansion ratios and at most L processed Fourier coefficients per branch. DPOT includes FFTs and grouped channel transforms; Axial-AFNO includes axial FFTs and low-rank transforms across a branches. For fixed channel and rank settings, both spectral mixers avoid the quadratic token-pair computation of standard self-attention. Axial factorization retains the asymptotic FFT order. The separate channel MLP adds $\mathcal{O}(LC^2)$ per layer at fixed expansion ratio to each reported mixer cost. The structural distinction from DPOT lies in the low-rank, axis- and wavenumber-aware parameterization. Actual speed and memory depend on resolution, retained modes, branch implementation, and model configuration.

C.3 Empirical Inference Cost

We benchmark the 4–20 Hz checkpoints of FFNO, PINO, DPOT, and CBNO on an NVIDIA RTX 4090 using a 70×70 grid, batch size 1, and FP32 precision. Table 6 reports per-sample latency and peak GPU memory measured with PyTorch, alongside the Marmousi relative errors from Table 1. CBNO has the lowest error but the highest latency among these four models. This comparison captures an accuracy–latency trade-off, separate from the A100-based acceleration over the CBS solver and the inversion timings in Table 2.

Table 6: **Inference cost of learned surrogates.** Latency and peak GPU memory are measured on an RTX 4090; RelError refers to the Marmousi results in Table 1. Lower is better for all three metrics.

Model	Latency (ms/sample)	Peak GPU memory (GiB)	RelError
FFNO	17.426	0.167	0.6035
PINO	**11.439**	0.156	0.2707
DPOT	25.004	0.157	0.6597
CBNO	29.183	**0.147**	**0.1464**

D Experimental Setup and Reproducibility

D.1 Baseline Method Descriptions

We compare CBNO with numerical solvers, physics-informed neural operators, and data-driven inversion networks.

Convergent Born Series (CBS) (Osnabrugge et al., 2016): Serving as the high-precision numerical reference solver, CBS solves the inhomogeneous Helmholtz equation via a preconditioned Born series. Unlike the standard Born series, which may diverge under strong scattering, CBS provides a convergent formulation for strongly scattering media under its convergence conditions and serves as the reference solver for error quantification and the physics engine for AD-FWI.

Physics-Informed Neural Networks (PINN) (Raissi et al., 2019): A coordinate-based MLP that directly minimizes the PDE residual with simplified absorbing boundary conditions.

Physics-Informed Neural Operator (PINO) (Li et al., 2024): Enhances the Fourier Neural Operator (FNO) with physics constraints. To ensure a fair comparison, we upgrade the standard PINO by replacing the generic PDE residual with a one-step CBS iteration loss, explicitly embedding scattering physics into the spectral learning process.

Factorized Fourier Neural Operator (FFNO) (Tran et al., 2023): Uses separable Fourier layers that apply spectral transformations independently along each spatial dimension and combine their outputs, together with improved residual connections.

Multigrid Neural Operator (MgNO) (He et al., 2024): Inspired by multigrid solvers, MgNO uses a U-Net architecture to process features iteratively across multiple resolutions, capturing both local interactions and global propagation patterns.

Denoising Pre-training Operator Transformer (DPOT) (Hao et al., 2024): Uses an AFNO-inspired Fourier-attention backbone for operator learning. Spatial features are transformed into the Fourier domain, processed by channel-grouped MLPs with weights shared across Fourier modes, and mapped back to the spatial domain by inverse FFTs.

InversionNet (Wu & Lin, 2020): A representative purely data-driven CNN encoder-decoder that maps seismic data directly to velocity models. We use this baseline to assess data-driven inversion on the evaluated external 2D benchmarks.

VeloGAN (Zhang & Lin, 2020): Introduces adversarial training to enforce structural realism. This baseline contrasts texture-focused generative approaches with our physics-consistent inversion, particularly regarding quantitative velocity.

D.2 Detailed Implementation Specifications

Experiments used a single NVIDIA A100 GPU (80GB memory), except for the additional inference benchmark in Appendix C.3, which used an RTX 4090. Below, we provide the architectural configurations and hyperparameter settings for each model.

CBNO (Ours). The 2D CBNO uses a 70×70 physical grid, embedding dimension 512, and 24 Axial-AFNO layers with 8 channel groups per spectral mixer. Its patch kernel is 10×10, with rank 8, 2 spatial-wavenumber bands, and a Fourier-mode cap of 64 per axis. The 3D configuration uses a 128^3 physical grid, embedding dimension 256, 8 layers with 4 channel groups, an 8^3 patch kernel, rank 8, and a mode cap of 32 per axis. Patch embedding applies overlapping convolutions to the PML-padded input, with stride and convolution padding equal to half the kernel size. Each axial branch retains the minimum of its mode cap and the available rFFT coefficients. Both configurations use GeLU and softmax-weighted multi-axis fusion (Appendix C.1); the exponential MLP time aggregation is used in 2D.

Physics-Informed Neural Networks (PINN). The PINN baseline is implemented as a coordinate-based Multi-Layer Perceptron (MLP) mapping spatial coordinates to wavefields. We utilize a SIREN-style architecture ("shallow" configuration) with a width of 50 and Sine activation functions to explicitly model high-frequency derivatives. The model is trained using the Adam optimizer with an initial learning rate of 1×10^{-3} and a StepLR scheduler (gamma=0.7), minimizing a composite loss with boundary constraints weighted by 100 relative to the PDE residual.

Physics-Informed Neural Operator (PINO). For the PINO baseline, we employ a Fourier Neural Operator (FNO) backbone configured with 12 layers. The architecture utilizes 32 Fourier modes in each spatial dimension and a hidden channel width of 92, which is projected to 256 channels for the final output. The model incorporates GeLU activation functions for both the FNO blocks and the MLP layers, operating within a target frequency range (e.g., 4-20 Hz) and utilizing a PML boundary padding of 30 grid points.

Factorized Fourier Neural Operator (FFNO). The FFNO baseline adopts a significantly deeper architecture compared to PINO, comprising 24 layers to capture complex wavefield scattering features. It utilizes a Factorized Fourier structure with 32 modes per dimension and a hidden dimension of 92 (projected to 256). Similar to the PINO setup, it employs GeLU activations and

processes the wavefield with a PML padding of 30 grid points, leveraging the factorization to maintain computational efficiency despite the increased depth.

Multigrid Neural Operator (MgNO). The MgNO baseline is structured with a multigrid U-Net architecture spanning 24 layers. It features a channel width of 32 and utilizes a V-cycle iteration pattern of [2, 1, 1, 1, 2] to capture multi-scale interactions efficiently. The network uses GeLU activation, processing input data with 4 input channels and generating 2 output channels, effectively modeling both local and global wave propagation.

Denoising Pre-training Operator Transformer (DPOT). The DPOT backbone uses AFNO-inspired Fourier attention. Inputs on the 70×70 physical grid are PML-padded and embedded using overlapping 10×10 convolutional kernels with stride and convolution padding 5. The network consists of 24 layers with 8 channel groups per spectral mixer, using an embedding dimension of 512 and an output layer dimension of 128. It performs spectral mixing through FFTs, channel-grouped MLPs with weights shared across Fourier modes, and inverse FFTs. The configuration uses an MLP ratio of 2.0 and an exponential MLP for time aggregation.

Numerical Solver (CBS & AD-FWI). For reference frequency-domain wavefields and the physics-based AD-FWI baseline, we use the Convergent Born Series (CBS) solver for the inhomogeneous Helmholtz equation. CBS operates directly in the frequency domain for each prescribed source frequency. We evaluate the forward solver at frequencies sampled from the 4–20 Hz range, using a PML of 30 grid points and an adaptive stopping tolerance of 10^{-5} unless otherwise specified. Time-domain visualizations combine predictions at multiple frequencies through an inverse Fourier transform with the prescribed source wavelet.

Inversion Hyperparameters. For all differentiable inversion experiments, we optimize the velocity model using Adam with a learning rate of 10 for 20,000 iterations. The initial velocity model is obtained by applying a Gaussian filter with standard deviation $\sigma = 10$ grid points to the target velocity model. After each optimizer step, the velocity values are clipped to the physical range $[600, 6000]$ m/s. Inversion uses frequency-domain observations within 4–20 Hz and a correlation-based data-misfit objective. Reflection-aware preprocessing is implemented using the same upgoing wavefield-separation operator as in Sec. 2.6, with a wavenumber-domain taper/window ratio of 0.2. The correlation loss used to construct the objectives in Eqs. (3) and (9) uses no extra pre-normalization by the observed-data standard deviation. For the CBS-based AD-FWI baseline, each Helmholtz solve uses a convergence tolerance of 10^{-5} and a maximum of $10{,}000$ solver iterations.

Correlation Loss. The total objective $\mathcal{J}_{\mathrm{corr}}$ is assembled from evaluations of $\mathcal{L}_{\mathrm{corr}}$. For one evaluation over B shots, let $\mathbf{p}_b$ and $\mathbf{d}_b$ be complex predicted and observed receiver-data vectors, flattening the frequency and receiver dimensions supplied to that evaluation. With $\widetilde{\mathbf{a}} = \mathbf{a} - \mathrm{mean}(\mathbf{a})\mathbf{1}$, the implementation computes

$$\begin{aligned} \rho_\epsilon(\mathbf{a}, \mathbf{b}) &= \frac{\widetilde{\mathbf{a}}^{\mathsf{T}} \widetilde{\mathbf{b}}}{\max\Big(\|\widetilde{\mathbf{a}}\|_2 \|\widetilde{\mathbf{b}}\|_2, \epsilon \Big)}, \\ \mathcal{L}_{\mathrm{corr}} &= 1 - \frac{1}{2B} \sum_{b=1}^{B} \left[\rho_\epsilon(\mathrm{Re}\, \mathbf{p}_b, \mathrm{Re}\, \mathbf{d}_b) + \rho_\epsilon(\mathrm{Im}\, \mathbf{p}_b, \mathrm{Im}\, \mathbf{d}_b) \right], \end{aligned} \tag{28}$$

where $\epsilon = 10^{-12}$. Real and imaginary parts are centered separately, then averaged across components and shots.

InversionNet. InversionNet (Wu & Lin, 2020) adopts a convolutional Encoder-Decoder architecture. The encoder comprises 8 convolutional blocks that downsample the 5-channel seismic input (1000×70) into a 512-dimensional latent vector, with channel dimensions progressively increasing as [32, 64, 128, 256, 512]. The decoder mirrors this process with 5 deconvolutional blocks, upsampling the latent features to a 70×70 velocity model. We use BatchNorm and LeakyReLU (alpha=0.2) for intermediate layers and a Tanh activation for the output. The model is trained using the AdamW optimizer with a learning rate of 1×10^{-4}, betas of (0.9, 0.999), and a batch size of 256 for 120 epochs. The objective function combines L1 and L2 losses with equal weighting (lambda=1.0).

VeloGAN. VeloGAN (Zhang & Lin, 2020) employs a Generative Adversarial Network framework, utilizing the same generator architecture as InversionNet. The discriminator is a 5-layer convolutional network (channels: [32, 64, 128, 256, 1]) that outputs a scalar Wasserstein critic score. We train the model using the WGAN-GP strategy with a gradient penalty coefficient of 10.0. The generator is optimized using a composite loss function that enforces both pixel-wise accuracy and structural realism: the weights for L1 and L2 losses are set to 100.0, while the adversarial loss weight is set to 1.0. Both the generator and discriminator use the AdamW optimizer with a learning rate of 1×10^{-4} and betas of (0.0, 0.9). Training runs for 500 epochs with a batch size of 64, updating the discriminator 5 times for every generator update.

D.3 Evaluation Metrics

The following metrics apply to real-valued reference and prediction arrays, denoted by u^*_{gt} and u^*_{pred}. For complex-valued wavefields, each metric is computed separately for the real and imaginary components, and the two values are averaged, except for the joint RMSE in Fig. 4(a–c) described below. For velocity inversion, the metrics are applied directly to the real-valued velocity models. Three-dimensional results are evaluated over the full volume. The total number of grid points is $N = H \times W$ in 2D and $N = H \times W \times D$ in 3D, where D is the third grid dimension.

Numerical Accuracy. We first evaluate the pixel-wise amplitude accuracy using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Relative Error (RelError). **RMSE** disproportionately penalizes large deviations, making it highly effective for detecting amplitude mismatches in high-energy regions. Conversely, **MAE** is the arithmetic mean of absolute errors and is less sensitive to outliers than RMSE:

$$\text{RMSE}(u^*_{\text{gt}}, u^*_{\text{pred}}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(u^*_{\text{gt},i} - u^*_{\text{pred},i}\right)^2}, \tag{29}$$

$$\text{MAE}(u^*_{\text{gt}}, u^*_{\text{pred}}) = \frac{1}{N}\sum_{i=1}^{N}\left|u^*_{\text{gt},i} - u^*_{\text{pred},i}\right|. \tag{30}$$

To measure the discrepancy relative to the signal energy, we compute the **Relative Error (RelError)** based on the L_2 norm ratio. This metric is scale-invariant and particularly useful for comparing performance across samples with varying energy levels:

$$\text{RelError}(u^*_{\text{gt}}, u^*_{\text{pred}}) = \frac{\|u^*_{\text{gt}} - u^*_{\text{pred}}\|_2}{\|u^*_{\text{gt}}\|_2} = \sqrt{\frac{\sum_{i=1}^{N}(u^*_{\text{gt},i} - u^*_{\text{pred},i})^2}{\sum_{i=1}^{N}(u^*_{\text{gt},i})^2}}. \tag{31}$$

Structural Fidelity. In seismic analysis, minimizing numerical error alone is insufficient, as slight phase shifts can induce large residuals even when structural alignment is correct. To capture physical consistency, we utilize Pearson Correlation (Corr) and Structural Similarity Index (SSIM).

Corr evaluates the linear dependence and phase synchronization between signals, independent of absolute amplitude scaling:

$$\text{Corr}(u^*_{\text{gt}}, u^*_{\text{pred}}) = \frac{\sum_{i=1}^{N}(u^*_{\text{gt},i} - \mu_{gt})(u^*_{\text{pred},i} - \mu_{pred})}{\sqrt{\sum_{i=1}^{N}(u^*_{\text{gt},i} - \mu_{gt})^2}\sqrt{\sum_{i=1}^{N}(u^*_{\text{pred},i} - \mu_{pred})^2}}. \tag{32}$$

SSIM provides a holistic assessment of perceptual quality by modeling the inter-dependencies between spatially adjacent pixels, quantifying similarity in luminance, contrast, and structure:

$$\text{SSIM}(u^*_{\text{gt}}, u^*_{\text{pred}}) = \frac{(2\mu_{gt}\mu_{pred} + C_1)(2\sigma_{gt,pred} + C_2)}{(\mu^2_{gt} + \mu^2_{pred} + C_1)(\sigma^2_{gt} + \sigma^2_{pred} + C_2)}. \tag{33}$$

Here, μ and σ represent the mean and standard deviation, $\sigma_{gt,pred}$ is the covariance, and C_1, C_2 are stability constants. Higher SSIM indicates better preservation of geological interfaces and faults.

Figure 4 aggregation. The quantitative panels (a–c) use RMSE computed jointly over the real and imaginary components: $\mathrm{RMSE}_{\mathrm{joint}} = \sqrt{(\mathrm{RMSE}_{\mathrm{Re}}^2 + \mathrm{RMSE}_{\mathrm{Im}}^2)/2}$. Panel (b) pools 8 velocity models and 8 sources per model at each frequency ($n = 64$ model–source pairs). Its shading is $\pm 1.96 s/\sqrt{n}$ around the mean, where s is the sample standard deviation of these RMSE values. This summarizes variation across the pooled evaluation records; training-seed variability is reported separately in Appendix E.3.

Table 7: **Comparison on In-Distribution (ID) Datasets.** We compare CBNO against supervised and physics-informed baselines on OpenFWI ID benchmarks. Best results are highlighted in **bold**, and second-best results are underlined. Rel. Change is computed from the displayed values against the strongest baseline for each metric; positive values indicate improvement and negative values indicate degradation.

Dataset	Model	RMSE ↓	MAE ↓	RelError ↓	Corr ↑	SSIM ↑
CurveVel-A	PINN	0.0028	0.0018	0.8551	0.533	0.410
	FFNO	0.0010	0.0006	0.2655	0.966	0.876
	MgNO	0.0007	0.0004	0.1799	**0.982**	**0.957**
	DPOT	0.0010	0.0005	0.2456	0.963	0.924
	PINO	0.0005	0.0003	0.2338	0.969	0.889
	Ours	**0.0003**	**0.0002**	**0.1573**	0.979	0.924
	Rel. Change	40.00%	33.33%	12.56%	-0.31%	-3.45%
CurveVel-B	PINN	0.0033	0.0020	0.8575	0.523	0.411
	FFNO	0.0020	0.0012	0.3991	0.906	0.762
	MgNO	0.0014	0.0008	0.2832	0.956	0.880
	DPOT	0.0017	0.0010	0.3344	0.941	0.851
	PINO	0.0010	0.0006	0.3090	0.954	0.843
	Ours	**0.0002**	**0.0001**	**0.1071**	**0.989**	**0.952**
	Rel. Change	80.00%	83.33%	62.18%	3.45%	8.18%
Style-A	PINN	0.0046	0.0028	0.8954	0.463	0.431
	FFNO	0.0031	0.0018	0.4648	0.888	0.775
	MgNO	0.0026	0.0015	0.3814	0.922	0.839
	DPOT	0.0031	0.0018	0.4700	0.885	0.790
	PINO	0.0008	0.0004	0.2226	0.972	0.907
	Ours	**0.0005**	**0.0002**	**0.1432**	**0.984**	**0.941**
	Rel. Change	37.50%	50.00%	35.67%	1.23%	3.75%
Style-B	PINN	0.0044	0.0028	0.8910	0.472	0.423
	FFNO	0.0028	0.0017	0.4354	0.896	0.788
	MgNO	0.0022	0.0012	0.3249	0.939	0.887
	DPOT	0.0027	0.0016	0.4157	0.909	0.833
	PINO	0.0010	0.0006	0.2418	0.971	0.903
	Ours	**0.0004**	**0.0002**	**0.1285**	**0.987**	**0.947**
	Rel. Change	60.00%	66.67%	46.86%	1.65%	4.87%
FlatFault-A	PINN	0.0016	0.0010	0.9928	0.418	0.302
	FFNO	0.0007	0.0004	0.3710	0.923	0.748
	MgNO	0.0008	0.0004	0.3906	0.909	0.822
	DPOT	0.0008	0.0005	0.4339	0.892	0.800
	PINO	0.0007	0.0004	0.5018	0.896	0.753
	Ours	**0.0002**	**0.0001**	**0.1997**	**0.973**	**0.906**
	Rel. Change	71.43%	75.00%	46.17%	5.42%	10.22%
FlatFault-B	PINN	0.0039	0.0024	0.9010	0.427	0.404
	FFNO	0.0022	0.0013	0.3946	0.914	0.810
	MgNO	0.0017	0.0010	0.2878	0.948	0.892
	DPOT	0.0024	0.0014	0.4160	0.898	0.812
	PINO	0.0008	0.0005	0.2299	0.974	0.903
	Ours	**0.0004**	**0.0002**	**0.1375**	**0.986**	**0.939**
	Rel. Change	50.00%	60.00%	40.19%	1.23%	3.99%

E Additional Results

To provide a comprehensive evaluation, we report detailed quantitative results on six in-distribution (ID) OpenFWI subsets in Table 7, encompassing diverse geological styles and fault configurations. Our assessment utilizes both pixel-wise error metrics (RMSE, MAE, RelError) and structure-aware measures (Corr, SSIM) to jointly evaluate numerical fidelity and structural consistency. Qualitatively, Fig. 9 compares seismic shot gathers, demonstrating how improved forward modeling accuracy

translates into cleaner residuals and superior preservation of weak reflection events. Furthermore, Fig. 10 visualizes our physics-aware up/down wavefield decomposition, which effectively isolates reflection-dominated energy to support reflection-aware inversion.

E.1 Robustness to Observation Noise and Receiver Dropout

Using the protocol of Table 2, we evaluate CBNO under Gaussian observation noise and receiver dropout, keeping the pretrained model and all other inversion settings unchanged. We add Gaussian noise to frequency-domain records at signal-to-noise ratios (SNRs) of 30, 20, and 10 dB. Separately, we randomly remove 10%, 20%, or 30% of receivers from the acquisition geometry. Table 8 reports the resulting inversion metrics; the clean reference corresponds to the CBNO OOD entry in Table 2.

Table 8: **CBNO inversion under observation noise and receiver dropout.** Noise and dropout are evaluated separately using the Table 2 protocol.

Setting	RMSE↓	RelError↓	SSIM↑
Clean	410.7400	0.1293	0.515
Noise 30 dB	411.4876	0.1295	0.485
Noise 20 dB	434.4249	0.1375	0.353
Noise 10 dB	560.0595	0.1813	0.134
Dropout 10%	450.6597	0.1432	0.438
Dropout 20%	461.9930	0.1471	0.413
Dropout 30%	467.5452	0.1491	0.399

At 30 dB, RMSE and RelError remain nearly unchanged, but both deteriorate as SNR decreases, with substantial degradation at 10 dB, where SSIM falls from 0.515 to 0.134. At 30% receiver dropout, RMSE and RelError increase by 13.8% and 15.3%, respectively, while SSIM is 0.399. These results show gradual degradation under receiver sparsity and sensitivity to severe noise. The tests cover observation noise and acquisition sparsity in synthetic acoustic settings, not boundary-condition errors, broader modeling mismatches, or field data.

E.2 Additional Controlled Ablations

Table 9 separates supervision and backbone comparisons for forward wavefield prediction. In the supervision group, we keep the Full CBNO architecture and all other experimental settings fixed and replace only the training objective. CBS supervision gives a relative error of 0.1536, compared with 0.6427 for supervised regression and 0.9685 for a differential PINN loss. These results support the benefit of CBS supervision in the evaluated setting.

The backbone group holds the one-step CBS loss ($K = 1$) and the remaining experimental settings fixed while comparing Axial-AFNO with FNO. The relative errors are 0.2180 and 0.2707, respectively. Relative to FNO, the reduction is $(0.2707 - 0.2180)/0.2707 \approx 19.5\%$. The backbone reference uses $K = 1$; Full CBNO ($K = 5$) is the reference for the supervision comparison.

Table 9: **Controlled forward-modeling comparisons.** Each group varies the stated component with the other settings held fixed. Shading identifies the separate reference configuration for each group.

Configuration	RelError↓
Supervision objective (architecture and other settings fixed)	
CBS — Full CBNO, $K = 5$	0.1536
Supervised regression	0.6427
Differential PINN loss	0.9685
Backbone (matched one-step CBS, $K = 1$)	
Axial-AFNO	0.2180
FNO	0.2707

Inversion comparison and aggregation. Table 3(b) compares full-wavefield and upgoing-wavefield inversion on the same test samples, with identical initialization and optimization settings; only the inversion objective differs. Each metric is computed per test sample and averaged within each dataset. The reported OOD result is the equal-weight mean of the Marmousi and Overthrust dataset-level metrics. The Upgoing row reproduces the CBNO OOD result in Table 2: RMSE 410.74, MAE 288.66, RelError 0.1293, Corr 0.899, and SSIM 0.515. The clean reference in Table 8 uses the same result and follows the Table 2 protocol.

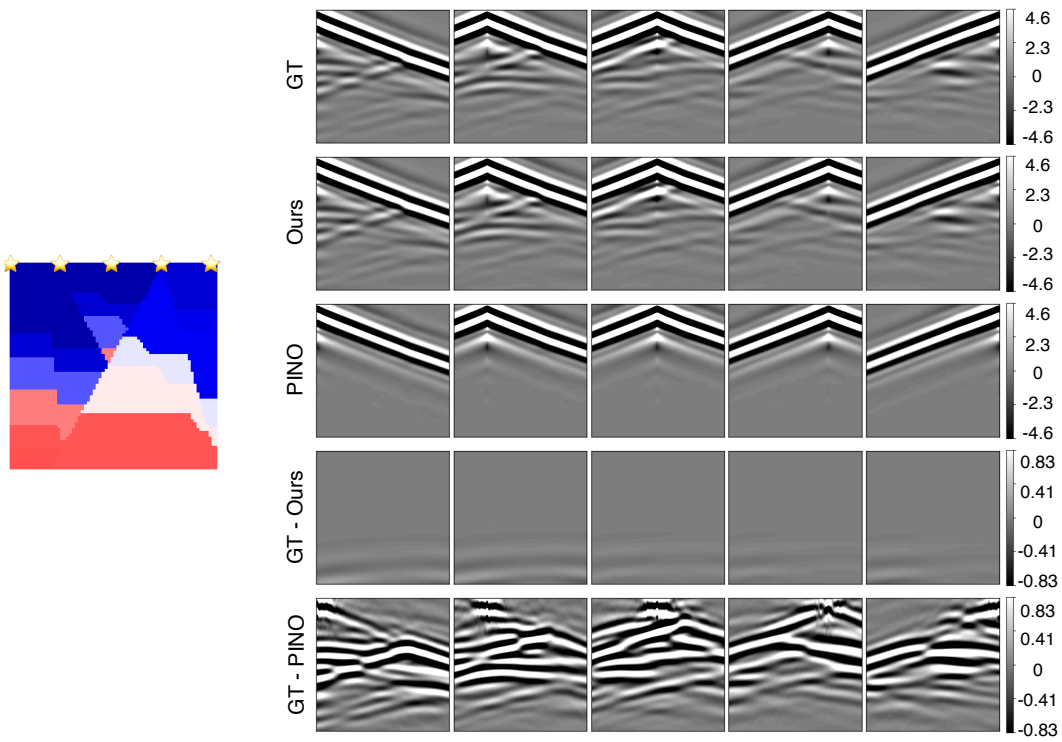

Figure 9: **Qualitative comparison of seismic shot gathers.** The rows from top to bottom represent: (1) Ground Truth seismic data; (2) Predictions by **CBNO (Ours)**; (3) Predictions by PINO; (4) Prediction error of CBNO (Ground Truth - CBNO); and (5) Prediction error of PINO (Ground Truth - PINO). **Analysis:** Our CBNO accurately reconstructs complex reflection events and hyperbolic patterns, resulting in a near-zero residual map (Row 4). Conversely, PINO fails to capture the fine-grained scattering details and weaker reflections, leading to significant structural residuals as shown in the last row.

E.3 Random-Seed Variability

We repeat Full CBNO training with seeds 42, 43, and 44, keeping all training and evaluation settings fixed. The resulting RelErrors are 0.1536, 0.1533, and 0.1537, respectively, giving a mean of 0.1535 and a sample standard deviation of 0.0002. These results indicate small variability across training seeds. Observation-noise and receiver-dropout results are reported separately in Appendix E.1.

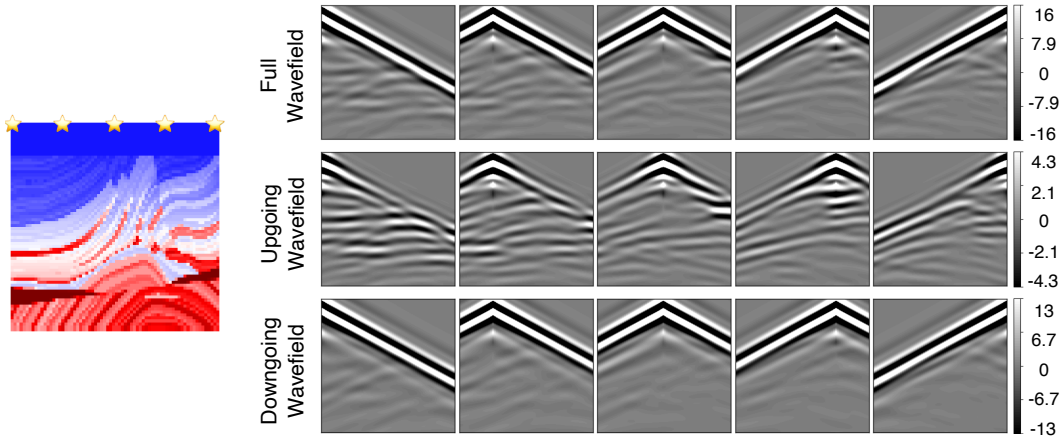

Figure 10: **Visualization of Up/Down wavefield separation.** The columns correspond to five different shot locations. **Top Row:** The full wavefield (u), containing both strong direct arrivals and weak reflections. **Middle Row:** The separated upgoing wavefield ($u^{\uparrow}$), which isolates the reflection energy used for our reflection-aware inversion. **Bottom Row:** The downgoing wavefield ($u^{\downarrow}$), dominated by direct arrivals.