WISER: multimodal variational inference for full-waveform inversion without dimensionality reduction

Ziyi Yin[1], Rafael Orozco[1], Felix J. Herrmann[1]
[1]Georgia Institute of Technology

Abstract

We present a semi-amortized variational inference framework designed for computationally feasible uncertainty quantification in 2D full-waveform inversion to explore the multimodal posterior distribution without dimensionality reduction. The framework is called **WISER**, short for full-**W**aveform variational **I**nference via **S**ubsurface **E**xtensions with **R**efinements. WISER leverages the power of generative artificial intelligence to perform approximate amortized inference that is low-cost albeit showing an amortization gap. This gap is closed through non-amortized refinements that make frugal use of acoustic wave physics. Case studies illustrate that WISER is capable of full-resolution, computationally feasible, and reliable uncertainty estimates of velocity models and imaged reflectivities.

1 Introduction

Full-waveform inversion (FWI) aims to estimate unknown multi-dimensional ($D \geq 2$) velocity models, denoted as $\mathbf{x}$, from noisy seismic data, $\mathbf{y}$, by inverting the nonlinear forward operator, $\mathcal{F}$, which relates $\mathbf{x}$ and $\mathbf{y}$ via $\mathbf{y} = \mathcal{F}(\mathbf{x}) + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ measurement noise [1]. FWI poses significant challenges, as it requires solving a high-dimensional, non-convex, and ill-posed inverse problem, with a computationally demanding forward operator in multiple dimensions. In addition, the inherent nonuniqueness of FWI results leads to multiple possible Earth models compatible with the observed data, underscoring the need for uncertainty quantification (UQ) to handle this multimodality.

The trade-off between accuracy and computational cost is a critical consideration in any high-dimensional inference routine with expensive forward operators [2], including FWI. To circumvent the costs associated with global optimization, several approaches have attempted localized UQ [3–6] based on the Laplace approximation. However, these approaches may not capture the full complexities of multimodal parameter spaces. In contrast, a Bayesian inference approach offers a costly but comprehensive resolution of the posterior distribution, $p(\mathbf{x} \mid \mathbf{y})$.

Bayesian inference algorithms are broadly categorized into two groups. The first, sampling-based methods, like Markov-chain Monte Carlo [McMC, 7], struggle with high-dimensional parameter spaces. To meet this challenge, they often rely on too restrictive low-dimensional parameterizations to reduce the number of sampling iterations [8–13], which could bias the inference results, rendering them impractical for multi-D UQ studies especially when solutions are nonunique.

The second category, optimization-based methods, like variational inference [VI, 14], seek to approximate the posterior distribution using classes of known parameterized distributions. VI can be subdivided into amortized and non-amortized methods. Amortized VI involves a computationally intensive offline training phase, leveraging advances in generative artificial intelligence (genAI), particularly with models like conditional diffusion [15] and conditional normalizing flows [CNFs, 16]. After training, amortized VI provides rapid sampling during inference [17, 18], exemplified by

---

[1]Correspondence: ziyi.yin@gatech.edu

the WISE framework [19] for multi-D FWI problems. However, these methods may suffer from an *amortization gap* — implying that the amortized networks may only deliver suboptimal inference for a single observation at inference time, particularly when trained with limited examples or when there exists a discrepancy between training and inference [20]. Conversely, non-amortized VI dedicates entire computational resources to the online inference [21–23]. They result in more accurate inference, but the costly optimization has to be carried out repeatedly for new observations, and integrating realistic priors can be challenging since the prior needs to be embedded involving density evaluations [24].

This paper introduces WISER as a semi-amortized VI framework to facilitate computationally feasible and reliable UQ for multi-D FWI without dimensionality reduction. Building on WISE, we train CNFs for efficient, suboptimal amortized inference, but then follow up with a crucial refinement step that only needs frugal use of the forward operator and its gradient. The refinement step aligns the posterior samples with the observation during inference, effectively bridging the amortization gap and enhancing inference accuracy.

Our contributions are organized as follows. We begin by outlining WISER in Algorithm 1. We explore the algorithm by explaining amortized VI with WISE, followed by computationally feasible multi-D physics-based refinement. The performance of WISER is demonstrated through realistic synthetic 2D case studies using the Compass model [25], showcasing improvements over WISE for both in- and out-of-distribution scenarios.

2 Amortized VI with WISE (lines 1—20)

WISER starts with an offline training phase that leverages conditional generative models to approximate the posterior distribution. This is achieved by WISE [19], which involves generating a training dataset (lines 3—9 of Algorithm 1) and training the CNFs (line 11—12).

2.1 Dataset generation (lines 3—9)

We begin by drawing N velocity models from the prior distribution, denoted by $p(\mathbf{x})$ (line 5). For each sample, $\mathbf{x}^{(i)}$, we simulate the observed data, $\mathbf{y}^{(i)}$, by performing the wave modeling and adding a random noise term (lines 6—7). Next, we compute common-image gathers [CIGs, 26] for each observed data with an initial smooth 1D migration-velocity model, $\mathbf{x}_0$, which can be rather inaccurate. These CIGs, represented by $\overline{\mathbf{y}}^{(i)}$, are produced by applying the adjoint of the extended migration operator, $\overline{\nabla\mathcal{F}}(\mathbf{x}_0)^\top$, to the observed data. Using CIGs as the set of physics-informed summary statistics not only preserves information from the observed seismic data [27] but also enhances the training of CNFs in the next stage [28, 29], as they help to decode the wave physics, translating prestack data to the image (subsurface-offset) domain.

2.2 Network training (lines 11—12)

CNFs are trained with pairs of velocity models and CIGs via minimization of the objective in line 12. The symbol $f_{\boldsymbol{\theta}}$ denotes the CNFs, characterized by their network weights, $\boldsymbol{\theta}$, and the Jacobian, $\mathbf{J}_{f_{\boldsymbol{\theta}}}$. The term "normalizing" within CNFs implies their functionality to transform realizations of velocity models, $\mathbf{x}^{(i)}$, into Gaussian noise from a standard multivariate normal distribution (as defined by the ℓ_2 norm), conditioned on the summary statistics (CIGs).

2.3 Online inference (lines 14—20)

The aforementioned data generation and CNF training procedures conclude the offline training phase. During online inference, amortized VI is enabled by leveraging the inherent invertibility of CNFs. For a given observation, $\mathbf{y}_{\text{obs}}$, the online cost is merely generation of a single set of CIGs (line 16). Subsequently, the posterior samples are generated by applying the inverse of the CNFs to Gaussian noise realizations, conditioned on these CIGs (lines 18—19)[2].

---

[2]We slightly abuse the notation to assume h_ϕ as an identity operator here.

Algorithm 1 WISER: full-**W**aveform variational **I**nference via **S**ubsurface **E**xtensions with **R**efinements

---

1: **Offline training phase**
2:
3: **Dataset generation**
4: **for** $i = 1 : N$ **do**
5:      $\mathbf{x}^{(i)} \sim p(\mathbf{x})$
6:      $\boldsymbol{\epsilon}^{(i)} \sim p(\boldsymbol{\epsilon})$
7:      $\mathbf{y}^{(i)} = \mathcal{F}(\mathbf{x}^{(i)}) + \boldsymbol{\epsilon}^{(i)}$
8:      $\overline{\mathbf{y}}^{(i)} = \overline{\nabla \mathcal{F}}(\mathbf{x}_0)^\top \mathbf{y}^{(i)}$
9: **end for**
10:
11: **Network training**

12: $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \quad \dfrac{1}{N} \sum_{i=1}^{N} \left(\dfrac{1}{2} \| f_{\boldsymbol{\theta}} \left(\mathbf{x}^{(i)}; \overline{\mathbf{y}}^{(i)} \right) \|_2^2 - \log |\det \mathbf{J}_{f_{\boldsymbol{\theta}}}| \right)$

13:
14: **Online inference phase**
15:
16: $\overline{\mathbf{y}}_{\text{obs}} = \overline{\nabla \mathcal{F}}(\mathbf{x}_0)^\top \mathbf{y}_{\text{obs}}$
17: **for** $i = 1 : M$ **do**
18:      $\mathbf{z}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$
19:      $\mathbf{x}_i = f_{\boldsymbol{\theta}^*}^{-1} \left(h_{\boldsymbol{\phi}} \left(\mathbf{z}_i \right); \overline{\mathbf{y}}_{\text{obs}} \right)$
20: **end for**
21:
22: **Physics-based refinements**
23: **for** $ii = 1 : \text{maxiter}_1$ **do**
24:      **for** $i = 1 : M$ **do**
25:          $\mathbf{g}_i = \nabla_{\mathbf{x}_i} \left[\dfrac{1}{2\sigma^2} \| \mathcal{F}(\mathbf{x}_i) - \mathbf{y}_{\text{obs}} \|_2^2 + \dfrac{1}{2\gamma^2} \| \mathbf{x}_i - f_{\boldsymbol{\theta}^*}^{-1} \left(h_{\boldsymbol{\phi}} \left(\mathbf{z}_i \right); \overline{\mathbf{y}}_{\text{obs}} \right) \|_2^2 \right]$
26:          $\mathbf{x}_i = \mathbf{x}_i - \tau \mathbf{g}_i$
27:      **end for**
28:      **for** $iii = 1 : \text{maxiter}_2$ **do**
29:          $\mathcal{L}(\boldsymbol{\phi}) = \sum_{i=1}^{M} \left[\dfrac{1}{2\gamma^2} \| \mathbf{x}_i - f_{\boldsymbol{\theta}^*}^{-1} \left(h_{\boldsymbol{\phi}} \left(\mathbf{z}_i \right); \overline{\mathbf{y}}_{\text{obs}} \right) \|_2^2 + \dfrac{1}{2} \| h_{\boldsymbol{\phi}} \left(\mathbf{z}_i \right) \|_2^2 - \log \left| \det \mathbf{J}_{h_{\boldsymbol{\phi}}} \right| \right]$
30:          $\boldsymbol{\phi} \leftarrow \mathrm{ADAM}\left(\mathcal{L}(\boldsymbol{\phi})\right)$
31:      **end for**
32: **end for**
33:
34: **Output**: $\{ f_{\boldsymbol{\theta}^*}^{-1} \left(h_{\boldsymbol{\phi}} \left(\mathbf{z}_i \right); \overline{\mathbf{y}}_{\text{obs}} \right) \}_{i=1}^{M}$ as samples of $p(\mathbf{x}|\mathbf{y}_{\text{obs}})$

---

3 Physics-based refinements (lines 22—32)

Consider a single observation, $\mathbf{y}_{\text{obs}}$, and its corresponding posterior samples, $\mathbf{x}_i \sim p(\mathbf{x} \mid \overline{\mathbf{y}}_{\text{obs}})$. The latent representations generated by the trained CNFs, $\hat{\mathbf{z}}_i = f_{\boldsymbol{\theta}^*}(\mathbf{x}_i; \overline{\mathbf{y}}_{\text{obs}})$, may not conform exactly to the standard Gaussian distribution during inference. To address this issue, we follow Siahkoohi et al. [17] to apply latent space corrections to fine-tune the trained CNFs. This involves integrating a shallower, yet invertible, network[3], specifically trained to map realizations of true Gaussian noise to the corresponding latent codes, $\hat{\mathbf{z}}_i$. Adhering to a transfer learning approach, we maintain the weights of the trained CNFs while solely updating the weights of the shallower network by minimizing the following objective:

---

[3]For linear inverse problems in seismic imaging, Siahkoohi et al. [17] show that an elementwise scaling and shift mechanism is adequate to bridge the gap. However, given the complex, non-convex nature of FWI, we employ $h_{\boldsymbol{\phi}}$ as a generic invertible network.

$$\underset{\boldsymbol{\phi}}{\text{minimize}} \quad \mathbb{E}_{\mathbf{z} \sim \mathrm{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{2\sigma^2} \|\mathcal{F} \circ f_{\boldsymbol{\theta}^*}^{-1}\left(h_{\boldsymbol{\phi}}\left(\mathbf{z}\right); \overline{\mathbf{y}}_{\mathrm{obs}}\right) - \mathbf{y}_{\mathrm{obs}}\|_2^2 + \frac{1}{2}\|h_{\boldsymbol{\phi}}\left(\mathbf{z}\right)\|_2^2 - \log \left| \det \mathbf{J}_{h_{\boldsymbol{\phi}}} \right| \right].$$
$$(1)$$

Here, the refinement network, $h_{\boldsymbol{\phi}}$, mitigates the amortization gap by adjusting the latent variable $\mathbf{z}$ before feeding it to the inverse of the trained CNFs, $f_{\boldsymbol{\theta}^*}^{-1}$. Intuitively, minimizing the first terms ties the posterior samples closer to the observed data. The second and third terms prevent the corrected latent space from being far from the Gaussian distribution, which implicitly takes advantage of the prior information existing in the amortized training phase.

Equation 1 offers a fine-tuning approach that leverages the full multi-D wave physics to refine the amortized VI framework for a single observation at inference phase. However, it introduces notable computational demands because it necessitates the coupling of the modeling operator and the networks. Specifically, every update to the network weights, $\boldsymbol{\phi}$, requires costly wave modeling operations. Given that network training typically involves numerous iterations, these computational demands can render it impractical for realistic FWI applications.

To relieve this computational burden, we adopt a strategy from Siahkoohi et al. [30] to reformulate Equation 1 into a weak form by allowing the network output to be only weakly enforced (in an ℓ_2 sense) to be the corrected velocity models. The objective function for this weak formulation reads:

$$\underset{\mathbf{x}_{1:M},\boldsymbol{\phi}}{\text{minimize}} \quad \frac{1}{M}\sum_{i=1}^{N}\left[\frac{1}{2\sigma^2}\|\mathcal{F}(\mathbf{x}_i) - \mathbf{y}_{\mathrm{obs}}\|_2^2 + \frac{1}{2\gamma^2}\|\mathbf{x}_i - f_{\boldsymbol{\theta}^*}^{-1}\left(h_{\boldsymbol{\phi}}\left(\mathbf{z}_i\right); \overline{\mathbf{y}}_{\mathrm{obs}}\right)\|_2^2 + \frac{1}{2}\|h_{\boldsymbol{\phi}}\left(\mathbf{z}_i\right)\|_2^2 - \log \left|\det \mathbf{J}_{h_{\boldsymbol{\phi}}}\right|\right].$$
$$(2)$$

We strategically decouple the computationally expensive forward operator, $\mathcal{F}$, from the more cheap-to-evaluate networks, $f_{\boldsymbol{\theta}^*}$ and $h_{\boldsymbol{\phi}}$. This is achieved in a penalty form with the assumption that the misfit between the network outputs and the posterior samples adheres to a Gaussian distribution, $\mathrm{N}(\mathbf{0}, \gamma^2\mathbf{I})$, where γ is a hyperparameter dictating the trade-off between data misfit and regularization. Setting γ to infinity equates this weak formulation to the constrained formulation in Equation 1. This weak formulation also supports optimization strategies for updating the velocity models with physical constraints [31, 32] and multiscale optimization techniques [33].

WISER takes full computational advantage of this weak formulation by employing a nested loop structure. The outer loop is dedicated to updating M velocity models, $\mathbf{x}_i$, through costly gradient descent steps (lines 24—27 of Algorithm 1), while the inner loop (lines 28—31) focuses on more updates (with the ADAM optimizer) to network weights, $\boldsymbol{\phi}$, without computationally expensive physics modeling. To achieve a balance, we conduct $\mathrm{maxiter}_2 = 128$ iterations in the inner loop. After refinements, WISER first evaluates the refined network on the latent variables to obtain refined latent codes. Subsequently, the amortized network uses the refined codes conditioned on the CIGs to compute the corrected posterior samples (line 34).

4 Case studies

Evaluation of WISER is conducted through synthetic case studies utilizing 2D slices of the Compass model and 2D acoustic wave physics. The parameter of interest is discretized into 512×256 gridpoints with a spatial resolution of $12.5\,\mathrm{m}$, resulting in over 10^5 degrees of freedom. The forward operator, $\mathcal{F}$, simulates acoustic data with absorbing boundaries. A Ricker wavelet with a central frequency of $15\,\mathrm{Hz}$ and an energy cut below $3\,\mathrm{Hz}$ is employed. We use 512 sources towed at $12.5\,\mathrm{m}$ depth and 64 ocean-bottom nodes (OBNs) located at jittered sampled horizontal positions [34]. We employ source-receiver reciprocity during the modeling and sensitivity calculations. The observed data, $\mathbf{y}^{\mathrm{obs}}$, is perturbed with band-limited Gaussian noise to achieve a signal-to-noise ratio (S/N) of $12\,\mathrm{dB}$. The training of the CNFs uses $N = 800$ pairs of velocity models and CIGs. To demonstrate WISER's efficacy in mitigating the amortization gap, we compare results from WISE and WISER under two scenarios during inference:

(i) observed shot data is generated using an in-distribution velocity model with the same forward operator;

4

(ii) observed shot data is produced by an out-of-distribution (OOD) velocity model and also a slightly altered forward operator.

4.1 Case 1: in distribution

The ground-truth velocity model is an unseen 2D slice from the Compass model, shown in Figure 1a. Following Algorithm 1, we initiate WISER by drawing $M = 16$ Gaussian noise realizations to create the initial set of 16 velocity models, depicted in Figure 1b. To minimize computational demands, stochastic gradients [35] are calculated in line 25 of Algorithm 1. Each particle's gradient is estimated using only 1 randomly selected OBN gather from the observed data. We also add box constraints to the velocity models to restrict their range to 1.48 to $4.44\,\mathrm{km/s}$. Following $\mathrm{maxiter}_1 = 80$ outer iterations—equivalent to 20 data passes or 2560 PDE solves[4]—we obtain the posterior samples from WISER in Figure 1c.

4.1.1 Observations

The conditional mean estimate (CM) from WISE lacks finer details, particularly beneath the unconformity at depths below $2.4\,\mathrm{km}$ (in red). This is attributed to the excessive variability in structural details of the posterior samples, visible on the right panel of Figure 1b.

In contrast, WISER generates more consistent and accurate posterior samples. In Figure 1c, the right panel shows that the uncertainty from WISER is reduced below the unconformity. The upper panel illustrates that the uncertainty is more focused on the dipping events at the unconformity, highlighting areas of poor illumination.

4.1.2 Impact on imaging

To assess the impact of uncertainty in velocity models on downstream tasks, we conduct high-frequency imaging using a Ricker wavelet with central frequency of $30\,\mathrm{Hz}$ and compare the imaged reflectivities derived from the posterior samples of both WISE and WISER, shown in Figure 1d and Figure 1e, respectively.

The imaged reflectivities produced by CM from WISER exhibit superior continuity and a better correlation with the CM migration-velocity model, particularly noticeable in Figure 1g under the unconformity. Also, reflectivity samples produced by WISER demonstrate improved alignment among themselves compared to those produced by WISE. In addition, notable vertical shifts observed in the imaged reflectivities from WISE to WISER indicate significant adjustments in the positioning of subsurface reflectors, underlining the necessity of the refinement procedure for precisely estimating migration-velocity models that locate subsurface reflectors more accurately.

4.2 Case 2: out of distribution

To test the robustness and adaptability of WISER when faced with unexpected variations at inference, we also evaluate WISER's performance under OOD scenarios. We introduce alterations to the velocity model depicted in Figure 1a through an elementwise perturbation shown in Figure 2a. This manipulation modifies the velocity values across different depth levels, resulting in a significant shift in their statistical distribution, illustrated in Figure 2b. We use the perturbed velocity as the unseen ground-truth velocity model in this case study, shown in Figure 2c. To further expand the amortization gap, we modify the encoding of the forward operator by introducing a higher amplitude of band-limited Gaussian noise (S/N $0\,\mathrm{dB}$).

These complexities present substantial challenges for WISE, leading to biased inference results as depicted in Figure 2d. The yellow histograms in Figure 2b show that the velocity values of the posterior samples from WISE closely resemble those of the original velocity model, despite the different distribution of the ground-truth velocity model. This indicates that WISE tends to incorporate an inductive bias from the training samples. In WISER, we conduct $\mathrm{maxiter}_1 = 160$ outer iterations, using $M = 16$ particles and 1 OBN per gradient. We also employ the frequency

---

[4]1 PDE solve means solving the wave equation for a single source. A gradient requires 2 PDE solves (forward and adjoint).
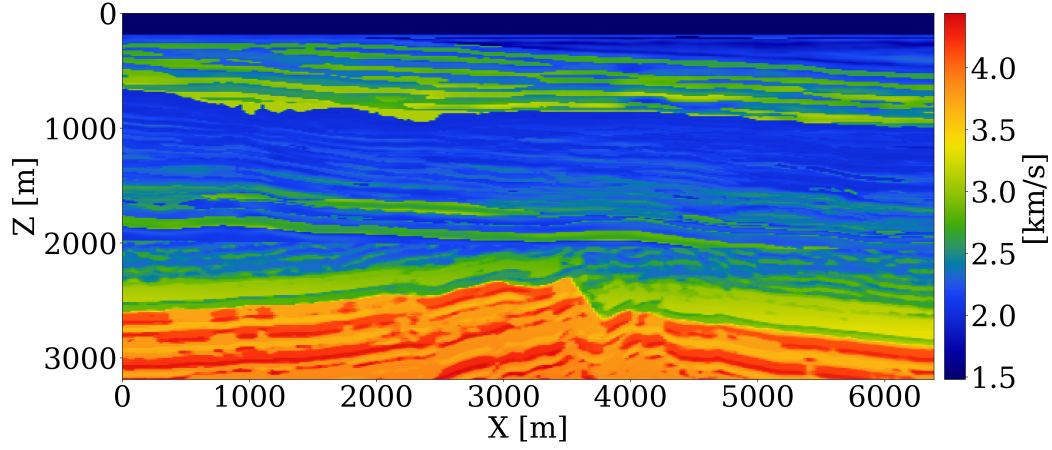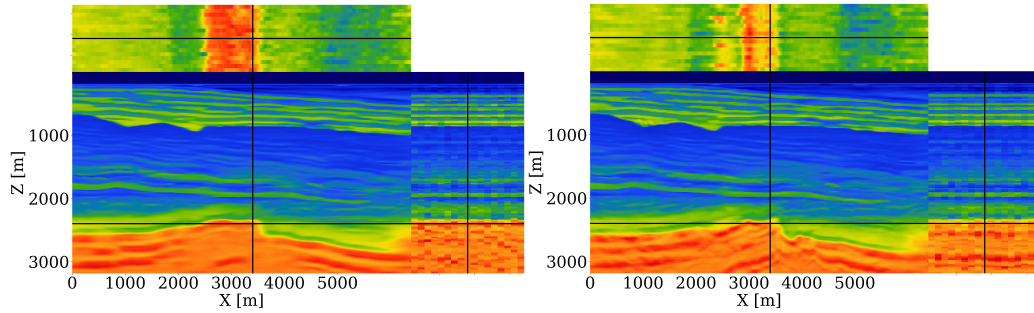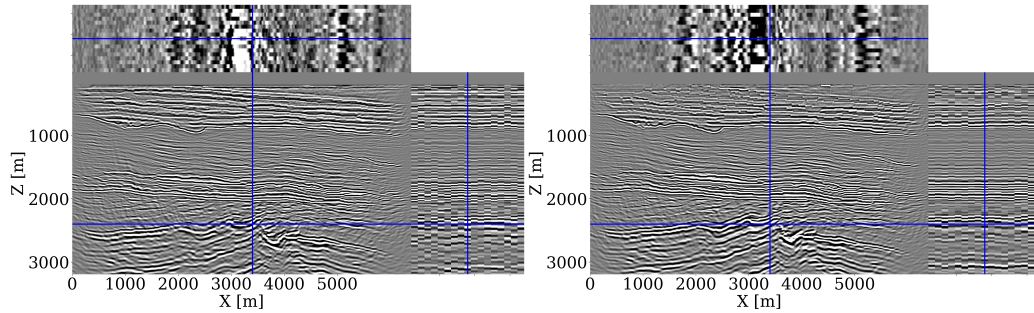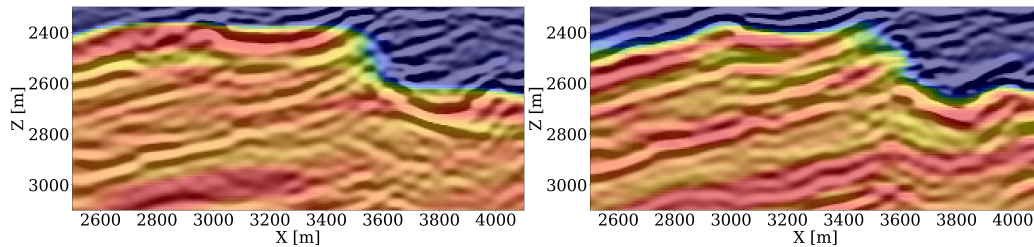
Figure 1: Comparison between WISE and WISER for an in-distribution case. *(a)* Unseen ground-truth velocity model. *(b)* Estimated velocity models from WISE. The conditional mean estimate (CM) is shown in the center. For posterior samples, horizontal traces at $Z = 2.7$ km and vertical traces at $X = 3.6$ km are displayed on the top and on the right, respectively. *(d)* Imaged reflectivity samples from WISE. *(f)* Zoom-in views of *(d)* overlaying on the CM of WISE. *(c)(e)(g)* are the counterparts from WISER, showcasing significant improvements.

continuation method [33] to compute the gradient in line 25 of Algorithm 1, transitioning gradually from low-frequency to high-frequency data. This results in 40 datapasses or 5120 PDE solves in total.

4.2.1 Observations

WISER produces more accurate posterior samples shown in Figure 2e. Furthermore, the statistical distribution of the velocity values in the WISER posterior samples (green histogram in Figure 2b) aligns better with the distribution of the unseen ground-truth velocity values (blue histogram in Figure 2b), demonstrating WISER's robustness against potential distribution shifts during inference.

To further showcase WISER's robustness under severe measurement noise, we compare a posterior sample from WISER (Figure 2k) to a velocity model estimated by FWI (Figure 2j), derived by minimizing only the data likelihood (the first term in line 25 of Algorithm 1), while starting from the same initial model as WISER. The FWI result is significantly impacted by noise, while the posterior samples from WISER remain relatively noise-free and capture all pertinent geological structures.

4.2.2 Impact on imaging

The imaging results from WISE (Figure 2f) and WISER (Figure 2g) reveal noticeable discrepancies in quality. The CM migration-velocity model from WISE leads to discontinuities in the imaged reflectivities, particularly at the horizontal layer around $1.8\,\mathrm{km}$ depth and more so below the unconformity. In contrast, the CM from WISER significantly improves the continuity of the imaged reflectivities across the entire seismic section. The imaged reflectivity samples from WISER not only show better consistency among themselves but also align more accurately with the estimated CM migration-velocity model, particularly visible in Figure 2i.

5 Discussion and conclusions

The primary contribution of WISER is to leverage both genAI and physics to achieve a semi-amortized VI framework for scalable ($\mathrm{D} \geq 2$) and reliable UQ for FWI even in situations where local approximations are unsuitable. At its core, WISER harnesses the strengths of both amortized and non-amortized VI: the amortized posterior obtained through offline training provides a low-fidelity but fast mapping, and the physics-based refinements offer reliable and accurate inference. Both approaches benefit from information preservation exhibited by CIGs, rendering our inference successful where conventional FWI fails due to cycle skipping.

Compared to McMC methods that rely on low-dimensional parameterizations, WISER does not impose intrinsic dimensionality reductions or simplifications of the forward model. Therefore, WISER is capable of delivering full-resolution UQ for realistic multi-D FWI problems. CNFs are primed for large-scale 3D inversion thanks to their invertibility, which allows for memory-efficient training and inference [36].

Compared to non-amortized VI methods, WISER also requires significantly less computational resources during inference. This is because WISE, as a precursor of WISER, already provides near-accurate posterior samples, making the refinement procedure computationally feasible. Zhang et al. [22] show that non-amortized VI, without access to realistic prior information, requires $\mathrm{O}(10^6)$ to $\mathrm{O}(10^8)$ PDE solves, while WISER only needs $\mathrm{O}(10^3)$ PDE solves. Apart from the computational cost reduction, WISER ensures that the posterior samples realistically resemble Earth models, thanks to the integration of the conditional prior information from WISE. Contrary to non-amortized VI, which needs density evaluations to embed the prior (i.e., $p(\mathbf{x})$ in line 5 of Algorithm 1) to produce realistic Earth models, WISER only needs access to samples of the prior distribution (i.e., $\mathbf{x}^{(i)}$ in line 5).

Opportunities for future research remain. Although case 2 demonstrates the robustness of WISER concerning OOD issues, these issues could be fundamentally addressed by diversifying the training set of WISE through a foundation model [37]. Also, our OOD case study has not yet explored scenarios where the likelihood term is more pathologically misspecified, such as the presence of unremoved shear wave energy outside the range of the forward operator, which calls for further investigations. Our approach will also benefit from calibration of the estimated posterior on which the authors report elsewhere, including application of WISE(R) in 3D.
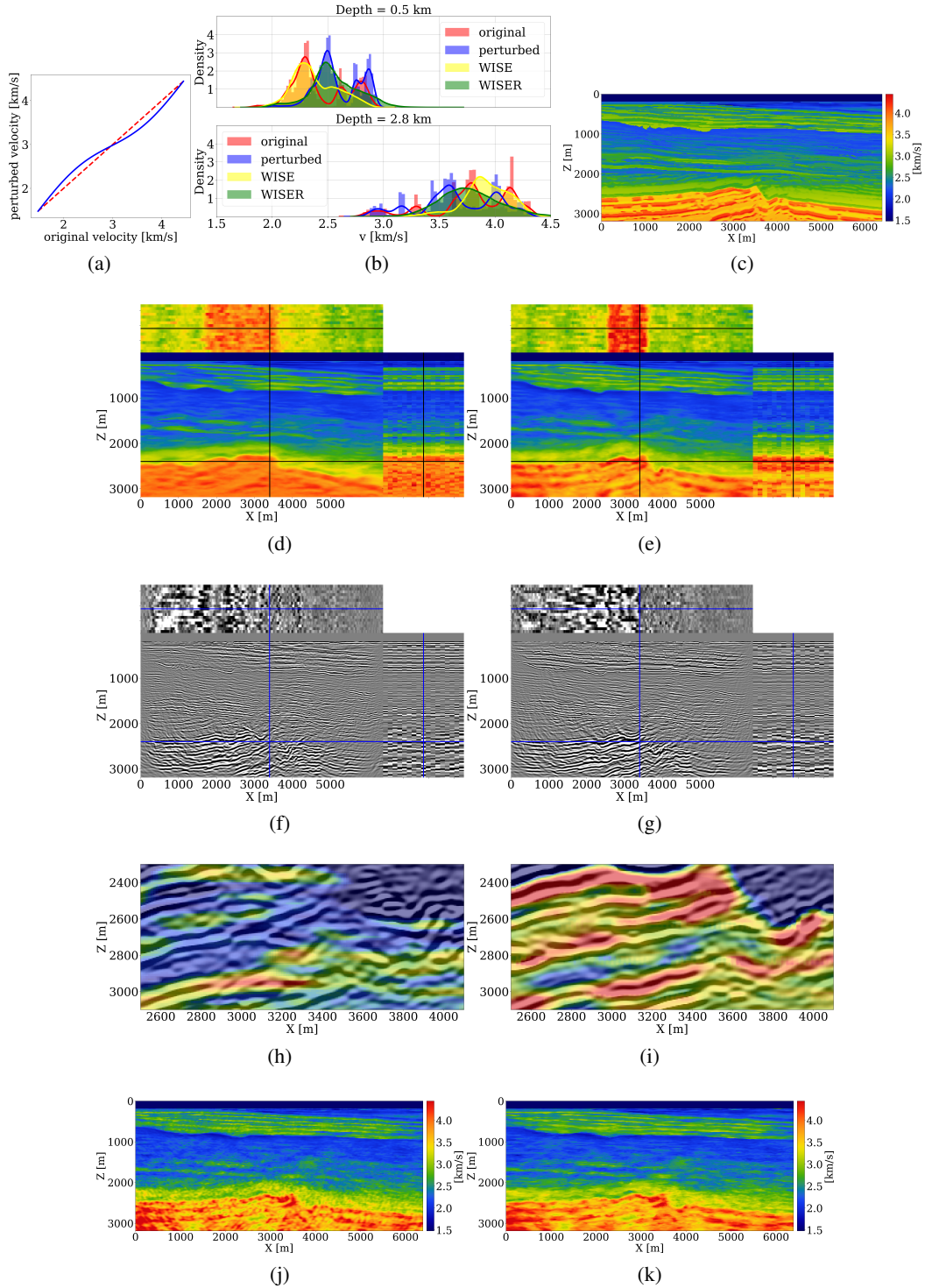
Figure 2: OOD case study. *(a)* Curves for velocity-value perturbations; *(b)* histograms of values at the depth of $0.5\,\text{km}$ and $2.8\,\text{km}$ in the original velocity model (Figure 1a), perturbed velocity model (Figure 2c), posterior samples of WISE, and WISER, shown in red, blue, yellow and green color, respectively. *(c)—(i)* Comparison between WISE and WISER. The ordering remains the same as in Figure 1. *(j)* FWI result starting with a posterior sample from WISE. *(k)* A posterior sample from WISER.

In conclusion, this paper sets the stage for deploying genAI models to facilitate high-dimensional Bayesian inference with computationally intensive multi-D forward operators. Deep learning and AI have been criticized for their reliance on realistic training samples, but WISER alleviates this reliance and still offers computationally feasible and reliable inference through a blend of offline training and online frugal physics-based refinements, preparing our approach for large 3D deployment.

6 Acknowledgement

7 Related material

The scripts to reproduce the experiments are available on the SLIM GitHub page https://github.com/slimgroup/WISER.jl.

8 Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to refine sentence structures and improve the readability of the manuscript. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] J. Virieux and S. Operto. An overview of full-waveform inversion in exploration geophysics. *GEOPHYSICS*, 74(6):WCC1–WCC26, 11 2009. doi: 10.1190/1.3238367. URL http://dx.doi.org/10.1190/1.3238367.

[2] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[3] Zhilong Fang, Curt Da Silva, Rachel Kuske, and Felix J Herrmann. Uncertainty quantification for inverse problems with weak partial-differential-equation constraints. *Geophysics*, 83(6): R629–R647, 2018.

[4] Scott D Keating and Kristopher A Innanen. Null-space shuttles for targeted uncertainty analysis in full-waveform inversion. *Geophysics*, 86(1):R63–R76, 2021.

[5] Muhammad Izzatullah, Matteo Ravasi, and Tariq Alkhalifah. Physics reliable frugal uncertainty analysis for full waveform inversion. *arXiv preprint arXiv:2305.07921*, 2023.

[6] 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.

[7] Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American statistical Association*, 91(434):883–904, 1996.

[8] Zhilong Fang, Hongjian Fang, and Laurent Demanet. Deep generator priors for bayesian seismic inversion. In *Advances in geophysics*, volume 61, pages 179–216. Elsevier, 2020.

[9] Zhouji Liang, Florian Wellmann, and Omar Ghattas. Uncertainty quantification of geologic model parameters in 3d gravity inversion by hessian-informed markov chain monte carlo. *Geophysics*, 88(1):G1–G18, 2023.

[10] Xiaolong Wei, Jiajia Sun, and Mrinal K Sen. Quantifying uncertainty of salt body shapes recovered from gravity data using trans-dimensional markov chain monte carlo sampling. *Geophysical Journal International*, 232(3):1957–1978, 2023.

[11] Xiaolong Wei, Jiajia Sun, and Mrinal Sen. 3d monte carlo geometry inversion using gravity data. *Geophysics*, 89(3):1–62, 2024.

[12] Xiaolong Wei, Jiajia Sun, and Mrinal K Sen. Reconstruction of multiple target bodies using transdimensional bayesian inversion with different constraints. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[13] Nirmit Dhabaria and Satish C Singh. Hamiltonian monte carlo based elastic full-waveform inversion of wide-angle seismic data. *Geophysical Journal International*, 237(3):1384–1399, 2024.

[14] Xin Zhang, Muhammad Atif Nawaz, Xuebin Zhao, and Andrew Curtis. An introduction to variational inference in geophysical inverse problems. In *Advances in geophysics*, volume 62, pages 73–140. Elsevier, 2021.

[15] Lorenzo Baldassari, Ali Siahkoohi, Josselin Garnier, Knut Solna, and Maarten V de Hoop. Conditional score-based diffusion models for bayesian inference in infinite dimensions. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

[17] Ali Siahkoohi, Gabrio Rizzuti, Rafael Orozco, and Felix J. Herrmann. Reliable amortized variational inference with physics-based latent distribution correction. *GEOPHYSICS*, 88 (3):R297–R322, 04 2023. doi: 10.1190/geo2022-0472.1. URL http://dx.doi.org/10.1190/geo2022-0472.1.

[18] Rafael Orozco, Mathias Louboutin, Ali Siahkoohi, Gabrio Rizzuti, Tristan van Leeuwen, and Felix Johan Herrmann. Amortized normalizing flows for transcranial ultrasound with uncertainty quantification. In *Medical Imaging with Deep Learning*, 2023. URL https://openreview.net/forum?id=LoJG-lUIlk.

[19] Ziyi Yin, Rafael Orozco, Mathias Louboutin, and Felix J Herrmann. Wise: full-waveform variational inference via subsurface extensions. *Geophysics*, 89(4):1–31, 2024.

[20] Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, pages 3403–3412. PMLR, 2018.

[21] Xuebin Zhao, Andrew Curtis, and Xin Zhang. Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, 228(1):213–239, 2022.

[22] Xin Zhang, Angus Lomas, Muhong Zhou, York Zheng, and Andrew Curtis. 3-d bayesian variational full waveform inversion. *Geophysical Journal International*, 234(1):546–561, 2023.

[23] Xin Zhang and Andrew Curtis. Bayesian variational time-lapse full waveform inversion. *Geophysical Journal International*, 237(3):1624–1638, 2024.

[24] Jakob Kruse, Gianluca Detommaso, Ullrich Köthe, and Robert Scheichl. Hint: Hierarchical invertible neural transport for density estimation and bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8191–8199, 2021.

[25] CE Jones, JA Edgar, JI Selvage, and H Crook. Building complex synthetic models to evaluate acquisition geometries and velocity inversion technologies. In *74th EAGE Conference and Exhibition incorporating EUROPEC 2012*, pages cp–293. European Association of Geoscientists & Engineers, 2012.

[26] Jie Hou and William W. Symes. Accelerating extended least-squares migration with weighted conjugate gradient iteration. *GEOPHYSICS*, 81(4):S165–S179, 07 2016. doi: 10.1190/geo2015-0499.1. URL http://dx.doi.org/10.1190/geo2015-0499.1.

[27] Fons ten Kroode. An omnidirectional seismic image extension. *Inverse Problems*, 39(3):035003, 2023.

[28] Stefan T Radev, Ulf K Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe. Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*, 33(4):1452–1466, 2020.

[29] Rafael Orozco, Ali Siahkoohi, Gabrio Rizzuti, Tristan van Leeuwen, and Felix Herrmann. Adjoint operators enable fast and amortized machine learning based bayesian uncertainty quantification. *Medical Imaging 2023: Image Processing*, 04 2023. doi: 10.1117/12.2651691. URL http://dx.doi.org/10.1117/12.2651691.

[30] Ali Siahkoohi, Gabrio Rizzuti, and Felix J Herrmann. Weak deep priors for seismic imaging. In *SEG Technical Program Expanded Abstracts 2020*, pages 2998–3002. Society of Exploration Geophysicists, 2020.

[31] Ernie Esser, Lluis Guasch, Tristan van Leeuwen, Aleksandr Y Aravkin, and Felix J Herrmann. Total variation regularization strategies in full-waveform inversion. *SIAM Journal on Imaging Sciences*, 11(1):376–406, 2018.

[32] Bas Peters, Brendan R Smithyman, and Felix J Herrmann. Projection methods and applications for seismic nonlinear inverse problems with multiple constraints. *Geophysics*, 84(2):R251–R269, 2019.

[33] Carey Bunks, Fatimetou M Saleck, S Zaleski, and G Chavent. Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473, 1995.

[34] Gilles Hennenfent and Felix J Herrmann. Simply denoise: Wavefield reconstruction via jittered undersampling. *Geophysics*, 73(3):V19–V28, 2008.

[35] Felix J Herrmann, Ian Hanlon, Rajiv Kumar, Tristan van Leeuwen, Xiang Li, Brendan Smithyman, Haneet Wason, Andrew J Calvert, Mostafa Javanmehri, and Eric Takam Takougang. Frugal full-waveform inversion: From theory to a practical algorithm. *The Leading Edge*, 32(9): 1082–1092, 2013.

[36] Rafael Orozco, Philipp Witte, Mathias Louboutin, Ali Siahkoohi, Gabrio Rizzuti, Bas Peters, and Felix J Herrmann. Invertiblenetworks.jl: A julia package for scalable normalizing flows. *arXiv preprint arXiv:2312.13480*, 2023.

[37] Hanlin Sheng, Xinming Wu, Xu Si, Jintao Li, Sibio Zhang, and Xudong Duan. Seismic foundation model (sfm): a new generation deep learning model in geophysics. *arXiv preprint arXiv:2309.02791*, 2023.