Response to Reivewer's Comments on Manuscript ID: egusphere-2026-1087

Dear Pouria Behnoudfar:

Thank you for your letter and for regarding our manuscript entitled **"Latent-Compression-Free Generative Diffusion with Geological Priors and Geophysical Regularization for Implicit Structural Modeling" (ID: egusphere-2026-1087)**. We sincerely appreciate the time and effort you have dedicated to reviewing our work.

Response to Comment:

General Comments

1. Uncertainty quantification. The manuscript frames LFD as a generative model for the distribution of implicit structural models, which implies that the method should be capable of representing and communicating uncertainty across plausible structural realizations. This is, after all, one of the principal motivations for using a generative rather than a deterministic model in this setting. However, the manuscript does not clearly characterize the uncertainty produced by the method: it is not clear whether multiple samples from LFD given the same conditioning (horizons/faults) are shown to produce meaningfully diverse, geologically plausible realizations, how the spread of these realizations compares to the known or expected geological uncertainty in the synthetic/real test cases, and whether the prior-guided losses (horizon loss, bending-energy loss), by construction, narrow the ensemble in ways that could underrepresent genuine structural uncertainty. Given that uncertainty representation is central to the value proposition of a diffusion/flow-matching approach over deterministic implicit modeling, I recommend that the authors add an explicit quantitative and visual analysis of ensemble spread/variability (e.g., pixel-wise standard deviation maps or probability-of-occurrence maps across multiple samples) and discuss how the imposed priors affect this spread.

Response:
Thank you for your valuable suggestion. We agree that, as a generative approach, LFD should be able to represent and communicate uncertainty in the generated structural models, and that the original manuscript lacked a quantitative analysis in this respect. To address this concern, we have added a new subsection to the Experiments section (Sect. 3.5, "Ablation Study on Uncertainty: Effects of Conditioning Sparsity and Prior-Guided Regularization Strength"), which systematically extends the uncertainty analysis along two dimensions:

(1) Quantitative characterization of multi-sample generation. With the fault and horizon conditioning fixed, we generate 20 independent realizations for each conditioning configuration by sampling different initial Gaussian noise fields, and compute the pixel-wise variance across these realizations to quantify uncertainty. We further systematically reduce the number of input horizons from 6 to 4, 2, and 1 to examine how the density of conditioning information affects realization spread. As shown in the newly added Figure 4, when horizon constraints are abundant (6 horizons), the variance remains uniformly low across the model domain, indicating that dense conditioning

effectively anchors the generation. As the number of horizons decreases, high-variance regions progressively expand, first emerging in unconstrained areas between horizons and eventually spanning the majority of the model domain when only a single horizon is provided. The mean variance increases monotonically from 0.0013 (6 horizons) to 0.0073 (1 horizon), quantitatively confirming that realization spread is directly governed by conditioning density. These results demonstrate that LFD generates a data-consistent distribution: when conditioning information is abundant, the model converges to consistent solutions; when conditioning information is sparse, the model appropriately broadens its output distribution to reflect increased geological ambiguity, rather than collapsing to a single, overconfident prediction.

(2) Effect of the prior-guided regularization strength on the ensemble. To directly address whether the prior-guided losses narrow the ensemble by construction, we further investigate how the strength of the proposed fault-aware bending-energy regularization affects the realization ensemble. For each bending-energy weight $\lambda_{Bend} \in \{0,0.1,1\}$ ($\lambda_{Bend}=0.1$ being the default used throughout this paper), we train a separate model under the same training settings and evaluate it on the full validation set (3200 cases). For each case, we extract the central horizon of the implicit field together with the corresponding fault mask as the conditioning input, and generate 20 realizations from different noise seeds. We report the resulting pixel-wise Mean Squared Error (MSE), the Horizon Consistency Error (HCE-6, evaluated on 6 horizons uniformly extracted from the ground-truth implicit field), and the mean realization variance, averaged over the validation set, in the added Table 4.

As shown in Table 4, MSE and HCE-6 do not vary monotonically with λ_{Bend} , but both are consistently lower with the bending-energy prior enabled than without it, indicating that the regularization yields realizations that are more stable and closer to the reference model, rather than drifting into implausible solutions. In particular, $\lambda_{Bend}=0.1$ achieves the lowest pixel-wise error among the three settings. At the same time, the prior does not reduce realization variance—the strongest setting ($\lambda_{Bend}=1$) yields the highest variance among the three, showing that this improved stability does not come at the cost of the model's ability to represent structural uncertainty. Taken together with the results in (1), these findings indicate that the prior-guided losses do not narrow the ensemble in ways that would underrepresent genuine structural uncertainty.

3.5 Ablation Study on Uncertainty: Effects of Conditioning Sparsity and Prior-Guided Regularization Strength

To investigate how the number of input horizon constraints affects generation quality and realization variability, we conduct an ablation study in which the number of input horizons is systematically reduced from 6 to 4, 2, and 1 (The top row in Figure 4). For each configuration, we generate 20 independent realizations by sampling different initial Gaussian noise fields while keeping the fault and horizon conditioning fixed, and compute the pixel-wise variance across realizations to quantify uncertainty.

The middle row of Figure 4 shows the mean implicit structural models over 20 realizations generated under each sparsity level. As the number of input horizons decreases, the generated models remain geologically plausible but exhibit increasing variability in stratigraphic geometry,

particularly in regions far from the remaining horizon constraints. This reflects the model's learned prior filling in the unconstrained space with a broader range of geologically consistent solutions.

The bottom row of Figure 4 shows the spatial distribution of pixel-wise variance across 20 realizations for each horizon configuration. With 6 input horizons, variance is uniformly low across the model domain, indicating that dense conditioning effectively anchors the generation. As the number of horizons is reduced to 4, 2, and finally 1, high-variance regions progressively expand, concentrating first in areas between horizons and eventually spanning the majority of the model domain. This spatial pattern is consistent with the intuition that uncertainty grows where conditioning data are absent. The mean variance plot (bottom-left panel of Figure 4) further confirms this trend quantitatively: mean variance increases monotonically from approximately 0.0013 (6 horizons) to 0.0073 (1 horizon), demonstrating that realization spread is directly controlled by the density of horizon constraints. Collectively, these results suggest that LFD generates a data-consistent distribution: when conditioning data are abundant, the model converges to consistent solutions; when data are sparse, the model appropriately broadens its output distribution to reflect the increased geological ambiguity, rather than collapsing to a single overconfident prediction.

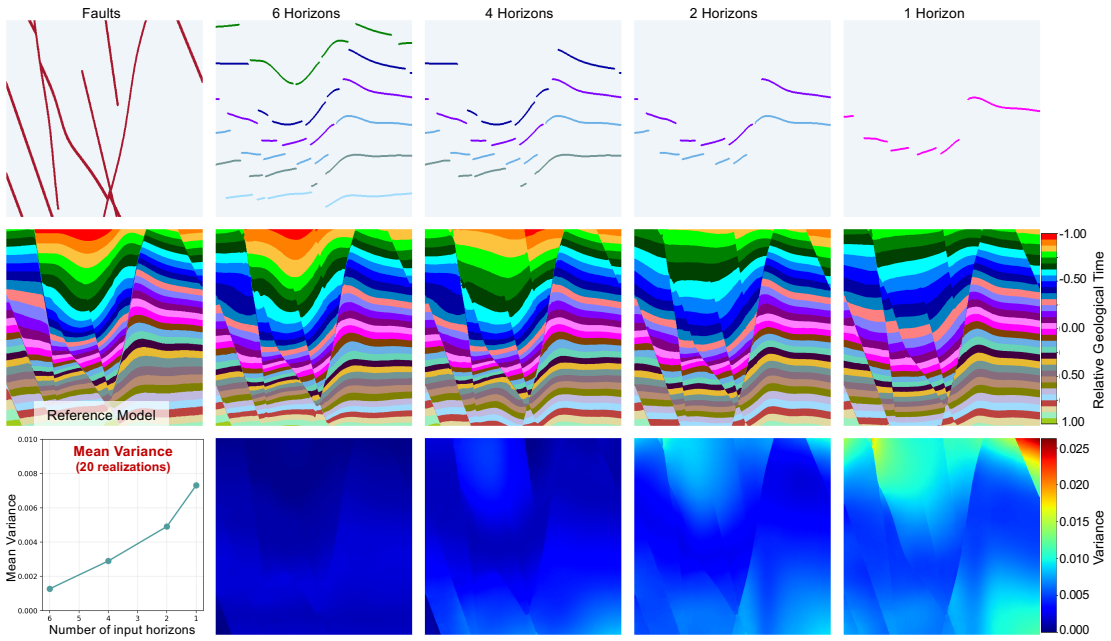

Figure 4: Ablation study on horizon sparsity and realization uncertainty. Top row: input conditioning, showing the shared fault interpretation (leftmost) and horizon inputs with 6, 4, 2, and 1 horizons. Middle row: mean implicit structural model over 20 realizations for each horizon configuration, with the reference model shown in the leftmost panel. Bottom row: pixel-wise variance across 20 realizations (leftmost panel: mean variance as a function of the number of input horizons), demonstrating that realization variance increases monotonically as horizon constraints become sparser.

Beyond conditioning sparsity, we further investigate how the strength of the proposed fault-aware bending-energy regularization affects the realization ensemble. For each bending-energy weight $\lambda_{Bend} \in \{0, 0.1, 1\}$ (λ_{Bend} =0.1 being the default used throughout this paper), we train a separate model under the same training settings and evaluate it on the full validation set (3200 cases). For

each case, we extract the central horizon of the implicit field together with the corresponding fault mask as the conditioning input, and generate 20 realizations from different noise seeds. We report the resulting pixel-wise Mean Squared Error (MSE), the Horizon Consistency Error (HCE-6, evaluated on 6 horizons uniformly extracted from the ground-truth implicit field), and the mean realization variance, averaged over the validation set, in Table 4.

Table 4. Ablation on fault-aware bending-energy weight λ_{Bend} under the sparsest conditioning setting (single input horizon), evaluated on the validation set (3200 cases, 20 realizations each). HCE-6 is computed on 6 horizons uniformly sampled from the ground-truth implicit field.

λ_{Bend}	MSE (1e-2) ↓	HCE-6 (1e-2) ↓	Mean Variance (1e-3)
0	3.73	3.65	7.63
0.1	2.93	2.84	7.27
1	3.50	3.36	8.61

As shown in Table 4, MSE and HCE-6 do not vary monotonically with λ_{Bend}, but both are consistently lower with the bending-energy prior enabled than without it, indicating that the regularization yields realizations that are more stable and closer to the reference model, rather than drifting into implausible solutions. In particular, λ_{Bend} =0.1 achieves the lowest pixel-wise error among the three settings. At the same time, the prior does not reduce realization variance; the strongest setting (λ_{Bend} =1) yields the highest variance among the three, showing that this improved stability does not come at the cost of the model's ability to represent structural uncertainty.

2. Reproducibility of the code. I attempted to run the code accompanying the manuscript and was unable to get it to execute. As the manuscript's central claims rely on specific implementation choices (e.g., the structure-enhanced Transformer conditioning, the fault-aware finite-difference stencils in Eq. 14, and the reported inference time of 1.56 s for a 512×512 model on an NVIDIA H20 GPU), I am not currently able to independently verify these claims. I would ask the authors to (a) verify that the uploaded code corresponds to the exact version used to produce the results in the manuscript, (b) provide a minimal working example with pinned dependency versions and, if possible, a small pretrained checkpoint or synthetic toy dataset that reproduces at least one figure/table end to end. Given GMD's emphasis on reproducibility of model development, this should be resolved before publication.

Response:
Thank you for raising this important issue, and for taking the time to attempt to run our code. We agree that the original repository lacked sufficient documentation to support reproduction of our results. In response, we have thoroughly revised the code repository as follows.

(a) We have verified and updated the repository to ensure that the uploaded code corresponds exactly to the version used to produce the results reported in the manuscript.

(b) We have substantially expanded the README documentation to include detailed environment setup instructions, covering pinned dependency versions, data preparation procedures, path

configuration, and inference guidelines. We have also uploaded a pretrained checkpoint together with the example real-survey data used in the manuscript to Zenodo, enabling the reviewer and other readers to reproduce the inference workflow and results without retraining.

The revised repository is now available on Zenodo (https://zenodo.org/records/20508635) and GitHub (https://github.com/ProgrammerZXG/LFD).

3. Uniform-grid assumption underlying Eq. 14. The fault-aware bending-energy term in Eq. 14 is written in terms of pixel-index derivatives ($\partial^2\hat{x}/\partial i^2$, $\partial^2\hat{x}/\partial i\partial j$, $\partial^2\hat{x}/\partial j^2$). This formulation implicitly assumes that the grid spacing is uniform and identical in both the i- and j-directions (i.e., an isotropic, regular sampling grid), since otherwise the index-space second differences do not correspond to a consistent physical curvature. This assumption should be stated explicitly, since implicit structural models are frequently defined on grids with unequal vertical/horizontal spacing, or on locally warped/flattened grids. I would ask the authors to: (a) explicitly state whether the ViT patch grid used by LFD is always uniform and isotropic in physical units; (b) if not, clarify whether the derivatives in Eq. 14 are normalized by the local grid spacing (e.g., divided by Δi^2, $\Delta i\Delta j$, Δj^2, respectively); and (c) provide at least one experiment or discussion addressing whether/how the method would behave on non-uniform grids, since this materially affects the physical meaning of the bending-energy penalty and, by extension, the geological plausibility of the generated structures.

Response:
Thank you for this important observation. We agree that Eq. 14, as originally written in terms of pixel-index derivatives, is not fully rigorous, since it does not explicitly account for the physical grid spacing in the i- and j-directions. We have revised Eq. 14 as follows:

$$\mathcal{L}_{\text{Bend}} = \mathrm{E}_{(i,j)\notin\Omega_{\text{Faults}}}\left[\left(\frac{1}{\Delta i^2}\frac{\partial^2\hat{\mathbf{x}}}{\partial i^2}\right)^2 + 2\left(\frac{1}{\Delta i\,\Delta j}\frac{\partial^2\hat{\mathbf{x}}}{\partial i\,\partial j}\right)^2 + \left(\frac{1}{\Delta j^2}\frac{\partial^2\hat{\mathbf{x}}}{\partial j^2}\right)^2\right],$$

where Δi and Δj denote the grid spacing in the i- and j-directions, respectively.

This revision makes the bending-energy term correspond to a consistent physical curvature. We clarify that the present study focuses on validating the effectiveness of the proposed x-prediction formulation, structure-enhanced conditioning, and prior-guided losses under a uniformly sampled grid, consistent with the fixed-size patch embedding used by the ViT backbone. Extending the framework to non-uniform or locally warped grids is an important direction, but would require a different patch embedding scheme to account for spatially varying grid spacing, which is beyond this scope. We have noted this as a limitation and a direction for future work in the Discussion section.

Limitation and 3D extension.
Similarly, the present study assumes a uniformly sampled grid, consistent with the fixed-size patch embedding used by the ViT backbone; extending the framework to non-uniform grids, where spacing varies spatially (e.g., locally warped or flattened grids), would require rethinking the patch embedding scheme, and represents another promising direction for future work.

4. Sensitivity to noise level, number of sampling steps, and initial state. The manuscript reports a single operating point for the flow-matching/diffusion sampler (e.g., a fixed number of integration steps and a fixed initialization) without a systematic sensitivity analysis. Since LFD is positioned as a generative model whose outputs should be both efficient and geologically reliable, I believe such an analysis is necessary. Specifically: (a) Noise level/integration step size: flow-matching and diffusion samplers are known to be sensitive to the discretization of the noise/time schedule, and coarser schedules can introduce integration error that is not obviously distinguishable from genuine multi-modal geological variability. The authors should report how the quality of the generated structural models (fidelity to horizons/faults, bending-energy statistics, and any other structural metrics used) varies as the step size/noise schedule is coarsened or refined. (b) Number of sampling steps: relatedly, the manuscript should report how output quality and inference time trade off as the number of steps is varied, since the headline efficiency result (1.56 s for a 512×512 model) is only meaningful in the context of a stated accuracy level, and readers need to know whether fewer steps could be used without materially degrading geological plausibility, or whether more steps are required to reach convergence. (c) Initial state: because LFD generates directly in data space rather than in a compact latent space, the choice of initial noise/state may have a more direct and visible effect on the final structural model than it would in latent diffusion. The authors should clarify whether the initial state is drawn from a fixed prior (e.g., standard Gaussian noise) for every sample, and, if so, present an ablation showing how the diversity and quality of generated structural models depend on the initial state. Without this sensitivity analysis, it is difficult to assess how robust the reported results are to these sampler-level choices, and whether the reported timing/quality figures represent a carefully chosen operating point or a somewhat arbitrary one.

Response:
Thank you for this insightful and important comment. We agree that reporting a single operating point without a systematic sensitivity analysis makes it difficult for readers to assess whether the reported efficiency and quality figures represent a carefully considered trade-off or a somewhat arbitrary choice, particularly given that LFD is positioned as a generative model expected to be both efficient and geologically reliable. We address each point in turn.

(a) Noise level/integration step size. We agree that flow-matching and diffusion samplers can indeed exhibit different sensitivities under different time-scheduling strategies. In our implementation, the ODE time interval is discretized uniformly ($\Delta t=1/K$, Eq. 6), so the integration step size and the number of sampling steps are jointly determined by the same parameter K: specifying one uniquely determines the other, and the two are not independent variables. We address this point through the systematic ablation on K presented in (b).

In addition, the original manuscript already includes an ablation on the noise scale (see "Impact of noise scale" in the Discussion section), in which we vary the noise scale from 2.0 to 0.1 while keeping all other settings fixed. The results show that decreasing the noise scale yields noticeably smoother and more geologically reasonable implicit fields, whereas larger noise scales commonly adopted in natural-image diffusion (e.g., 1.0 or 2.0) lead to visibly noisier generations. This finding also provides a useful reference for other generative tasks in geophysics: the noise scale should be

tuned to the characteristics of the target data rather than directly adopted from conventions established for natural images.

Impact of noise scale in Sec Discussion

We find that the noise scale is a critical factor for implicit structural modeling. In early experiments, using commonly adopted noise scales for natural-image diffusion (e.g., 1.0 or 2.0) often led to overly noisy generations even with prolonged training. This is likely because implicit structural models exhibit relatively smooth distributions, for which excessively strong corruption can hinder stable denoising. We perform an ablation study on the noise scale and observe that decreasing the noise scale yields noticeably smoother and more geologically reasonable implicit fields. This finding provides practical guidance for deploying diffusion models in geophysics: the noise scale should be tuned to the target task and data characteristics to achieve optimal generation quality.

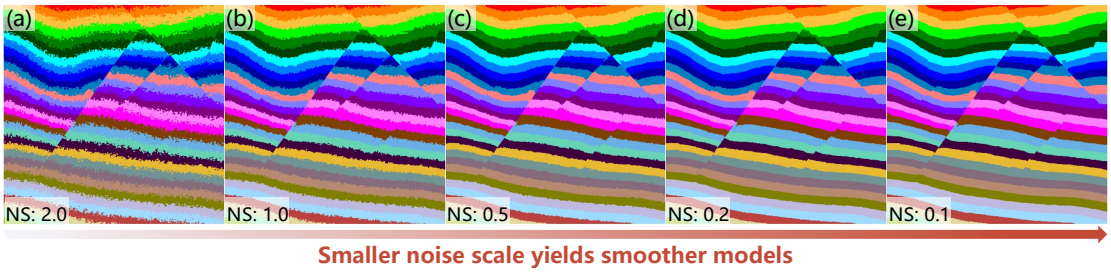

Figure 6. Effect of noise scale (NS) on generation. The model is trained with noise scale NS=2.0. During inference, we vary only the noise scale while keeping all other settings fixed. Panels (a--e) correspond to NS=2.0,1.0,0.5,0.2,0.1. As NS decreases, the generated implicit structural models become smoother and exhibit less residual noise.

(b) Number of sampling steps. We conduct an ablation on the number of ODE sampling steps $K \in \{5,10,20,30,50\}$, evaluated on the full validation set, using the fault mask together with the central horizon extracted from the implicit field as conditioning, and reporting the average inference time (see "Impact of sampling steps" in the Discussion section). The results show that inference time scales approximately linearly with K, while both MSE and HCE improve only marginally beyond K=20, with diminishing returns thereafter, indicating that geologically plausible, high-quality results can already be obtained with substantially fewer steps (e.g., 10 or 20). The default setting used throughout this paper (K=50) is intended to showcase the best achievable generation quality; in practice, the number of sampling steps can be reduced to achieve substantially faster inference while maintaining comparable quality. We thank the reviewer for raising this point, which has led us to characterize the applicable range of the reported efficiency figure more precisely and rigorously.

Impact of sampling steps in Sec Discussion

In addition to the noise scale, we examine the sensitivity of generation quality to the number of ODE sampling steps $K \in \{5,10,20,30,50\}$, evaluated on the full validation set, using the fault mask together with the central horizon extracted from the implicit field as conditioning, and reporting the average inference time. As shown in Table 5, inference time scales approximately linearly with K, while both MSE and HCE-6 improve only marginally beyond K=20, with diminishing returns thereafter. The default setting used throughout this paper (K=50) is chosen to showcase the best

achievable generation quality; in practice, the number of sampling steps can be reduced to obtain substantially faster inference while maintaining comparable quality.

Table 5. Ablation on the number of ODE sampling steps K.

K	MSE (1e-2) ↓	HCE-6 (1e-2) ↓	Time (s)
5	3.18	3.016	0.15
10	3.16	3.021	0.32
20	3.13	3.025	0.64
30	3.08	2.950	0.97
50	3.04	2.967	1.56

(c) Initial state. We confirm that the initial state z_0 is drawn from a fixed standard Gaussian prior, $z_0 \sim N(0, I)$ for every generated sample. The horizon-sparsity ablation presented in **Sec 3.5 Ablation Study on Uncertainty** already provides the relevant analysis: for each conditioning configuration, we generate 20 realizations from different initial noise samples and report the resulting pixel-wise variance. The results show that these realizations exhibit meaningful diversity that scales with the sparsity of the conditioning, rather than collapsing to a near-identical output regardless of the sampled initial state, indicating that the model's generation quality is not overly sensitive to the specific noise realization drawn at inference time.

The Second Paragraph of Sec 3.5

We note that, since these 20 realizations are generated from different initial noise samples, this also indicates that the sampled initial noise has a visible influence on the resulting structural details, with different noise samples leading to different specific realizations, while the overall outputs remain geologically plausible.

Specific Comment: The Second-Order Finite-Difference Bending-Energy Loss (Eq. 14)

1. Grid metric is not accounted for. As noted above, Eq. 14 is written entirely in index space. Even setting aside non-uniform grids, vertical and horizontal sampling intervals in seismic/structural volumes are typically different by construction, so the mixed term implicitly assumes an isotropic metric that generally does not hold. The authors should clarify whether/how the loss accounts for anisotropic or spatially varying grid spacing, and, if it does not, discuss the resulting bias.

Response:

Thank you for raising this point. We agree that Eq. 14, as originally written in terms of pixel-index derivatives ($\partial^2\hat{x}/\partial i^2$, $\partial^2\hat{x}/\partial i\partial j$, $\partial^2\hat{x}/\partial j^2$), is not fully rigorous, since it does not explicitly account for the physical grid spacing in the i- and j-directions and implicitly assumes an isotropic metric (i.e., $\Delta i = \Delta j$). As the reviewer points out, this assumption generally does not hold, particularly for structural volumes, where the vertical and horizontal sampling intervals typically differ by construction. We have revised Eq. 14 to incorporate the corresponding grid-spacing terms, normalizing the second-order derivatives by Δi^2, $\Delta i \cdot \Delta j$, and Δj^2, respectively, where Δi and Δj denote the grid spacing in the i- and j-directions. This revision makes the bending-energy term correspond to a consistent physical curvature, without implicitly assuming that the spacing in the two directions is equal.

We further clarify that the revised formulation assumes the grid is regular (i.e., Δi and Δj are constant throughout the domain) but does not require it to be isotropic (i.e., $\Delta i = \Delta j$), which is consistent with the synthetic and field datasets used in this study. Non-uniform grids, where spacing varies spatially (e.g., locally warped or flattened grids), remain outside the scope of the current formulation. We have explicitly noted this as a limitation and a direction for future work in the Discussion section ("Limitation and 3D extension").

2. The "clean" prediction is still an uncertain estimate at high noise levels, and this is where curvature penalties are most fragile. The manuscript states that "as the model directly predicts the clean data, geological observations and prior constraints can be imposed in a straightforward and differentiable manner", and applies this prediction. I would like to point out that even though is nominally the "clean" target, at early sampling steps / high noise levels the network's prediction of given a heavily corrupted input is fundamentally underdetermined: many plausible clean structural models are consistent with the same noisy observation, and the network's point estimate at this stage is typically a blurred, low-confidence average over these possibilities rather than a single coherent geological structure. Its discrete second-order curvature at this stage therefore reflects the shape of the network's predictive uncertainty, not a genuine geological feature (or lack thereof). Penalizing that curvature is a much less well-posed operation than penalizing the curvature of a final, high-confidence output, and could bias the network toward predicting overly smooth estimates at high noise levels purely to reduce this loss term, independent of whether the final sample should be smooth. I would ask the authors to clarify: (a) at which noise levels/timesteps it is applied during training (all, or only near the clean end of the schedule), and (b) whether the loss weighting is scheduled/annealed as a function of noise level to account for the fact that it is far less reliable at high noise levels. If the loss is applied uniformly across all noise levels, I would ask for evidence (e.g., a comparison of results with vs. without this weighting) that it is not simply driving the network toward generically smoother predictions rather than toward geologically faithful ones.

Response:
Thank you for raising this insightful point, which prompted us to take a closer look at how this loss term behaves throughout training.

(a) At which noise levels/timesteps this loss is applied. In the current implementation, the bending-energy loss is computed and applied at every training iteration, regardless of the sampled timestep t (i.e., the noise level) — that is, the loss is applied uniformly across all noise levels rather than only near the clean end of the schedule.

(b) Whether the loss weighting is scheduled/annealed as a function of noise level. In the current implementation, the weight λ_{Bend} is a fixed constant and is not scheduled or annealed with respect to noise level.

We agree with this mechanistic observation: at high noise levels, the network's prediction of the clean data $\hat{x}$ is essentially an averaged estimate over multiple plausible structures, which is mathematically expected behavior. To illustrate this phenomenon directly, we performed a qualitative

analysis in which the same ground-truth sample x and the same noise sample ε are fixed while only the timestep t is varied, examining the network's single-step prediction of $\hat{x}$ at different noise levels (see Figure 7). The results show that as t increases (i.e., as the noise level decreases), the prediction error decreases monotonically, and the clarity and structural detail of the prediction improve accordingly, confirming that predictions are indeed relatively more blurred at high noise levels.

Nevertheless, our results show that, even at high noise levels, the network is still able to recover a geologically plausible structure rather than a severely blurred or incoherent one, indicating that the reliability of the prediction at this stage, while lower than at low noise levels, remains sufficient to provide a meaningful geological signal. This suggests that applying the bending-energy regularization at high noise levels is still meaningful and does not severely compromise the final results. This is further supported quantitatively: as shown in Table 4, reconstruction metrics (MSE, HCE-6) improve rather than deteriorate as the loss weight increases from 0 to 1, while the variance across realizations does not decrease monotonically with increasing weight, with $\lambda_{\text{Bend}}=1$ in fact yielding the highest variance among the three settings. This indicates that the loss does not exhibit a systematic tendency to drive the network toward globally smoother outputs, and that its practical impact under the current weighting and training configuration does not compromise the geological plausibility of the final generated results.

We also acknowledge that a more refined weighting schedule, in which the strength of the bending-energy regularization decays with the timestep, could enable finer control over this trade-off. The corresponding discussion has been added to the Discussion section.

Reliability of clean-data prediction at high noise levels. Since LFD directly predicts the clean data $\hat{x}$ at every timestep, it is worth examining how the reliability of this prediction varies with the noise level, as this bears on the validity of applying the horizon and bending-energy losses uniformly across all timesteps during training. Using a single validation sample, we fix the same implicit field and noise realization while varying only the timestep t, and visualize the resulting single-step prediction $\hat{x}$ (Figure 7). As t increases from 0.1 to 0.9 (i.e., as the noise level decreases), the prediction error decreases by nearly two orders of magnitude (from 9.77×10^{-4} to 1.47×10^{-5}), and the predicted field becomes visibly sharper. This confirms that predictions are more blurred at high noise levels. In principle, geological and geophysical regularization terms such as the horizon and bending-energy losses would be more directly justified at low noise levels, where $\hat{x}$ more closely approximates the implicit model. However, our results show that, even at high noise levels, the network is still able to recover a geologically plausible structure rather than a severely blurred or incoherent one. This suggests that, for the implicit structural modeling task considered here, applying these prior losses at high noise levels remains reasonably well-justified. In this paper, the same regularization weight is applied uniformly across all noise levels t during training. Building on this observation, a promising direction for future work is to design a more refined weighting schedule in which the strength of the bending-energy regularization decays with the timestep, which may enable finer control over the trade-off between reconstruction fidelity and geological plausibility.

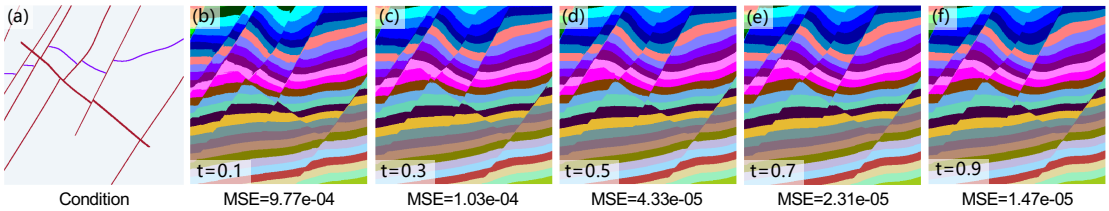

Figure 7 Single-step clean-data prediction at different noise levels. (a) Input conditioning. (b-f) Predicted x at t = 0.1, 0.3, 0.5, 0.7, 0.9, using the same ground-truth sample and the same noise realization across all panels. Prediction error (MSE, relative to the ground-truth field) decreases monotonically as t increases, indicating that predictions become progressively more reliable as the noise level decreases.

3. Inconsistent stencil order near fault-mask boundaries. "Fault-aware stencils that do not cross fault pixels" implies one-sided or truncated finite differences adjacent to. Standard central differences are second-order accurate, whereas one-sided or truncated stencils are typically only first-order accurate or have a different truncation-error constant. This means the effective smoothness penalty is not spatially homogeneous: pixels near a fault are regularized with a different bias/strength than interior pixels, purely as an artifact of discretization rather than of the underlying geology. Please specify the exact one-sided/modified stencils used near fault boundaries and confirm that they remain numerically consistent (i.e., that they converge to the correct derivative as), rather than being a central stencil with missing terms simply set to zero, which would not be consistent.

Response:

Thank you for raising this point. We clarify that the fault-aware treatment adopted in this work is not a one-sided or truncated finite-difference scheme, as the reviewer's comment surmised, but rather a stencil-validity masking strategy: for each second-order derivative term ($\partial^2\hat{x}/\partial i^2$, $\partial^2\hat{x}/\partial i\partial j$, $\partial^2\hat{x}/\partial j^2$), the curvature contribution at a given location is computed only if all sampling points involved in the corresponding stencil lie in the non-fault region; if any point of the stencil falls on a fault pixel, the curvature term at that location is set to zero and does not contribute to the loss.

Specifically, we do not replace the interior stencil with a lower-order one-sided or truncated approximation near fault boundaries. Consequently, the concern raised by the reviewer, namely that pixels near a fault would be regularized with a different bias or strength than interior pixels purely as an artifact of discretization, does not arise in our implementation. Pixels near a fault either use the exact same standard central-difference stencil as interior pixels, or, when the stencil crosses the fault, are entirely excluded from this loss term. In neither case does the accuracy of the stencil itself change near a fault.

We agree that this "exclusion" strategy, as opposed to substituting a lower-order one-sided scheme, means that a narrow band of pixels immediately adjacent to a fault (spanning the width of the finite-difference stencil) is not constrained by this regularization term. We note, however, that this effect is highly localized, spanning only a few pixels corresponding to the stencil width. Moreover, this region remains constrained by the other training objectives, including the flow matching loss and the horizon loss (where a horizon constraint passes through the region). We further note that a fault

itself represents a genuine discontinuity in the geological structure, so not enforcing a continuous-curvature constraint immediately at the fault also avoids imposing an incorrect continuity assumption at a location where the field is not expected to be continuous. Consistent with this, we do not observe any visible artifacts or geologically implausible structures near fault boundaries in our results on field data (Figure 3), suggesting that this localized effect does not compromise the geological plausibility of the generated structures in practice. We have clarified this implementation detail in the revised manuscript.

Specifically, for each second-order derivative term, we adopt a stencil-validity masking strategy: the curvature contribution at a given location is computed only if all sampling points involved in the corresponding standard central-difference stencil lie in the non-fault region; if any point of the stencil falls on a fault pixel, the curvature term at that location is set to zero rather than being approximated with a lower-order one-sided or truncated scheme. As a result, the accuracy of the stencil itself does not vary near fault boundaries.

4. Risk of over-smoothing genuine, non-fault geological curvature. A quadratic penalty on bending energy favors globally low curvature (thin-plate-like behavior). However, fold hinges, drag folds, and tight synclines/anticlines away from faults are geologically real, legitimate high-curvature features that are not associated with a discontinuity. A blanket quadratic penalty risks over-smoothing these features unless the loss weight is very carefully tuned. I would ask for an ablation showing the sensitivity of the results to the weight of , and specifically its effect on fold-heavy synthetic or real test cases, together with a discussion of whether a more robust penalty (e.g., an L1 curvature term or edge-preserving weighting) was considered and, if so, why it was not adopted.

Response:
Thank you for raising this important point. We agree that a quadratic bending-energy penalty inherently favors globally low curvature (thin-plate-like behavior), and that, if the loss weight is not carefully chosen, there is in principle a risk of over-smoothing genuine, non-fault-related high-curvature geological features, such as fold hinges, drag folds, and tight synclines/anticlines.

Sensitivity ablation on the loss weight and its effect on high-curvature scenarios. We conducted a systematic ablation on the bending-energy weight $\lambda_{Bend} \in \{0, 0.1, 1\}$ under the sparsest conditioning setting (a single input horizon) (Table 4). The results show that reconstruction accuracy metrics (MSE, HCE-6) improve as the weight increases, while the realization variance does not decrease monotonically with increasing weight. In addition, we constructed several high-curvature test cases in which the fold hinges are located away from any fault (including tight anticlines and asymmetric folds), and generated results under the same three weights for comparison (Figure 8). The results show that, even under the strongest setting ($\lambda_{Bend}=1$), the model is still able to accurately preserve the high-curvature geometry at these fold hinges, with no noticeable over-smoothing or structural collapse observed. This indicates that, within the range of weights tested in this study, the regularizer does not come at the cost of genuine, non-fault-related high-curvature geological features.

Whether a more robust penalty (e.g., an L1 curvature term or edge-preserving weighting) was considered. We acknowledge that this work does not adopt an L1 curvature penalty or edge-

preserving weighting, nor have we systematically compared such alternatives against the current quadratic penalty. The quadratic (thin-plate-like) bending-energy penalty adopted here follows the conventional form of this type of regularizer commonly used in implicit structural modeling (e.g., Renaudeau et al., 2019; Belhachmi et al., 2025). We agree that L1 curvature terms or edge-preserving weighting schemes, which are more tolerant of localized high-curvature features, could further reduce the risk of over-smoothing genuine fold structures, and represent a promising direction for systematic investigation. We have identified this comparison as a direction for future work and noted it accordingly in the Discussion section.

Preservation of high-curvature features. While the bending-energy loss suppresses geologically implausible artifacts, a quadratic penalty on curvature could, in principle, also over-smooth genuine, non-fault high-curvature features such as fold hinges. To examine this risk, we constructed several high-curvature test cases in which the fold hinges are located away from any fault (including a tight anticline, an asymmetric fold, and a double-hinge fold without any fault present), and generated results under three bending-energy weights ($\lambda_{Bend} \in \{0, 0.1, 1\}$) for comparison (Figure 8). Even under the strongest setting (λ_{Bend} =1), the model accurately preserves the high-curvature geometry at these fold hinges, with no noticeable over-smoothing or structural collapse observed. Together with the quantitative results in Table 4, where reconstruction accuracy improves and realization variance does not decrease monotonically with λ_{Bend}, this indicates that, within the tested weight range, the regularizer does not compromise high-curvature geological features while still preserving meaningful realization diversity.

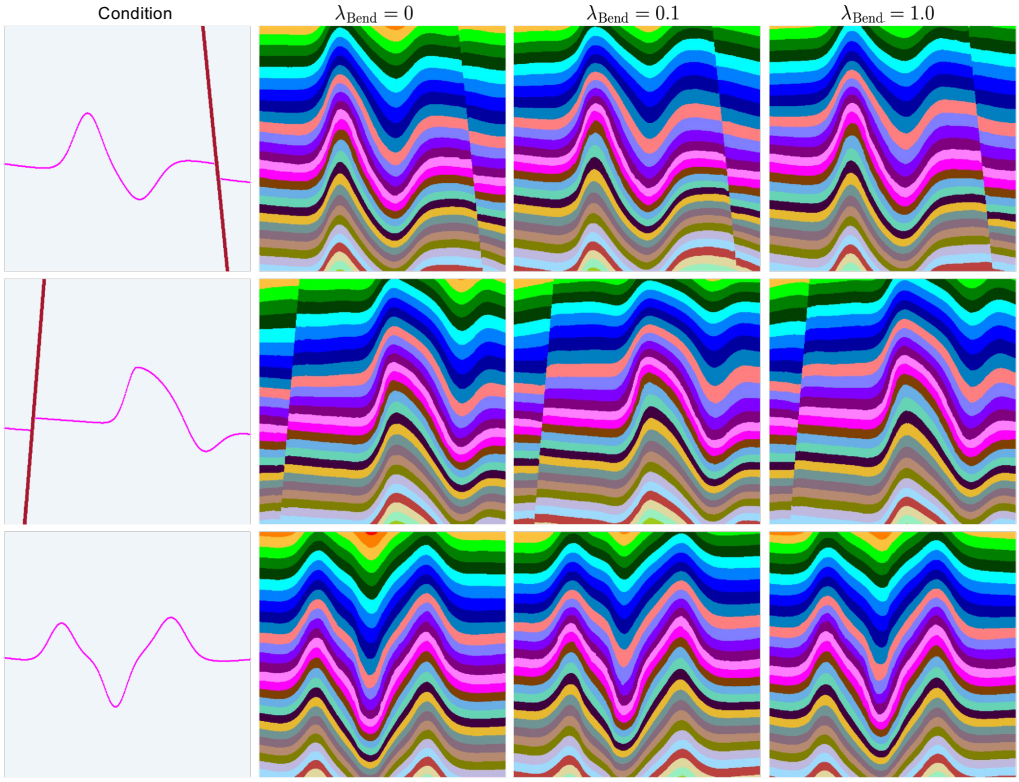

Figure 8. Sensitivity of high-curvature fold hinges to the bending-energy weight. Left column: input conditioning for three test cases, each featuring a fold hinge positioned sufficiently far from any fault: a tight anticline (top), an asymmetric fold (middle), and a double-hinge fold without any fault present (bottom). Right three columns:

ensemble-mean structural models, averaged over 20 realizations from different noise seeds, generated under λ_{Bend} = 0, 0.1, 1.0, respectively.

We would like to express our sincere appreciation for your thorough assessment of our manuscript and for the valuable comments and suggestions provided. We have carefully revised the manuscript accordingly and believe that these changes have substantially enhanced its clarity and overall quality. We hope that the revised manuscript has satisfactorily addressed all the concerns raised.

Kind regards,

Zhixiang Guo