MULTIMODAL PHYSICAL POSTERIOR INFERENCE FOR EARLY PREDICTION OF SURFACE DEFORMATION FROM INSAR TIME SERIES

ABSTRACT

Early observations of mining-induced subsidence reveal surface motion but leave its eventual magnitude and growth rate uncertain. This ambiguity makes forecasting from short interferometric synthetic aperture radar (InSAR) sequences difficult. We introduce Multimodal Physical Posterior Inference (MPPI), which uses geological and mining context to constrain physical parameters that early deformation maps cannot fully identify. MPPI combines visual patterns with textual descriptions of observation statistics and geological conditions to infer a posterior over maximum subsidence and growth rate. A physics-based decoder converts these parameters into future deformation fields and propagates posterior samples into prediction intervals. We pre-train the inference network on simulated sequences and fine-tune it on real InSAR observations. Across six coal-mining regions in Shanxi Province, China, MPPI achieves 4 cm RMSE for both maximum-subsidence estimation and deformation forecasting. Ablations identify geological information as the strongest complementary source, with textual representation providing further gains. Probabilistic parameter learning halves forecast RMSE relative to deterministic regression, while posterior sampling improves the continuous ranked probability score. These results support combining multimodal evidence with physical-parameter inference to forecast deformation from incomplete observations.

1 INTRODUCTION

Accurate prediction of surface deformation supports geological-hazard assessment and risk mitigation. Interferometric synthetic aperture radar (InSAR) provides wide-area measurements of historical deformation at high spatial resolution and relatively low cost, making it a useful basis for forecasting (Osmanoğlu et al., 2016). We study the setting in which only an early prefix of an InSAR deformation sequence is available, often before the deformation reaches its inflection point. At this stage, the observations reveal the current spatial pattern but provide incomplete evidence about the eventual magnitude and rate of deformation.

Existing approaches expose complementary limitations in this setting. Model-driven methods specify a temporal or physical law, estimate its parameters from InSAR observations, and use the calibrated model for forecasting. Their interpretability is valuable, but the estimates depend on source parameters and on whether the chosen law matches the observed process (Hu et al., 2022; Dong et al., 2026; Weston et al., 2012; Masterlark, 2007). Data-driven methods learn a direct mapping from historical observations to future states (Han et al., 2019; Lim & Zohren, 2021; Benidis et al., 2022). They can represent complex spatiotemporal patterns, yet the limited number of InSAR sequences makes the mapping difficult to learn without overfitting. Knowledge-guided models combine these paradigms by adding physical relationships to network training (Hettige et al., 2024; Li et al., 2025b). For mining-induced deformation, the S-shaped temporal evolution motivates Verhulst-based prediction with prior-parameter inversion (Xiang et al., 2026), but an early prefix may still be insufficient to identify the carrying capacity and growth rate from the deformation sequence alone.

The central difficulty is parameter identifiability rather than only feature extraction. The future field is governed by latent physical quantities, while the early prefix constrains them only partially. Visual deformation patterns provide evidence about basin morphology and its evolution; statistical descriptions summarize the observed numerical state; geological and mining information provides

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.

context that is not visible in the deformation maps. Multimodal forecasting methods have begun to combine visual and textual evidence (Wang et al., 2024; Wu et al., 2026), but they generally do not specify how these heterogeneous signals should constrain physically meaningful deformation parameters. This motivates a formulation in which multimodal evidence conditions a posterior over temporal parameters and a physical decoder converts the inferred parameters into future maps.

We propose Multimodal Physical Posterior Inference (MPPI) for early-stage deformation forecasting. MPPI conditions a posterior over the temporal parameters $\boldsymbol{\theta} = (W_{\max}, r)$ on the observation prefix, visual deformation patterns, statistical text, and geological text. PVL and PTL project these sources into a common parameter-prior space. MPIN fuses the resulting local and global representations through cross-attention and a learned gate, while PPM decodes the posterior estimate with Verhulst temporal dynamics and a probability-integral spatial kernel. The spatial parameters $\boldsymbol{\eta} = (q, \rho, \phi)$ are optimized together with the physical fitting objective and used by the decoder for spatial reconstruction. The main contributions are as follows:

- We formulate early-stage deformation forecasting as posterior inference over the temporal parameters $\boldsymbol{\theta} = (W_{\max}, r)$, making the parameter ambiguity explicit when the observation prefix precedes the inflection point.
- We develop MPPI to implement this formulation: PVL and PTL provide visual and textual priors, MPIN fuses them with prefix evidence, and PPM combines the inferred temporal parameters with the fitted spatial parameters $\boldsymbol{\eta}$ to generate physically structured forecasts.
- We evaluate parameter inversion, map-level forecasting, and the contribution of statistical and geological text on six real coal-mining regions in Shanxi Province. Under the reported protocol, MPPI obtains 4 cm RMSE for $W_{\max}$ estimation and 4 cm RMSE for deformation forecasting.

2 Related Work

2.1 Time-Series Prediction Based on Model-Driven Methods

Model-driven approaches forecast the deformation of the surface by combining a temporal function, which describes dynamic subsidence at individual locations, with a static spatial prediction model. Common temporal functions include the Knothe, Weibull, and Logistic models (Berardino et al., 2003; Tsoularis & Wallace, 2002). For example, a Weibull temporal function can be integrated with the probability-integral model, where time series of LOS deformation obtained from InSAR are utilized to invert the parameters of the prediction model and to forecast mining-induced surface deformation. An alternative strategy renders the parameters of a static model time-dependent. For instance, a simplified Boltzmann function can dynamically model the subsidence coefficient within the probability-integral model and support the prediction of future deformation (Cercignani, 1988). Although these methods provide physically interpretable forecasts, they rely heavily on assumptions designed by experts and accurate source parameters. When the assumed temporal law or parameterization fails to capture the actual geological process, the predictions can become unreliable.

2.2 Time-Series Prediction Based on Data-Driven Methods

Data-driven methods learn a mapping from historical observations to future states without explicitly modeling the underlying physical processes. Early studies primarily utilize recurrent architectures, such as RNNs and LSTMs, to extract temporal dependencies, while incorporating spatial heterogeneity or regional clustering into the forecasting process (Sherstinsky, 2020; Shiri et al., 2023). More recently, Transformer architectures have become widely adopted, as the mechanism of self-attention can capture long-range dependencies. The Informer model reduces the computational cost of modeling long sequences through sparse attention (Zhou et al., 2021); TimeSformer applies spatiotemporal attention to videos and spatiotemporal sequences (Bertasius et al., 2021); PatchTST divides a time series into consecutive patches to enhance the extraction of local patterns (Nie et al., 2022); and ESTformer further models temporal changes at multiple scales (Li et al., 2025a). These approaches can learn complex spatiotemporal features from historical sequences of deformation. However, they usually require relatively large datasets for training and may exhibit poor generalization when only a limited number of InSAR observations are available.

2.3 Time-Series Prediction with Knowledge-Guided Models

Knowledge-guided models incorporate physical laws, differential equations, or conservation relationships into the architectures and objective functions of neural networks, thereby reducing the dependence of data-driven models on large sets of training data. Representative PINN approaches have been applied to the prediction of tunnel settlement, landslide deformation, and mining-induced deformation (Zhang et al., 2026; 2023). The PhyDNet model combines physical dynamics with convolutional networks, whereas PeRCNN integrates the structures of partial differential equations with data-driven features to improve spatiotemporal prediction (Le Guen & Thome, 2020; Rao et al., 2023). These methods improve physical interpretability, but the performance of such approaches generally depends on accurate prior equations, boundary conditions, and physical parameters. When the physical assumptions fail to align with the actual geological environment, or when key parameters are difficult to obtain, fixed physical constraints can restrict the ability of the model to adapt to complex processes of subsidence.

2.4 Time-Series Prediction with Multimodal Models

Driven by the recent development of vision-language models, multimodal methods have been increasingly explored for the forecasting of time series. These methods commonly convert time series into line plots, grayscale images, or other two-dimensional representations to extract structural characteristics, such as trends, periodicity, and local changes. Statistical descriptions, task information, or domain knowledge are subsequently utilized to provide complementary temporal semantics. Representative examples include Time-VLM, which combines visual and textual information for forecasting (Zhong et al., 2025), as well as GPT4TS and Time-LLM, which convert time series into textual representations and utilize pretrained language models for prediction (Zhou et al., 2023; Jin et al., 2024). Although these approaches demonstrate potential in few-shot scenarios, they still face limitations in continuous numerical modeling, the representation of spatial structures, and the effective utilization of domain knowledge. These limitations motivate the development of a multimodal framework that directly links visual and textual evidence to physically meaningful parameters of deformation.

3 Method

To address limited parameter identifiability before the deformation inflection point, MPPI formulates forecasting as posterior inference over the temporal parameters $\boldsymbol{\theta} = (W_{\max}, r)$. The observation prefix is interpreted jointly with visual and textual priors, and the posterior estimate is decoded by a physics-based predictor. The spatial parameters $\boldsymbol{\eta} = (q, \rho, \phi)$ are optimized by the physical fitting objective and used by the decoder for spatial reconstruction. The framework follows one inference path: PVL and PTL construct priors, MPIN fuses them with prefix evidence, and PPM generates future deformation maps.

PVL divides early InSAR deformation maps into patches and uses a frozen visual encoder to represent spatial evolution. Its output captures basin expansion, local gradients, and deformation intensity, providing visual evidence for identifying $\boldsymbol{\theta}$ before the inflection point.

PTL organizes deformation statistics, geological and mining conditions, and the forecasting task into structured prompts. A frozen text encoder converts each prompt into a semantic embedding, providing complementary evidence about feasible temporal parameters.

MPIN fuses the visual and textual priors with prefix evidence and predicts a posterior distribution over $\boldsymbol{\theta} = (W_{\max}, r)$. It addresses the ambiguity that remains when the prefix does not contain the inflection point.

PPM combines a posterior estimate of $\boldsymbol{\theta}$ with the spatial parameters $\boldsymbol{\eta}$ and generates future maps through temporal growth and spatial diffusion.

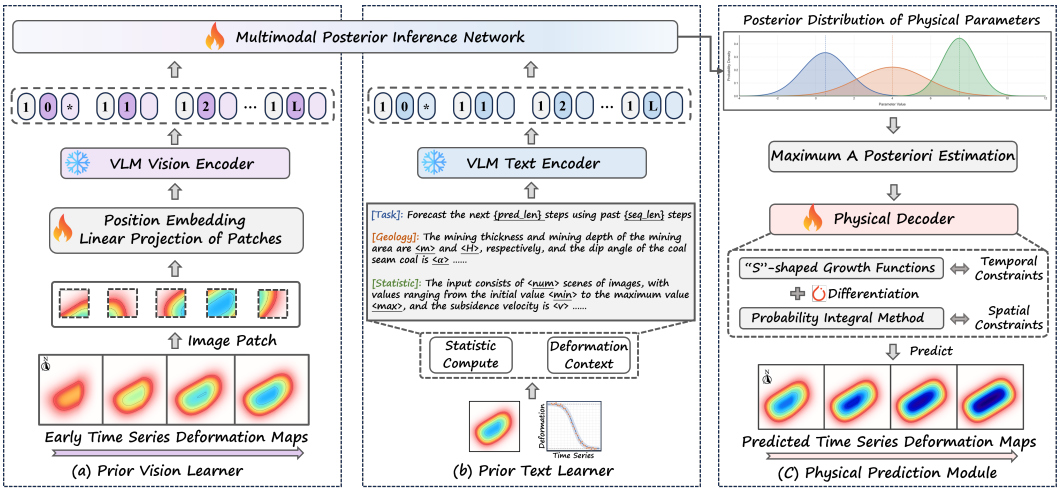

Figure 1: The framework of Multimodal Physical Posterior Inference framework (MPPI)

3.1 Prior Vision Learner (PVL)

Early spatial deformation patterns contain evidence about subsequent evolution. Let the observed deformation prefix be

$$\mathcal{D} = \{D_t\}_{t=1}^{L_{\text{obs}}}, \qquad D_t \in \mathbb{R}^{H \times W}, \tag{1}$$

where L_{obs} is the number of observed times and $H \times W$ is the spatial resolution. Each deformation map is robustly normalized using statistics computed from the observation prefix:

$$\widetilde{D}_t = \frac{D_t - \mu_{\text{obs}}}{\sigma_{\text{obs}} + \epsilon}, \tag{2}$$

where μ_{obs} and σ_{obs} denote the prefix mean and standard deviation, and ϵ is a numerical-stability constant. We partition each map into N non-overlapping patches. After flattening, each patch is linearly projected into a d-dimensional latent space:

$$z_{t,i} = \text{vec}(\widetilde{D}_{t,i})\mathbf{W}_p + \mathbf{e}_i^{\text{pos}} + \mathbf{e}_t^{\text{time}}, \tag{3}$$

where $\mathbf{W}_p$ is the patch-projection matrix, $\mathbf{e}_i^{\text{pos}}$ is the spatial-position embedding, and $\mathbf{e}_t^{\text{time}}$ encodes the temporal index of the patch. The temporal embedding is essential because concatenating patches without it preserves geometric adjacency but discards temporal order.

A frozen RemoteCLIP visual encoder f_{v} extracts visual features (Liu et al., 2024). Let Z denote the sequence of patch tokens. The encoder output is

$$T_{\text{v}} = f_{\text{v}}(Z), \tag{4}$$

Lightweight projection heads map T_{v} into a common parameter-prior space, yielding local visual priors $P_{\text{v}}^{\text{local}}$ and a global visual prior $p_{\text{v}}^{\text{global}}$. The local representation supports cross-modal interaction, while the global representation summarizes basin morphology, influence range, and evolution stage.

3.2 Prior Text Learner (PTL)

PTL converts early deformation observations and geological information into semantic priors for temporal-parameter inference. It extracts numerical statistics from deformation patches and organizes them with geological conditions and task information into structured prompts. The visual and textual embeddings are projected into a common parameter-prior space, but remain separate local and global representations until MPIN fuses them through cross-attention and a learned gate.

For each patch i, we extract statistics that describe subsidence evolution, including its mean, extrema, expansion scale, and temporal change rate:

$$\mathbf{s}_{t,i} = \left[\mu_{t,i},\ d_{t,i}^{\min},\ d_{t,i}^{\max},\ a_{t,i},\ v_{t,i}\right], \tag{5}$$

where $\mu_{t,i}$ is the patch mean, $d_{t,i}^{\min}$ and $d_{t,i}^{\max}$ are the minimum and maximum deformation values, $a_{t,i}$ denotes the spatial expansion scale, and $v_{t,i}$ denotes the temporal change rate. We further construct a global statistical summary by aggregating these patch-level descriptors:

$$\mathbf{s}^{\text{global}} = \text{Agg}_{t,i}(\mathbf{s}_{t,i}). \tag{6}$$

In addition to explicit statistics, we introduce geological and mining conditions $\mathcal{G}_i$ for each patch, including seam dip, mining thickness m_i, mining depth U_i, strike s_i, dip direction d_i, and mining speed v_i. The complete semantic-statistical representation is therefore

$$\mathbf{q}_i = [\mathbf{s}_i, \mathcal{G}_i, \Delta t], \tag{7}$$

where Δt is the observation interval. A structured textual prompt is generated by a template function $\mathcal{T}(\cdot)$:

$$\mathbf{x}_i = \mathcal{T}(\mathbf{q}_i). \tag{8}$$

We use a frozen Qwen3-4B text encoder f_{t} to obtain language-side prior representations (Bai et al., 2025; Yang et al., 2025):

$$T_{\text{t},i} = f_{\text{t}}(\mathbf{x}_i). \tag{9}$$

Lightweight projection heads map these embeddings to the parameter-prior space, yielding local textual priors $P_{\text{t}}^{\text{local}}$ and a global textual prior $p_{\text{t}}^{\text{global}}$. Compared with directly processing raw numerical sequences, PTL provides a more stable semantic representation in which early-observation statistics can participate in posterior inference in a compositional and cross-modal manner.

3.3 Multimodal Posterior Inference Network (MPIN)

Given visual and textual priors, MPIN combines prior semantics with observation evidence to infer the posterior over $\boldsymbol{\theta} = (W_{\max}, r)$. It predicts the temporal parameters that control subsidence evolution instead of directly regressing a future deformation field. The spatial parameters $\boldsymbol{\eta}$ are not outputs of MPIN. Let $\mathcal{D}_{1:L_{\text{obs}}}$ denote the observed prefix. MPIN learns the conditional posterior

$$q_\phi(\boldsymbol{\theta} \mid \mathcal{D}_{1:L_{\text{obs}}}, P_{\text{v}}^{\text{local}}, P_{\text{t}}^{\text{local}}, p_{\text{v}}^{\text{global}}, p_{\text{t}}^{\text{global}}). \tag{10}$$

In addition to local visual and textual priors, MPIN extracts evidence from the raw observation prefix. Each deformation map is spatially compressed and then encoded temporally:

$$\mathbf{o}_t = \text{Pool}(D_t), \qquad \mathbf{h}_{1:L_{\text{obs}}} = f_{\text{seq}}(\mathbf{o}_{1:L_{\text{obs}}}), \tag{11}$$

where $\mathbf{h}_t$ preserves prefix dynamics and the aggregated vector $\mathbf{h}^{\text{global}} = \text{Agg}_t(\mathbf{h}_t)$ provides global observation evidence.

Conditioned on the observation representation, MPIN performs cross-modal interaction between the visual and textual priors using multi-head cross-attention. Let $\mathbf{Q} = \mathbf{h}_{1:L_{\text{obs}}}\mathbf{W}_Q$, while $\mathbf{K}_{\text{v}} = P_{\text{v}}^{\text{local}}\mathbf{W}_{K,\text{v}}$ and $\mathbf{V}_{\text{v}} = P_{\text{v}}^{\text{local}}\mathbf{W}_{V,\text{v}}$ (and analogously for the textual prior). The attended features are

$$\mathbf{c}_{\text{v}} = \text{MHA}(\mathbf{Q}, \mathbf{K}_{\text{v}}, \mathbf{V}_{\text{v}}), \qquad \mathbf{c}_{\text{t}} = \text{MHA}(\mathbf{Q}, \mathbf{K}_{\text{t}}, \mathbf{V}_{\text{t}}). \tag{12}$$

A learned gate adaptively fuses the two priors:

$$\mathbf{g} = \sigma\Big(\mathbf{W}_g[\mathbf{h}^{\text{global}}; p_{\text{v}}^{\text{global}}; p_{\text{t}}^{\text{global}}] + \mathbf{b}_g\Big), \qquad \mathbf{c} = \mathbf{g} \odot \mathbf{c}_{\text{v}} + (1 - \mathbf{g}) \odot \mathbf{c}_{\text{t}}. \tag{13}$$

The fused feature is combined with the observation evidence through a normalized residual connection, producing $\widetilde{\mathbf{h}} = \text{LayerNorm}(\mathbf{c} + \mathbf{h}_{1:L_{\text{obs}}})$.

We globally aggregate $\widetilde{\mathbf{h}}$ and concatenate it with global observation, visual, and textual priors:

$$\mathbf{u} = \Big[\text{Agg}_t(\widetilde{\mathbf{h}}); \mathbf{h}^{\text{global}}; p_{\text{v}}^{\text{global}}; p_{\text{t}}^{\text{global}}\Big]. \tag{14}$$

Two parallel multilayer perceptrons predict the posterior mean and log variance,

$$\boldsymbol{\mu}_\theta = \text{MLP}_\mu(\mathbf{u}), \qquad \boldsymbol{\ell}_\theta = \text{MLP}_\ell(\mathbf{u}), \tag{15}$$

with positive variance enforced by

$$\boldsymbol{\sigma}_\theta = \mathrm{softplus}\big(\tfrac{1}{2}\boldsymbol{\ell}_\theta\big) + \epsilon, \qquad \boldsymbol{\sigma}_\theta^2 = \boldsymbol{\sigma}_\theta \odot \boldsymbol{\sigma}_\theta. \tag{16}$$

The resulting diagonal Gaussian posterior is

$$q_\phi(\boldsymbol{\theta} \mid \cdot) = \mathcal{N}\big(\boldsymbol{\mu}_\theta, \mathrm{diag}(\boldsymbol{\sigma}_\theta^2)\big)\,. \tag{17}$$

A reparameterized sample $\boldsymbol{\theta} = \boldsymbol{\mu}_\theta + \boldsymbol{\sigma}_\theta \odot \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, is passed to PPM for stochastic forecasting. For posterior-mean forecasts, PPM uses $\boldsymbol{\mu}_\theta$, which equals the MAP estimate for this diagonal Gaussian. Table 3 compares this decoding rule with posterior sampling and direct deterministic parameter regression.

3.4 Physical Prediction Module (PPM)

After MPIN infers the posterior distribution over $\boldsymbol{\theta}$, PPM performs physics-based decoding rather than directly regressing future deformation values. The spatial parameters $\boldsymbol{\eta}$ are optimized by the physical fitting objective and used by the decoder for spatial reconstruction. Here $W_{\max}$ denotes maximum subsidence and r the temporal growth rate. For posterior-mean forecasts, we use the maximum a posteriori estimate

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} q_\phi(\boldsymbol{\theta} \mid \cdot). \tag{18}$$

For the diagonal Gaussian posterior, this estimate equals the posterior mean, $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \boldsymbol{\mu}_\theta$.

For temporal evolution, we use the Verhulst differential equation to describe the S-shaped growth of subsidence (Ausloos, 2011). Let $S(t)$ denote cumulative subsidence. Its temporal response satisfies

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = rS(t)\left(1 - \frac{S(t)}{W_{\max}}\right). \tag{19}$$

To avoid introducing an additional initial parameter, PPM anchors the prediction at the last observed time t_{obs} and cumulative subsidence $S(L_{\mathrm{obs}})$. For a future time $t > L_{\mathrm{obs}}$, the response is written as

$$S(t) = \frac{W_{\max}}{1 + \left(\frac{W_{\max}}{S(L_{\mathrm{obs}})} - 1\right)\exp[-r(t - L_{\mathrm{obs}})]}. \tag{20}$$

The differential equation and its anchored solution provide the temporal consistency used by the decoder; an additional velocity output is not required for forecasting.

For spatial evolution, we adopt the probability-integral formulation to describe the diffusion of the subsidence basin (Zhou et al., 2026). Let $\mathbf{x}$ denote a surface location. The spatial parameters are $\boldsymbol{\eta} = (q, \rho, \phi)$, where q is the subsidence coefficient, ρ is the influence radius, and ϕ is the principal-direction angle. Let Ω denote the area of the mining-influence domain. We use the probability-integral formulation

$$K(\mathbf{x}; \boldsymbol{\eta}) = \frac{q\rho^2}{\Omega} \int_\Omega \exp\left[-\frac{\pi \|\mathbf{R}_\phi^\top(\mathbf{x} - \mathbf{u})\|^2}{\rho^2}\right] \mathrm{d}\mathbf{u}, \tag{21}$$

where $\mathbf{R}_\phi$ is the rotation matrix determined by ϕ and $\mathbf{u}$ is an infinitesimal position in the influence domain. This normalized kernel describes spatial attenuation and anisotropic diffusion from the mining area to the surface.

PPM combines the temporal response parameterized by $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ with the spatial response parameterized by $\boldsymbol{\eta}$. The predicted deformation at any future time is

$$\widehat{D}(\mathbf{x}, t) = S(t)K(\mathbf{x}; \boldsymbol{\eta}), \tag{22}$$

and the sequence of L_{pred} future maps is

$$\widehat{\mathcal{D}}_{\mathrm{future}} = \{\widehat{D}(\mathbf{x}, t)\}_{t=L_{\mathrm{obs}}+1}^{L_{\mathrm{obs}}+L_{\mathrm{pred}}}. \tag{23}$$

This construction bounds the temporal amplitude by $W_{\max}$ and uses K to constrain the spatial extent and morphology of the subsidence basin.

To jointly enforce observation fitting and physical consistency, PPM is trained with

$$\mathcal{L} = \mathcal{L}_{\text{obs}} + \lambda_t \mathcal{L}_t + \lambda_s \mathcal{L}_s, \tag{24}$$

where

$$\mathcal{L}_{\text{obs}} = \frac{1}{L_{\text{obs}}} \sum_{t=1}^{L_{\text{obs}}} \left\| \widehat{D}_t - D_t \right\|_2 \tag{25}$$

is the observation loss, where $\widehat{D}_t = S(t)K(\mathbf{x}; \boldsymbol{\eta})$ reconstructs the observed map at time t. The temporal and spatial regularizers are

$$\mathcal{L}_t = \frac{1}{L_{\text{obs}}} \sum_{t=1}^{L_{\text{obs}}} \left| \frac{\partial S(t)}{\partial t} - rS(t) \left(1 - \frac{S(t)}{W_{\max}} \right) \right|, \tag{26}$$

$$\mathcal{L}_s = \frac{1}{L_{\text{obs}}} \sum_{t=1}^{L_{\text{obs}}} \left\| D_t - S(t)K(\mathbf{x}; \boldsymbol{\eta}) \right\|_2 . \tag{27}$$

The coefficients λ_t and λ_s control the strengths of the temporal and spatial physical regularization. During training, $\boldsymbol{\theta}$ is supervised through the posterior head, while $\boldsymbol{\eta}$ is optimized through the observation and spatial-reconstruction terms. At inference time, PPM uses either $\widehat{\boldsymbol{\theta}}_{\text{MAP}}$ or samples from $q_\phi(\boldsymbol{\theta} \mid \cdot)$, with the fitted $\boldsymbol{\eta}$ held fixed for the future rollout. For posterior sampling, each parameter sample generates a future deformation trajectory. At each location and forecast time, the 5th and 95th percentiles of these trajectories define a pointwise 90% prediction interval. These intervals propagate uncertainty in the temporal parameters through PPM and form the shaded bands in Figure 6.

4 Experiments

4.1 Pre-training Based on Simulated Deformation

To stabilize parameter learning given the limited availability of real-world mining data and a scarcity of observations prior to the inflection point, we construct a simulated pre-training dataset based on physical models. During pre-training, the model learns a mapping from multimodal observations to temporal parameters $\boldsymbol{\theta} = (W_{\max}, r)$, establishing a foundation for stable parameter identification prior to fine-tuning on real data. The procedure adheres to the proposed framework: PVL receives sequences of simulated subsidence maps, whereas PTL receives the corresponding textual features. The VLM encoders within both modules remain frozen, and only MPIN along with the associated parameter posterior head are optimized. The simulated data are jointly generated via the probability integral method and an S-shaped temporal function. We randomly sample geological and mining conditions to construct a large pre-training dataset that covers diverse subsidence magnitudes, basin morphologies, and temporal profiles. The training objective combines a parameter regression loss with a distribution learning loss. The former encourages agreement between the inferred parameters and the simulated ground truth, whereas the latter aligns the posterior distribution with the ground truth parameter labels. This combined approach supports both accurate parameter estimation and uncertainty modeling. This procedure enables the model to learn the relationships among early subsidence patterns, textual priors, and physical parameters within a controlled simulation domain. Consequently, this provides a more stable initialization for fine-tuning on real InSAR data and improves the robustness of parameter posterior inference in small-sample settings.

4.2 Datasets and Baselines

Datasets: We evaluate MPPI across six regions affected by underground coal mining in Shanxi Province, China, including areas within Datong, Luliang, and Jincheng. As illustrated in Figure 2, these six mining regions exhibit an uneven spatial distribution across the province and have reached the late stage of subsidence. The deformation time series are derived from Sentinel-1 SAR imagery using SABS-InSAR. The observation periods collectively span the years 2017–2026, and detailed statistics of the datasets are provided in Table 1.

Baselines: We compare MPPI against mathematical models, conventional deep learning models, knowledge-guided models, and advanced multimodal forecasting models. The mathematical base-

lines comprise the Weibull, Logistic, and Gompertz temporal functions. The deep learning baselines include ConvLSTM, SwinLSTM, Informer, TimeSformer, PatchTST, and ESTformer. The knowledge-guided baselines consist of PhyDNet and PeRCNN. The advanced time-series forecasting baselines include TimeCAP (2025), TimeLLM (2024), GPT4TS (2023), and LLMTime (2023). We evaluate the prediction accuracy utilizing the root mean squared error (RMSE).

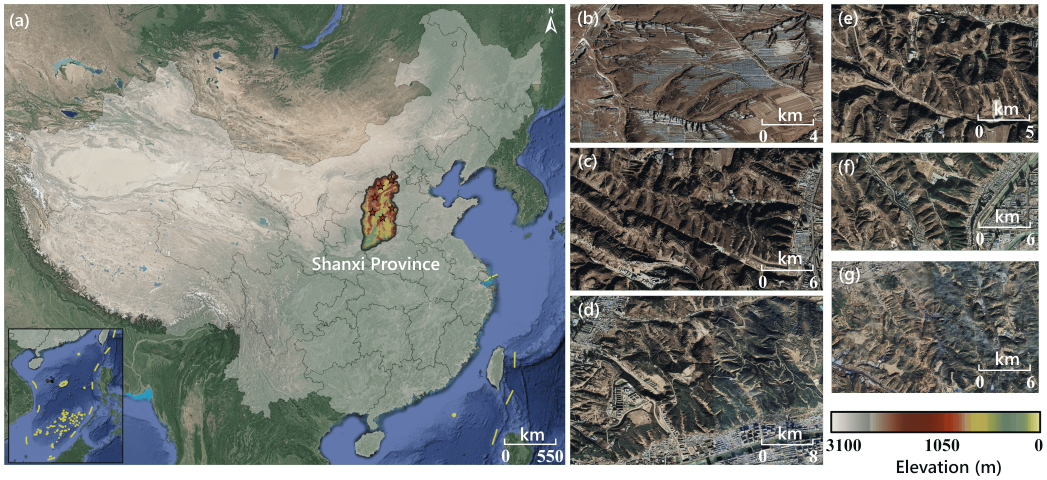

Figure 2: Real-world mining regions used for evaluation. (a) Geographic location and elevation of Shanxi Province, China. (b)–(g) Optical images of the six mining regions. Red markers indicate their locations within Shanxi Province.

Table 1: Dataset statistics for the six mining regions.

Region	Period	Patch size	Total images	Training samples	Test samples
ROI1	2017–2024	128×128	164	42	122
ROI2	2017–2025	128×128	182	76	106
ROI3	2017–2026	128×128	200	84	116
ROI4	2018–2025	128×128	212	82	130
ROI5	2017–2023	128×128	146	55	91
ROI6	2019–2022	128×128	58	38	20

4.3 Evaluation of Physical-Parameter Inversion

The results of the growth-rate inversion indicate that MPPI successfully recovers the deformation evolution across the six real-world mining regions. As illustrated in Figure 3, the estimated maps demonstrate close agreement with the observations, and the difference maps are primarily characterized by near-zero residuals. Errors are predominantly concentrated near local boundaries and fine spatial structures, without exhibiting large-scale spatial displacements. These findings suggest that the multimodal posterior inference extracts physically meaningful rate information from the simulated sequences of subsidence maps and the corresponding textual priors, rather than merely reproducing an approximate average spatial pattern.

Furthermore, the histograms demonstrate that the estimated distribution of r generally follows the observed distribution in terms of both range and dominant peak locations. In regions where the curves exhibit close agreement around the primary peaks, the growth-rate estimates remain relatively stable near the centers of deformation. For regions characterized by broader or more pronounced multimodal distributions, the estimated curves remain consistent with the observations over the principal range; however, the accuracy is constrained in areas exhibiting locally extreme deformation and along the boundaries of deformation. The normalized root mean squared error (NRMSE) of the growth-rate estimates across the six regions is 13%.

Figure 4 presents the inversion results for $W_{\max}$. Across the six regions, the estimated centers of deformation coincide with the observed centers, and the directions of the deformation evolution are largely consistent. The difference maps are predominantly characterized by near-zero residuals, which indicates accurate estimates of maximum subsidence, except in a limited number of areas fea-

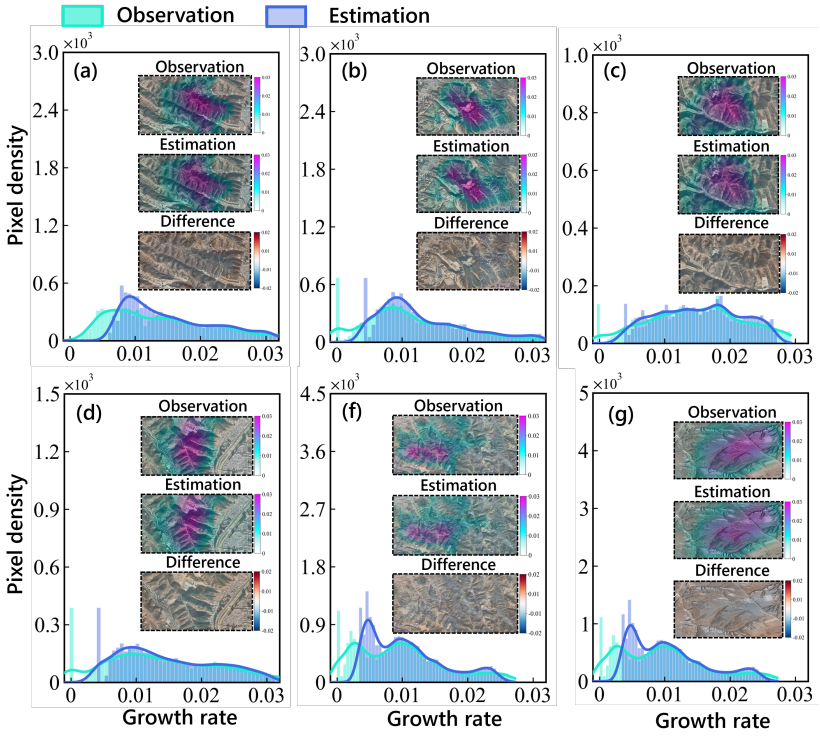

Figure 3: Distributions of the deformation growth rate r estimated by MPPI for ROI1–ROI6 and the corresponding differences from the observations.

turing complex surface conditions. The histograms reveal strong agreement between the estimated and observed distributions of $W_{\max}$ across the study regions, thereby suggesting that MPPI effectively captures the spatial patterns of the deformation evolution. Although the overall shapes of the distributions are preserved, deviations persist in portions of ROI3 and ROI6 that exhibit abrupt deformation or complex surface conditions. The average RMSE between the estimated and reference values of $W_{\max}$ across the six regions is 4 cm.

4.4 Deformation Forecasting in Six Real-World Mining Regions

As illustrated in Figure 5, MPPI consistently predicts the evolution of subsidence across the initial, accelerating, and residual stages within six real-world mining regions. Across ROI1–ROI6, the predicted centers of subsidence, the directions of deformation spread, and the affected areas exhibit strong agreement with the observations. The model successfully captures both the regions of peak subsidence and the progressive spatial expansion of the subsidence basins throughout successive stages. The error maps presented on the right indicate that the majority of deviations occur along the boundaries of the basins and in areas characterized by complex surface conditions, particularly within ROI2, ROI3, and ROI6. Furthermore, the predicted structures exhibiting large gradients of deformation and the geometries of the basins during the residual stage remain closely aligned with the observations. Larger local errors manifest in ROI4 and ROI5, where the deformation is more heterogeneous or involves multiple centers of subsidence. Comparisons across different stages indicate that MPPI preserves the overall spatial structure and the temporal evolution of the subsidence without overfitting fine-scale fluctuations at the boundaries. This behavior aligns with the expected role of physical constraints in nonlinear forecasting scenarios with limited samples.

4.5 Ablation Study

Prior information and representation. Table 2 compares seven configurations to examine the contributions of prior information and its representation. Starting from a vision-only model, we add statistical variables, geological variables, or both, either encoded numerically through an MLP or expressed as text through PTL. The vision-only model yields a $W_{\max}$ estimation RMSE of 18 cm, a

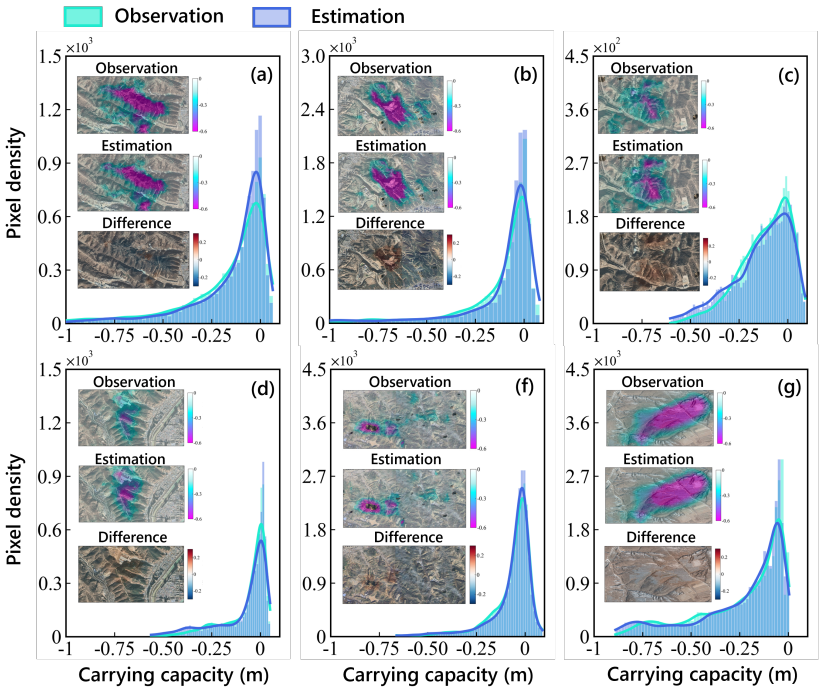

Figure 4: Distributions of the maximum subsidence $W_{\max}$ estimated by MPPI for ROI1–ROI6 and the corresponding differences from the observations.

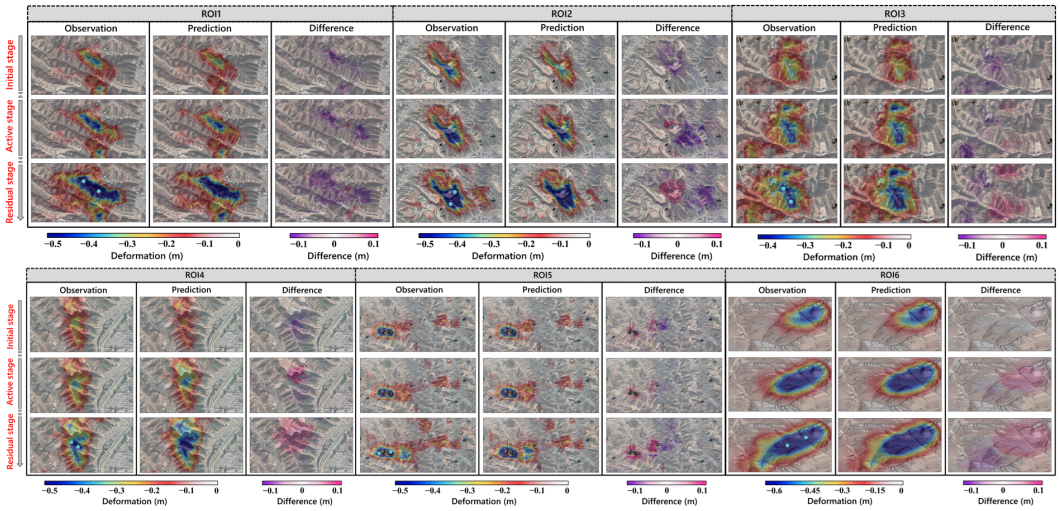

Figure 5: Comparison of deformation predicted by MPPI with observations and the corresponding differences in ROI1–ROI6 during the initial, accelerating, and residual stages. The two cyan circles in each region mark randomly selected surface points used to assess prediction accuracy.

growth-rate NRMSE of 39%, and a deformation-forecasting RMSE of 25 cm, reflecting the difficulty of inferring the full subsidence trajectory from early deformation maps.

Statistical variables encoded through an MLP reduce these errors to 16 cm, 35%, and 22 cm, corresponding to relative reductions of approximately 11%, 10%, and 12%. Thus, explicit summaries of the observed deformation provide useful but modest gains even without text encoding. Expressing the same statistics as text through PTL further reduces the errors to 15 cm, 31%, and 19 cm. The additional reductions are larger for growth-rate estimation and forecasting, suggesting that the textual representation helps extract information about temporal evolution from the observation statistics.

Geological information produces larger gains. Encoding geological variables through an MLP lowers the three errors to 12 cm, 30%, and 16 cm, outperforming both statistical configurations. This

identifies geological information itself as the stronger individual source of complementary evidence. Geological PTL further reduces the errors to 9 cm, 25%, and 11 cm. This improvement is consistent with textual representations helping organize relationships among geological and mining attributes, which jointly constrain the magnitude and evolution of subsidence.

Combining statistical and geological variables through an MLP yields errors of 8 cm, 22%, and 10 cm, outperforming every configuration that uses either source alone. The two sources therefore provide complementary information: statistics characterize the observed deformation state, while geological variables describe the conditions governing its subsequent evolution. The full MPPI model, which represents both sources through PTL and fuses them with visual priors, further lowers the errors to 4 cm, 13%, and 4 cm. Relative to the all-variable MLP, these correspond to reductions of 50%, approximately 41%, and 60%. This comparison supports the joint added value of textual representation and cross-modal fusion when both sources of prior information are available.

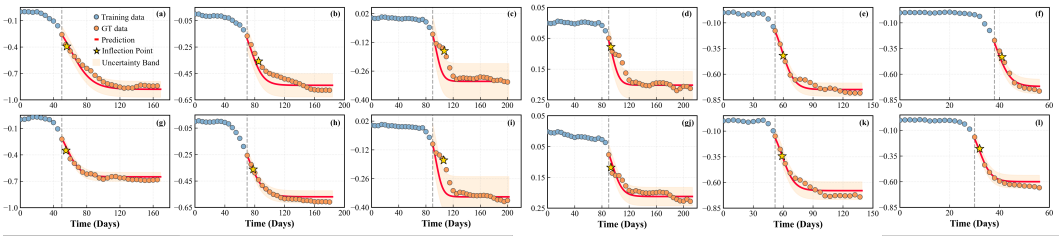

Figure 6: Comparison of deformation time series predicted by MPPI with observations at two selected points in each of ROI1–ROI6. Shaded bands are pointwise 90% prediction intervals, given by the 5th and 95th percentiles of future deformation obtained by propagating posterior parameter samples through PPM. The point locations are marked by cyan circles in Figure 5.

Table 2: Average parameter-estimation and deformation-forecasting errors across six ROIs for different sources and representations of prior knowledge. Numerical variables are encoded through an MLP; textual descriptions are encoded through PTL. Lower values are better.

Prior knowledge	$W_{\max}$ estimation RMSE (cm)	Growth-rate estimation NRMSE (%)	Deformation forecasting RMSE (cm)
Vision only	18	39	25
Vision + statistical variables via MLP	16	35	22
Vision + statistical text via PTL	15	31	19
Vision + geological variables via MLP	12	30	16
Vision + geological text via PTL	9	25	11
Vision + all variables via MLP	8	22	10
Vision + statistical and geological text (Ours)	**4**	**13**	**4**

Posterior inference and deterministic parameter regression. Table 3 examines the contribution of probabilistic parameter learning and uncertainty propagation while keeping the encoders, multimodal fusion, PPM, and inputs unchanged. Deterministic-$\boldsymbol{\theta}$ directly regresses $W_{\max}$ and r. The Gaussian-posterior configuration predicts a mean and variance with probabilistic supervision and passes the parameter mean to PPM. Posterior sampling instead propagates samples from the learned parameter distribution through PPM to obtain a distribution over future deformation.

Replacing deterministic regression with the Gaussian posterior reduces $W_{\max}$ RMSE from 6.0 to 4.0 cm and growth-rate NRMSE from 18.0% to 13.0%. Forecast RMSE decreases from 8.0 to 4.0 cm, and MAE from 5.4 to 3.5 cm. These results support the benefit of probabilistic parameter learning even when prediction uses the posterior mean. The two posterior configurations share a parameter NLL of 0.82 and the same parameter-estimation errors, consistent with changing the decoding rule while retaining the learned parameter distribution.

Propagating posterior samples reduces forecast MAE further to 3.2 cm and the continuous ranked probability score (CRPS) to 2.6 cm, compared with 3.5 cm for posterior-mean prediction. For a point forecast represented as a degenerate distribution, CRPS equals absolute error, explaining its

equality with MAE in the two point-prediction configurations. The lower CRPS of posterior sampling supports the value of retaining predictive dispersion through the physical decoder. Sampling increases forecast RMSE from 4.0 to 4.7 cm, however, indicating a trade-off between squared-error accuracy and the improvements in MAE and probabilistic score.

The sampled 90% prediction intervals attain a prediction interval coverage probability (PICP) of 88.0%, two percentage points below the nominal level, with a mean prediction interval width (MPIW) of 16.0 cm. PICP measures the fraction of observations inside the intervals, and MPIW measures their average width. The correlation between interval width and absolute forecast error is 0.61, indicating that wider intervals tend to accompany larger errors. Together, these results show that posterior propagation provides informative uncertainty estimates, with slight undercoverage at the evaluated 90% level.

Table 3: Ablation of parameter learning and posterior propagation with the same encoders, fusion, PPM, and inputs. Lower errors, NLL, and CRPS are better. Coverage is evaluated against the nominal 90% level; interval width is interpreted jointly with coverage. N/A indicates that the metric is not applicable to the evaluated output.

(a) Parameter estimation and point forecasting

Configuration	$W_{\max}$ RMSE (cm)	r NRMSE (%)	Forecast RMSE (cm)	Forecast MAE (cm)
Deterministic-$\boldsymbol{\theta}$	6.0	18.0	8.0	5.4
Gaussian posterior, mean	**4.0**	**13.0**	**4.0**	3.5
Posterior sampling	**4.0**	**13.0**	4.7	**3.2**

(b) Probabilistic forecasting and interval diagnostics

Configuration	Parameter NLL	CRPS (cm)	90% PICP (%)	90% MPIW (cm)	Width–error ρ
Deterministic-$\boldsymbol{\theta}$	N/A	5.4	N/A	N/A	N/A
Gaussian posterior, mean	0.82	3.5	N/A	N/A	N/A
Posterior sampling	0.82	**2.6**	88.0	16.0	0.61

5 DISCUSSION

Table 4 presents a comparison of the forecasting RMSE between MPPI and the baseline models across six mining regions. The mean errors discussed below represent the arithmetic averages of the RMSE values across the six regions.

Comparison with mathematical models: The Weibull, Logistic, and Gompertz models rely on predefined temporal functions that describe the monotonic evolution; however, they exhibit limited flexibility in accommodating local fluctuations and transitions between stages. The corresponding mean RMSE values are 15.6, 14.9, and 16.7 cm, respectively, all of which exceed the 4.2 cm achieved by MPPI. On average, the Logistic model yields a slightly superior performance compared to the other two mathematical models. These results indicate that a single analytical function is insufficient to fully capture the evolution of deformation in this task, particularly in scenarios where only limited historical observations are available.

Comparison with deep learning models: Although deep learning models possess the capability to extract temporal dependencies and spatiotemporal features, the associated mean RMSE values range from 33.2 to 39.5 cm in these experiments. Consequently, the resulting errors exceed those of both the mathematical baselines and MPPI. Within this group, ESTformer achieves the lowest mean RMSE at 33.2 cm, followed by TimeSformer at 36.6 cm and PatchTST at 37.4 cm. The performance across these models varies across the regions of interest, which suggests that numerical time-series modeling alone exhibits a limited capacity to account for regional differences and external semantic information under the conditions of small sample sizes. MPPI achieves the lowest error in every region of interest, thereby demonstrating a more consistent predictive performance across the six evaluated regions.

Comparison with knowledge-guided models: PeRCNN and PhyDNet yield mean RMSE values of 43.3 and 43.5 cm, respectively, which exceed the values associated with the other deep learning base-

lines. Although the integration of physical constraints enhances interpretability, the effectiveness of such constraints relies on the agreement between the prior assumptions and the observed processes. Furthermore, restrictive constraints may limit the adaptation to complex patterns of deformation. This comparison corroborates the effectiveness of combining data-driven representations, multimodal priors, and the inference of physical parameters in this task. However, the prediction errors alone do not sufficiently establish which physical assumptions are responsible for the suboptimal performance of the baselines.

Comparison with multimodal time-series forecasting models: TimeCAP, TimeLLM, GPT4TS, and LLMTime achieve mean RMSE values of 12.4, 9.0, 11.6, and 15.4 cm, respectively. Although these models outperform the conventional deep learning baselines, the associated errors remain higher than the 4.2 cm achieved by MPPI. Existing approaches primarily rely on prompting or the adaptation of language models, whereas MPPI explicitly incorporates multimodal evidence into the inference of the parameters that govern the deformation. Through the integration of information regarding local changes, long-term trends, and regional conditions, MPPI achieves the lowest RMSE across all six regions of interest. Consequently, these results validate the utility of linking multimodal priors to physically meaningful parameters for the purpose of forecasting from limited observations.

Table 4: Time-series forecasting results across six mining regions, measured by RMSE (cm). Lower values indicate better performance.

Method	ROI1	ROI2	ROI3	ROI4	ROI5	ROI6
MPPI (Ours)	2.8	5.6	4.3	4.5	3.7	4.3
TimeCAP (2025)	10.4	13.2	12.5	12.1	12.8	13.6
TimeLLM (2024)	7.2	10.8	7.5	10.5	8.1	9.9
GPT4TS (2023)	9.6	13.6	10.7	12.9	10.2	12.5
LLMTime (2023)	10.5	18.4	15.3	15.8	16.4	15.9
ESTformer (2023)	25.4	39.8	32.7	34.6	33.6	33.2
PatchTST (2023)	29.6	42.5	37.8	38.2	37.9	38.3
TimeSformer (2023)	27.6	36.9	41.1	37.2	37.8	39.2
Informer (2021)	28.6	41.8	41.9	42	41.5	40.8
SwinLSTM (2021)	32.1	39.5	40.5	43.6	39.5	41.9
ConvLSTM (2015)	29.7	41.4	42.9	40.8	38.4	41.7
PeRCNN (2021)	31.2	49.6	43.5	45.2	41.7	48.4
PhyDNet (2020)	34.2	51.3	41.1	46.3	42.6	45.2
Weibull	12.8	24.6	13.5	15.9	12.2	14.8
Logistic	14.2	20.8	14.1	14.9	12.3	12.9
Gompertz	13.4	24.1	15.6	16.5	14.2	16.4

6 Conclusion

We introduced MPPI for small-sample InSAR deformation forecasting, linking multimodal priors to posterior inference over maximum subsidence and growth rate. Its central insight is to use visual patterns, observation statistics, and geological context to constrain the physical quantities governing future deformation. A physical decoder then translates the inferred temporal parameters and fitted spatial parameters into deformation forecasts. Experiments on six coal-mining regions show improved parameter estimation, while baseline comparisons yield the lowest forecasting RMSE in each region. The modality ablations identify geological text as the stronger individual source of complementary information, with statistical and geological text together producing the lowest errors. These findings support physical-parameter inference as a useful interface between multimodal representations and deformation dynamics. Extending evaluation to broader geological settings and assessing posterior uncertainty calibration are concrete next steps.

AI Use Statement

Generative AI assistance was used for language translation and LaTeX formatting. The authors reviewed the resulting text, equations, citations, and formatting and take responsibility for the final manuscript.

REFERENCES

M. Ausloos. Gompertz and verhulst frameworks for growth and decay description. *arXiv preprint*, 2011.

S. Bai, Y. Cai, R. Chen, et al. Qwen3-vl technical report. *Preprint*, 2025.

Konstantinos Benidis et al. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36, 2022.

P. Berardino, G. Fornaro, R. Lanari, and E. Sansosti. A new algorithm for surface deformation monitoring based on small baseline differential sar interferograms. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11):2375–2383, 2003.

G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? *International Conference on Machine Learning*, 2021.

C. Cercignani. *The Boltzmann Equation and Its Applications*. Springer, 1988.

Xiaoyong Dong et al. High-precision prediction method for mine deformation based on gnss rtk and stacking ensemble learning. *Scientific Reports*, 16(1):12277, 2026.

Zhongyang Han et al. A review of deep learning models for time series prediction. *IEEE Sensors Journal*, 21(6):7833–7848, 2019.

Kethmi Hirushini Hettige et al. Airphynet: Harnessing physics-guided neural networks for air quality prediction. In *International Conference on Learning Representations*, 2024.

Qingfeng Hu et al. Quantitative and dynamic predictive model for mining-induced movement and deformation of overlying strata. *Engineering Geology*, 311:106876, 2022.

M. Jin, S. Wang, L. Ma, et al. Time-llm: Time series forecasting by reprogramming large language models. *ICLR*, 2024.

V. Le Guen and N. Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. *CVPR*, 2020.

D. Li, Z. Zeng, Z. Wang, et al. Estformer: Transformer utilising spatiotemporal dependencies. *Pattern Recognition*, 2025a.

Ruikun Li et al. Predicting the energy landscape of stochastic dynamical system via physics-informed self-supervised learning. In *International Conference on Learning Representations*, 2025b.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

F. Liu, D. Chen, Z. Guan, et al. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

T. Masterlark. Magma intrusion and deformation predictions: Sensitivities to the mogi assumptions. *Journal of Geophysical Research: Solid Earth*, 112(B6), 2007.

Y. Nie, N. H. Nguyen, P. Sinthong, et al. A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations*, 2022.

B. Osmanoğlu, F. Sunar, S. Wdowinski, et al. Time series analysis of insar data: Methods and trends. *ISPRS Journal*, 115:90–102, 2016.

C. Rao, P. Ren, Q. Wang, et al. Encoding physics to learn reaction-diffusion processes. *Nature Machine Intelligence*, 2023.

A. Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

F. M. Shiri, T. Perumal, N. Mustapha, et al. A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru. *Artificial Intelligence Review*, 2023.

A. Tsoularis and J. Wallace. Analysis of logistic growth models. *Mathematical Biosciences*, 179 (1):21–55, 2002.

Yi Wang et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, 2024.

J. Weston, A. M. Ferreira, and G. J. Funning. Systematic comparisons of earthquake source models determined using insar and seismic data. *Geophysical Journal International*, 532:61–81, 2012.

Xingjian Wu et al. Aurora: Towards universal generative multimodal time series forecasting. In *International Conference on Learning Representations*, 2026.

Y. Xiang, Z. Ouyang, J. Li, et al. Insar-based prediction of time-series displacements using a new physics-informed neural network with prior parameter inversion. *Remote Sensing*, 18:2516, 2026.

A. Yang, A. Li, B. Yang, et al. Qwen3 technical report. *Preprint*, 2025.

Z. Zhang, Q. Pan, Z. Yang, et al. Physics-informed deep learning method for predicting tunnelling-induced ground deformations. *International Journal of Geomechanics*, 18(9):4957–4972, 2023.

Z. Zhang, D. Dias, C. Anitescu, et al. Physics-informed neural network for elastic-plastic mesh-free modelling of tunnelling-induced deformation. *International Journal for Numerical and Analytical Methods in Geomechanics*, 21(1):349–370, 2026.

S. Zhong, W. Ruan, M. Jin, et al. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *arXiv preprint*, 2025.

H. Zhou, S. Zhang, J. Peng, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. *AAAI Conference on Artificial Intelligence*, 2021.

J. Zhou, C. Song, Z. Li, et al. A new dynamic probability integral model for three-dimensional deformation monitoring of loess mining subsidence using insar. *Remote Sensing*, 2026.

T. Zhou, P. Niu, L. Sun, et al. One fits all: Power general time series analysis by pretrained lm. *NeurIPS*, 2023.