GENERATIVE PRIORS CONDITIONED ON NATURAL LANGUAGE FOR BAYESIAN INVERSION IN PDES

Anonymous authors
Paper under double-blind review

ABSTRACT

Inferring quantities of interest (QoI) from data is a central task in Science and Engineering. In such contexts, we often have access to both quantitative data and qualitative data. Quantitative data may be represented by noisy sensor measurements, simulation data, re-analysis data; qualitative data may be in the form of text descriptions of experimental setups, expected experiment outcomes, and human-perceived system behaviours. The task we address in this paper is the following. Given a training set of paired qualitative text and quantitative QoI data, we learn to exploit the inherent correlation between the two modalities to learn a highly informative data-driven natural-language-conditional Bayesian prior, such that when presented with a new physical system, we can coherently combine (i) the training dataset, (ii) qualitative text describing the new system, and (iii) a small number of noisy sensor readings from that new system, to perform inference and uncertainty quantification (UQ) over the QoI. To achieve this task, we develop two parallel approaches, one uses conditional diffusion and the other conditional autoencoders, and compare both against classical Bayesian methodology, unconditional generative models and deterministic supervised methods. Each approach has specific strengths and tradeoffs; conditional autoencoder offers theoretical tractability, allows for fast posterior sampling, and provides better-calibrated UQ, whereas conditional diffusion is explored for greater expressiveness and capturing complex posteriors with irregular QoI fields. The approach is tested on the steady-state heat equation, damped Helmholtz equation, and UK weather reanalysis data.

1 INTRODUCTION

Recovering a full field of quantities of interest (QoI) from a small number of noisy measurements is a ubiquitous problem setup in science and engineering applications, such as in weather prediction (Kalnay, 2003; Bannister, 2017; Manshausen et al., 2025), medical image reconstruction (Hammernik et al., 2018; Arridge et al., 2019; Song et al., 2022), and structural health monitoring (Girolami et al., 2021; Febrianto et al., 2022). Such problems are typically severely ill-posed as the observation process is not a one-to-one map. Leveraging data helps attenuate the ill-posedness of these tasks. A direct approach is to learn the inverse map in a supervised fashion, regressing the full field directly from observations (Erichson et al., 2020; Fukami et al., 2021; Santos et al., 2023). Such methods are fast at test time, however, the map they learn is tied to the observation process it was trained under; they are also deterministic, whereas uncertainty quantification (UQ) is essential for decision making. Bayesian inference (Stuart, 2010; Cotter et al., 2010; Bui-Thanh & Girolami, 2014) provides UQ by construction: a likelihood function measuring data fit after an observation process and a prior encoding prior beliefs on QoI data, together give a posterior distribution.

One of the challenges in Bayesian inference is prior specification, with very sparse observations the prior determines most of the reconstruction, and one that is too vague leaves the posterior diffuse with relatively poor accuracy. Prior elicitation (O'Hagan et al., 2006; Mikkola et al., 2024) converts subjective domain knowledge into a prior distribution. Large Language Models (LLMs) (Brown et al., 2020) have recently been prompted to supply that knowledge in place of an expert (Capstick et al., 2025; Selby et al., 2025; Huang, 2025). However, such methods are confined to a few scalar parameters and are correspondingly insufficient to capture the complex spatial structures in physical fields. A promising direction is to replace hand-crafted priors with learned generative models trained on historical dataset (Daras et al., 2024). Such priors are expressive and scalable, and are now

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.

used widely in imaging (Bora et al., 2017; Chung et al., 2023; Feng et al., 2023) and in physical setups (Shu et al., 2023; Shysheya et al., 2024; Akyildiz et al., 2025; Glyn-Davies et al., 2025; Vadeboncoeur et al., 2026b).

System-specific information can also improve the generative prior, such as a system's geometry (Vadeboncoeur et al., 2026a). Domain expert knowledge can also serve as additional information. In many practical scenarios, prior information can be naturally expressed through human knowledge in natural language, such as a description of the experimental setup ("a strong forcing acts in the north-east"), or a human-perceived physical condition ("the north-west is cool with a light north-westerly breeze"). Such statements are imprecise, but they carry high-level physical structural patterns.

In this work, we propose a methodology that encodes such domain knowledge via text into a data-driven physical prior. Trained on paired text and field data, a conditional generative model learns the correlation between the two modalities, so that a qualitative description of a new system induces an informative prior over its field. The learned prior constrains the search space, which is further used in posterior inference from sparse observational data to enable high-fidelity field reconstruction with UQ. Contributions in this paper are as follows.

C1. We propose to learn a text-conditional prior on physical fields using generative models, combining qualitative descriptions with sparse noisy quantitative measurements to perform posterior inference over quantities of interest.

C2. We provide a complete pipeline: fine-tuning a pretrained text encoder, training the text-conditional prior, and posterior inference.

C3. Two parallel approaches based on autoencoders and diffusion models are developed. With autoencoders, Bayesian inference is performed in a low-dimensional learned latent space, where, given the pushforward form of the prior, the posterior admits asymptotically exact sampling; with diffusion models, posterior inference relies on approximate likelihood guidance.

C4. We analyse the advantages and tradeoffs of two models. For autoencoders we provide theoretical guarantees and error bounds, and obtain more consistently calibrated posteriors at a lower sampling cost, while diffusion models offer greater expressiveness on irregular fields.

C5. We test the methods on the steady-state heat equation, the damped Helmholtz equation, and UK weather reanalysis data, comparing against Gaussian processes, unconditional data-driven prior models and supervised models, with both accuracy and posterior coverage reported.

Section 1.1 illustrates our problem setup; Section 1.2 reviews related works; Section 2 presents the proposed pipeline of two parallel approaches; Section 3 provides additional implementation details; Section 4 reports and discusses results on three experiments; Section 5 draws final conclusions.

1.1 Problem Setup

Consider a collection of N high-fidelity experimental setups where each experiment produces quantitative data $u \in \mathcal{U}$, which is associated with a set of physical parameters θ,[1] of which only vague and qualitative information is available in the form of the text $s \in \mathcal{S}$, where $\mathcal{S}$ is a set of text sequences. We think of s as being obtained from a human expert relating an observed physical quantity to a text description. Importantly, s is an imprecise and random specification of information of relevance to u. The set $\{u^{(n)}, s^{(n)}\}_{n=1}^{N}$ will be our training dataset. One may then Euclideanise text using a map $\mathsf{G} : \mathcal{S} \to \mathcal{T}$, where $\mathcal{T}$ is d_τ-dimensional Euclidean space. In our experiments, G is a fine-tuned text encoder. Table 4 summarises our notation, with an example of each symbol.

Now consider a new, partially observed experiment, of which we observe a small amount of quantitative data, $y^{(o)} \in \mathbb{R}^{d_y}$ for $o > N$, and text description $s^{(o)}$. We wish to leverage the dataset $\{u^{(n)}, s^{(n)}\}_{n=1}^{N}$ in the form of a text-conditional prior, to help infer $u^{(o)}$ from $(y^{(o)}, s^{(o)})$ where

$$y^{(o)} = \mathsf{H}^{(o)} u^{(o)} + \xi^{(o)}, \quad \xi^{(o)} \sim \eta, \tag{1}$$

with $\mathsf{H}^{(o)}$ representing the system-dependent observation operator, η a known noise distribution, taken to be $\mathcal{N}(0, \sigma_y^2 \mathrm{I}_{d_y})$ in our experiments. During inference, we observe $y^{(o)}, s^{(o)}$, we know $\mathsf{H}^{(o)}, \eta$, and do not know $u^{(o)}, \xi^{(o)}$. Example 1.1 is a toy example of our problem setup.

[1] The relation between u and θ is left unspecified: θ is not a fixed representation of the system, and could take any form such as a parameterisation of the experimental setup or a coarse property of u itself.

Example 1.1. *The quantity of interest which we would like to estimate, $u^{(o)}$, is the field which describes the heat distribution across the surface of a plate; the observed quantitative data, $y^{(o)}$, is a small number of noisy point-wise sensor measurements; the unknown physical parameter $\theta^{(o)}$ is the temperature at the boundaries of the plate, of which we only know a vague, qualitative description such as* ``The left boundary is quite hot, while the bottom boundary is cold'', *etc., this is $s^{(o)}$. The task: from a corpus of data, $\{u^{(n)}, s^{(n)}\}_{n=1}^{N}$, learn a text-conditioned prior for u, and use this to infer $u^{(o)}$ from $(y^{(o)}, s^{(o)})$.*

1.2 Related Works

Generative Priors for Inverse Problems. Generative models are widely used as data-driven priors in inverse problems. One approach uses Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), learning a generator that maps a low-dimensional latent vector to the field and performing inference in that latent space. Approaches include gradient descent (Bora et al., 2017) and Markov chain Monte Carlo (MCMC) sampling (Patel & Oberai, 2021; Patel et al., 2022; Meng et al., 2022). GABI (Vadeboncoeur et al., 2026a) follows a similar idea with a graph autoencoder. There exists theory about the relation between learned priors and the resulting posteriors (Hosseini & Huang, 2026). Diffusion models (Ho et al., 2020; Song et al., 2021) are the dominant alternative as a generative prior, with many training-free schemes for conditioning on measurements at inference time (Song et al., 2022; Kawar et al., 2022; Feng et al., 2023; Mardani et al., 2024; Feng & Bouman, 2024). In our diffusion approach we follow Chung et al. (2023) and approximate the likelihood score via Tweedie's formula (Efron, 2011; Song et al., 2023; Chung et al., 2024a; Rozet & Louppe, 2023; Boys et al., 2024), which yields a biased posterior approximation; asymptotically exact SMC alternatives (Wu et al., 2023; Cardoso et al., 2024) instead require maintaining an ensemble of particles at a cost. In all of these works the prior itself is unconditional, and we propose to condition on text descriptions. Furthermore, mostly developed for imaging, these methods are also seldom assessed for UQ, which in physical settings is of paramount importance.

Conditional Generative Models. Block triangular maps (Zech & Marzouk, 2022a;b) provide a principled construction for conditional sampling, realised with monotone GANs (Baptista et al., 2024) and Wasserstein autoencoders (Tolstikhin et al., 2018; Al-Jarrah et al., 2026). Our autoencoder method draws inspiration from these works. Furthermore, conditional diffusion models have shown superior performance in text-to-image generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022), establishing the paradigm of conditioning a generative model on text embeddings, though these target forward generation. A few works bring text into inverse problems (Chung et al., 2024b; Kim et al., 2025), which use a pretrained latent diffusion model (LDM) (Rombach et al., 2022) as a prior and introduce the text at inference time as a guidance or regularisation. Both target imaging rather than physical fields.

Natural Language and Physical Systems. In the scientific domain, although generative models have been widely applied to forward and inverse problems in physical systems (Shu et al., 2023; Huang et al., 2024; Shysheya et al., 2024; Bastek et al., 2025; Jiang et al., 2024; Jacobsen et al., 2025; Zheng et al., 2025; Shan et al., 2026), the integration of natural language with physical simulation remains underexplored. Emerging multimodal approaches including PROSE (Liu et al., 2024b) and the Multimodal PDE Foundation Model (Negrini et al., 2025) combine numerical and symbolic or textual data to approximate solution operators for ODEs and PDEs, and Text2PDE (Zhou et al., 2025) uses LDM to perform physical simulations from linguistic prompts, all targeting forward simulation. On the inverse side, Saito et al. (2026) recently condition a Darcy-flow solver on geological descriptions encoded by a sentence embedding, using a deterministic U-Net without UQ. Learning a natural-language-conditioned prior to guide the inverse recovery of physical fields, in a form that supports UQ, remains open.

2 Methodology

This section presents our full pipeline, from fine-tuning a pretrained text encoder (Step 1) to conditional prior learning and posterior inference (Steps 2 and 3). Step 1 is shared by both models; Steps 2 and 3 are then described for autoencoder (AE) and score-based diffusion model (SD) in turn.

2.1 STEP 1: FINE-TUNING

As discussed in Section 1.1, the training data consists of pairs of physical fields and text descriptions $\{u^{(n)}, s^{(n)}\}_{n=1}^{N}$ drawn from a joint distribution $\mathbb{P}_{u,s}$. To make the text usable by a model we encode it with $\mathsf{G}^\phi : \mathcal{S} \to \mathcal{T}$, mapping the text to a fixed-dimensional latent representation τ. This induces a joint distribution over fields and text embeddings, $\mathbb{P}_{u,\tau} = (\mathsf{Id} \times \mathsf{G}^\phi)_\# \mathbb{P}_{u,s}$. [2] We take G^ϕ to be a pretrained sentence transformer (Reimers & Gurevych, 2019).[3] However, proximity in $\mathcal{T}$ reflects semantic similarity rather than physical similarity, so we fine-tune the encoder on the training dataset before training prior models. Parameters ϕ in the encoder are obtained by

$$\phi^\star \in \arg\min_{\phi} \mathbb{E}_{(u,s),(u',s')\sim\mathbb{P}_{u,s}} \left| \mathsf{score}_{\text{field}}(u,u') - \mathsf{score}_{\text{text}}(\mathsf{G}^\phi(s), \mathsf{G}^\phi(s')) \right|^2, \tag{2}$$

a cosine-similarity loss in which $\mathsf{score}_{\text{text}}(\tau,\tau') = (\tau^\top \tau')/(\|\tau\|\|\tau'\|)$ is the cosine similarity between text embeddings. $\mathsf{score}_{\text{field}}(\bullet,\bullet)$ is chosen appropriately for each physical setup, detailed in Appendix G, in each case it returns a value in $[0,1]$ quantifying how similar two physical fields are.

2.2 AUTOENCODER

2.2.1 STEP 2: LEARNING CONDITIONAL PRIOR – AUTOENCODER

A map $\mathsf{P} : \mathcal{Z} \times \mathcal{T} \to \mathcal{U} \times \mathcal{X}$ with $\mathsf{P}(z,\tau) = \big(\mathsf{P}_{\mathcal{U}}(z, \mathsf{P}_{\mathcal{X}}(\tau)),\ \mathsf{P}_{\mathcal{X}}(\tau)\big)$ is block-triangular (Baptista et al., 2024; Zech & Marzouk, 2022a;b). Our autoencoder consists of an encoder and a decoder,

$$\mathsf{E} : \mathcal{U} \times \mathcal{T} \to \mathcal{Z}, \qquad \mathsf{D} : \mathcal{Z} \times \mathcal{T} \to \mathcal{U}, \tag{3}$$

both taking the text embedding as input; these are the networks that are trained and evaluated. Distributions, however, are compared on the joint space, for which we write the maps

$$\widetilde{\mathsf{E}} : \mathcal{U} \times \mathcal{T} \to \mathcal{Z} \times \mathcal{T}, \qquad \widetilde{\mathsf{D}} : \mathcal{Z} \times \mathcal{T} \to \mathcal{U} \times \mathcal{T}, \tag{4}$$

so that $\widetilde{\mathsf{E}}(u,\tau) = (\mathsf{E}(u,\tau),\tau)$ and $\widetilde{\mathsf{D}}(z,\tau) = (\mathsf{D}(z,\tau),\tau)$, with no new parameters because the text is passed through unchanged. They have the form of a block-triangular map with $\mathsf{P}_{\mathcal{X}} = \mathsf{Id}_{\mathcal{T}}$. The maps E, D are learned by minimising

$$\mathsf{J}_{\mathsf{AE}}(\mathsf{D},\mathsf{E}) = \mathbb{E}_{u,\tau\sim\mathbb{P}_{u,\tau}} \left\| u - (\mathsf{D} \circ \widetilde{\mathsf{E}})(u,\tau) \right\|^2 + \mathsf{d}\big(\widetilde{\mathsf{E}}_\#(\mathbb{P}_{u,\tau}),\ \mathbb{Q}_z \times \mathbb{P}_\tau\big), \tag{5}$$

where $\mathsf{d}(\bullet,\bullet)$ is a discrepancy on probability measures over $\mathcal{Z} \times \mathcal{T}$, and we choose maximum mean discrepancy (MMD), detailed in Appendix D. $\mathbb{Q}_z$ is a probability measure on the latent space, taken to be $\mathcal{N}(0,\mathrm{I})$.

Theorem 2.1. *Let $\mathcal{U}, \mathcal{T}, \mathcal{Z}$ be separable Banach spaces, let $(\mathcal{U} \times \mathcal{T}, \mathcal{B}(\mathcal{U} \times \mathcal{T}), \mathbb{P}_{u,\tau})$ and $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mathbb{Q}_z)$ be probability spaces, and let D, E in (3) be continuous. If $\mathsf{J}_{\mathsf{AE}}(\mathsf{D},\mathsf{E}) = 0$, then $\mathbb{P}_{u|\tau} = \mathsf{D}(\bullet,\tau)_\# \mathbb{Q}_z$ for $\mathbb{P}_\tau$-a.e. τ.*

Theorem 2.1 shows that if the loss (5) is brought to zero, the AE method recovers the correct conditional distribution $\mathbb{P}_{u|\tau}$ from the joint data distribution $\mathbb{P}_{u,\tau}$. Proof is provided in Appendix B.1.

Theorem 2.2. *We have the following bound on the squared 2-Wasserstein distance between the data distribution and the generative model:*

$$\begin{aligned} \mathsf{W}_2^2(\mathbb{P}_{u,\tau}, \widetilde{\mathsf{D}}_\#(\mathbb{Q}_z \times \mathbb{P}_\tau)) \le & 2\mathbb{E}_{u,\tau\sim\mathbb{P}_{u,\tau}} \|u - (\mathsf{D} \circ \widetilde{\mathsf{E}})(u,\tau)\|^2 \\ & + 2(\|\mathsf{D}\|_{\mathrm{Lip}}^2 + 1)\ \mathsf{W}_2^2(\widetilde{\mathsf{E}}_\#(\mathbb{P}_{u,\tau}), \mathbb{Q}_z \times \mathbb{P}_\tau). \end{aligned} \tag{6}$$

This bounds the error between the joint distributions by terms related to ones that appear in our loss (5). The next results bridge the gap between MMD and Wasserstein, for distributions with compact support. Proofs are available in Appendix B.2 and B.3.

Corollary 2.3. *We have the following bound for latent distributions with compact support:*

$$\begin{aligned} \mathsf{W}_2^2(\mathbb{P}_{u,\tau}, \widetilde{\mathsf{D}}_\#(\mathbb{Q}_z \times \mathbb{P}_\tau)) \le & 2\mathbb{E}_{u,\tau\sim\mathbb{P}_{u,\tau}} \|u - (\mathsf{D} \circ \widetilde{\mathsf{E}})(u,\tau)\|^2 \\ & + 2C(\|\mathsf{D}\|_{\mathrm{Lip}}^2 + 1)\mathsf{MMD}_{\mathrm{RBF}}^{2\alpha}(\widetilde{\mathsf{E}}_\#(\mathbb{P}_{u,\tau}), \mathbb{Q}_z \times \mathbb{P}_\tau), \end{aligned} \tag{7}$$

for some $C > 0, \alpha \in (0,1]$, where α depends critically on $\dim(\mathcal{Z} \times \mathcal{T})$.

[2] Notational conventions are collected in Appendix A.

[3] We use *all-MiniLM-L6-v2* which provides text embeddings of 384 dimensions from `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

Algorithm 1 Posterior sampling — autoencoder

Require: trained $\mathsf{D}, \mathsf{G}^{\phi}$; text s, observation y, operator H, noise σ_y; M_o NUTS warmup steps, M NUTS posterior draws

$\tau \leftarrow \mathsf{G}^{\phi}(s)$
$z_0 \sim \mathbb{Q}_z$
for $i = 1$ **to** $M_o + M$ **do**
  $z_i \leftarrow \text{NUTS-STEP}\big(\mathbb{P}_{z|y,\tau},\ z_{i-1}\big)$
    ▷ posterior target defined in Lemma 2.4
end for
for $m = 1$ **to** M **do**
  $u^{(m)} \leftarrow \mathsf{D}(z_{M_o+m}, \tau)$
end for
return $\{u^{(m)}\}_{m=1}^{M}$

Algorithm 2 Posterior sampling — diffusion

Require: trained $\mathsf{V}, \mathsf{G}^{\phi}$; text s, observation y, operator H, noise σ_y; grid $T = t_L > \cdots > t_0 = \varepsilon$, cap c, M posterior draws; write u_i for u_{t_i}

$\tau \leftarrow \mathsf{G}^{\phi}(s)$
for $m = 1$ **to** M **do**
  $u_L \sim \mathcal{N}(0, \mathrm{I})$
  for $i = L$ **to** 1 **do**
    $\hat{u}_0 \leftarrow \text{TWEEDIE ESTIMATE}(u_i, t_i, \tau)$
    $\delta \leftarrow \text{LIKELIHOOD GUIDANCE}(\hat{u}_0, y, c)$
    $u_{i-1} \leftarrow \text{SDE STEP}(u_i, t_i, \tau, \delta)$
  end for
  $u^{(m)} \leftarrow \text{TWEEDIE ESTIMATE}(u_0, t_0, \tau)$
end for
return $\{u^{(m)}\}_{m=1}^{M}$

2.2.2 STEP 3: BAYESIAN INFERENCE – AUTOENCODER

After learning the conditional prior $\mathbb{P}_{u|\tau}$, given observations y from a new system obtained from (1), the conditional autoencoder approach allows for approximate sampling of the posterior $\mathbb{P}_{u|y,\tau}$ by first sampling the posterior in the latent space $\mathbb{P}_{z|y,\tau}$ using standard MCMC tools, such as NUTS (Hoffman et al., 2014), and then pushing the obtained samples through the decoder. This result follows directly from Vadeboncoeur et al. (2026a). Algorithm 1 shows the inference process for AE.

Lemma 2.4. *Let $\mathbb{P}_{u|y,\tau}$ be the posterior on $\mathcal{U}$, with likelihood proportional to $\exp(-\Phi(u; y))$,* [4] *and prior $\mathbb{P}_{u|\tau} = \mathsf{D}(\bullet, \tau)_{\#}\mathbb{Q}_z$ as in Theorem 2.1. Then, the posterior $\mathbb{P}_{u|y,\tau} = \mathsf{D}(\bullet, \tau)_{\#}\mathbb{P}_{z|y,\tau}$ with the latent posterior $\mathrm{d}\mathbb{P}_{z|y,\tau}(z) \propto \exp(-\Phi(\mathsf{D}(z, \tau); y))\mathrm{d}\mathbb{Q}_z(z)$.*

2.3 SCORE-BASED DIFFUSION MODEL

2.3.1 STEP 2: LEARNING CONDITIONAL PRIOR – SCORE-BASED DIFFUSION MODEL

The second construction targets $\mathbb{P}_{u|\tau}$ via diffusion. We adopt the variance-preserving (VP)-SDE of Song et al. (2021), the continuous-time limit of DDPM (Ho et al., 2020), which progressively perturbs a field $u_0 \sim \mathbb{P}_{u|\tau}$ into noise,

$$\mathrm{d}u = -\tfrac{1}{2}\beta(t)\, u\, \mathrm{d}t + \sqrt{\beta(t)}\, \mathrm{d}w, \qquad t \in [0, T], \tag{8}$$

where $\beta(t) > 0$ is the noise schedule and w a Wiener process on $\mathcal{U}$. Linearity of (8) gives a Gaussian perturbation kernel $p_{0t}(u_t|u_0) = \mathcal{N}(\sqrt{\bar{\alpha}(t)}u_0, (1 - \bar{\alpha}(t))\mathrm{I})$ with $\bar{\alpha}(t) = \exp(-\int_0^t \beta(r)\mathrm{d}r)$, so that $\bar{\alpha}(T) \approx 0$ and $u_T \sim p_T \approx \mathcal{N}(0, \mathrm{I})$. The forward process does not depend on τ and the text enters only through the reverse process. Samples from $\mathbb{P}_{u|\tau}$ are obtained by integrating the reverse-time SDE from u_T down to $t = 0$, following

$$\mathrm{d}u = \big[-\tfrac{1}{2}\beta(t)\, u - \beta(t)\, \nabla_{u_t} \log p_t(u_t|\tau)\big]\, \mathrm{d}t + \sqrt{\beta(t)}\, \mathrm{d}\bar{w}, \tag{9}$$

which requires the conditional score $\nabla_{u_t} \log p_t(u_t|\tau)$. We approximate it with a network $\mathsf{V} : \mathcal{U} \times [\varepsilon, T] \times \mathcal{T} \to \mathcal{U}$ trained by denoising score matching, with objective

$$\mathsf{J}_{\mathsf{SD}}(\mathsf{V}) = \mathbb{E}_{t\sim\mathrm{Unif}[\varepsilon,T]}\Big\{\lambda(t)\, \mathbb{E}_{u_0,\tau\sim\mathbb{P}_{u,\tau}}\, \mathbb{E}_{u_t|u_0}\Big[\big\|\mathsf{V}(u_t, t, \tau) - \nabla_{u_t} \log p_{0t}(u_t|u_0)\big\|_2^2\Big]\Big\}, \tag{10}$$

where $\varepsilon \simeq 0$ is a small positive constant to avoid exploding. In practice, we train the equivalent ϵ-parameterisation $\epsilon_{\mathsf{V}} = -\sigma(t)\, \mathsf{V}$ with $\lambda(t) = \sigma^2(t) = 1 - \bar{\alpha}(t)$, which avoids the variance blow-up of the score target as $t \to 0$ (Ho et al., 2020; Zhang & Chen, 2023; Rozet & Louppe, 2023), but for notational convenience we retain the score notation V throughout. At inference, (9) is discretised into L steps via the Euler-Maruyama scheme on a uniform temporal grid $T = t_L > \cdots > t_0 = \varepsilon$, and we write u_i for u_{t_i}.

[4] $\Phi(u; y) = \frac{1}{2\sigma_y^2}\|y - \mathsf{H}u\|_2^2$ is the potential associated with the likelihood in (1).

2.3.2 STEP 3: BAYESIAN INFERENCE – DIFFUSION POSTERIOR SAMPLING

To sample from $\mathbb{P}_{u|y,\tau}$, we replace the prior score in (9) by the posterior score using Bayes' rule

$$\nabla_{u_t} \log p_t(u_t|y,\tau) = \underbrace{\nabla_{u_t} \log p_t(u_t|\tau)}_{\approx \mathsf{V}(u_t,t,\tau)} + \nabla_{u_t} \log p_t(y|u_t,\tau). \tag{11}$$

The first term is the trained network; the second is intractable, since $p_t(y|u_t,\tau) = \mathbb{E}_{u_0|u_t,\tau}[p(y|u_0)]$ requires marginalising the denoising posterior $p(u_0|u_t,\tau)$, and DPS (Chung et al., 2023) evaluates the likelihood at its Tweedie point estimate $\hat{u}_0(u_t,\tau) = \mathbb{E}[u_0|u_t,\tau]$, available in closed form as,

$$\hat{u}_0(u_t,\tau) = \tfrac{1}{\sqrt{\bar{\alpha}(t)}}\big(u_t + (1-\bar{\alpha}(t))\ \mathsf{V}(u_t,t,\tau)\big), \tag{12}$$

so that, under the Gaussian likelihood given in (1),

$$\nabla_{u_t} \log p_t(y|u_t,\tau)\ \approx\ \nabla_{u_t} \log p\big(y|\hat{u}_0(u_t,\tau)\big)\ =\ -\tfrac{1}{2\sigma_y^2}\,\nabla_{u_t}\big\|y - \mathsf{H}\,\hat{u}_0(u_t,\tau)\big\|_2^2. \tag{13}$$

Substituting (13) into (11) and discretising the resulting reverse SDE gives our sampler. Standard DPS replaces the factor $1/(2\sigma_y^2)$ in (13) by a tuned step size $\zeta_i = \zeta'/\|y - \mathsf{H}\hat{u}_0\|_2$, but we retain (13) as derived, so that prior and likelihood keep their relative scale. The cost is that the approximation of the denoising posterior as a point estimate is crude at large t, which over-weights the data misfit and can cause divergence. We therefore apply a norm clip, controlled by a constant c, to the likelihood guidance contribution at each sampling step, which is only active at large t. Algorithm 2 provides the inference process, with a detailed version in Algorithm 3 in Appendix. Tables 5 and 9 in Appendix G show this yields better calibrated posteriors than standard DPS, though the approximation in (13) still leaves diffusion posteriors worse calibrated than autoencoders, shown further in experiments.

3 IMPLEMENTATION

Text generation. Text descriptions are generated synthetically by prompting LLMs with information from the corresponding field. They are free-form natural language, rather than a fixed template, so as to imitate the way a person would describe a QoI. The same QoI could be described in different words and sentence structures. Descriptions in the test set are held out from fine-tuning and training. More details about text generation and prompts used in each experiment are in Appendix G.

Network Architectures. The autoencoder uses modified Fourier neural operators (FNO) (Li et al., 2021) in steady-state heat and Helmholtz experiments and a convolutional network in the UK weather experiment; the score network V is a U-Net (Ronneberger et al., 2015) throughout. Text embeddings enter the autoencoder by concatenation; in the score network they are concatenated with the sinusoidal embedding of the diffusion time t, as the input to FiLM-style adaptive group normalisation (AdaGN) (Perez et al., 2018; Dhariwal & Nichol, 2021). This work focuses on how natural language is incorporated into the prior and in the training and inference pipeline, rather than architectural developments. We compare our proposed methodology (**Cond AE**, **Cond Diffusion**) against Gaussian processes (**GP**) (Rasmussen, 2003), and a deterministic supervised model (Fukami et al., 2021), which is a direct map $(y,\tau) \mapsto u$. We also test text-free and text-conditioned variants (**Uncon Supervised**, **Cond Supervised**, more details in Appendix C), and unconditional generative priors trained without text (**Uncon AE**, **Uncon Diffusion**).

Inference-time Sampling. Observations are drawn at different locations for each system. M posterior draws are kept per system. For the autoencoder, inference runs in the latent space with NUTS (Hoffman et al., 2014) using Pyro (Bingham et al., 2019), with a single chain of M_o warmup steps followed by M retained draws. Both warmup and sampling are counted in the reported inference cost. For the diffusion model, the M draws come from M independent integrations of the guided reverse SDE, each of L steps. More details about training and inference are in Appendix G.

4 NUMERICS

4.1 STEADY-STATE HEAT EQUATION

Setup: The first experiment is on the steady-state heat equation on a unit square, with setup in Appendix G.1.1. For each system, we randomly select two of the four sides and assign zero tem-

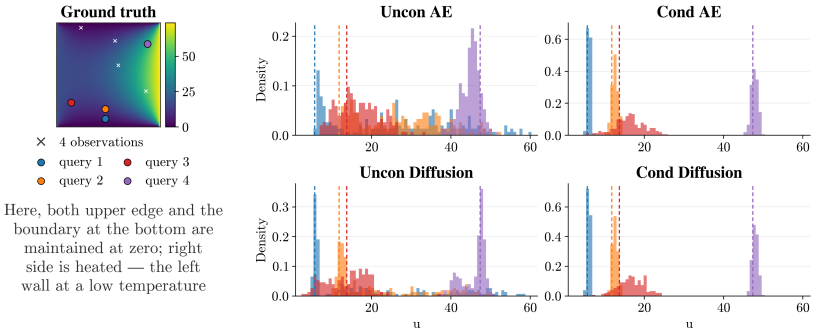

Figure 1: **Posterior marginals of steady-state heat equation at four unobserved query locations for $\mathbf{n_{obs} = 4}$.** Dashed vertical lines denote the ground truth at four query points.

Table 1: **Inference results on the steady-state heat equation at $\mathbf{n_{obs} = \{4, 8\}}$.** We report NMSE of posterior mean (with quartiles in Appendix Table 6), and coverages of the ground truth by the centered 68% and 95% posterior intervals (target values: 68% and 95%). NMSE entries are mean ± 1 std across 200 test fields; coverages report the mean, values in parentheses are the mean absolute difference from the target coverage (smaller is better). Pred. is the mean per-field inference time.

	$n_{\text{obs}} = 4$			$n_{\text{obs}} = 8$			Cost		
Model	NMSE ($\times 10^{-3}$)	$\%C_{68}$	$\%C_{95}$	NMSE ($\times 10^{-3}$)	$\%C_{68}$	$\%C_{95}$	Param(M)	Train(hr)	Pred(s)
GP	178.87 ± 103.15	71.7 (14.8)	93.6 (5.5)	97.03 ± 59.79	64.6 (13.8)	89.7 (7.5)	–	–	0.0004
Uncon Supervised	54.66 ± 203.76	–	–	4.87 ± 12.89	–	–	5.13	1.91	0.0003
Cond Supervised	9.52 ± 17.02	–	–	3.54 ± 6.18	–	–	5.13	0.93	0.0003
Uncon AE	46.45 ± 132.21	61.5 (22.4)	88.7 (12.0)	4.85 ± 15.95	67.6 (17.8)	92.6 (8.5)	9.48	3.63	28.51
Cond AE	6.13 ± 11.70	61.6 (23.4)	88.2 (8.2)	1.40 ± 3.01	62.5 (20.8)	88.1 (7.9)	9.87	2.40	10.25
Uncon Diffusion	37.71 ± 107.23	64.9 (25.2)	89.2 (12.2)	2.45 ± 8.95	59.8 (26.0)	87.0 (14.3)	0.99	2.75	32.15
Cond Diffusion	7.12 ± 17.64	62.3 (25.2)	89.8 (12.0)	1.44 ± 3.54	64.3 (22.7)	90.3 (11.2)	1.09	1.76	32.21

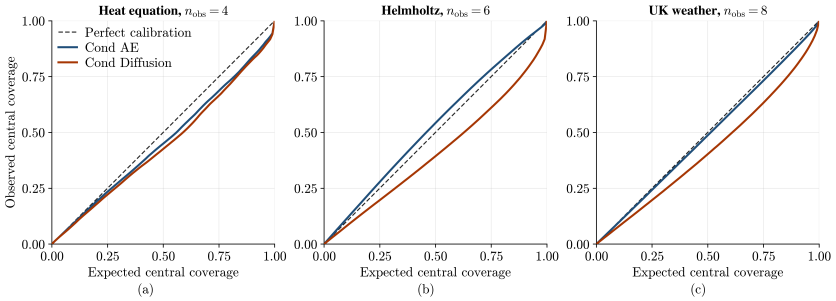

Figure 2: **Performance of posterior calibration.** We plot calibration curves (observed coverage over target central coverage) for three experiments, where the diagonal represents perfect calibration, with more details about calibration curve calculation in Appendix F.

perature; the remaining two are set to a constant drawn from $\text{Unif}[10, 80]$. Accompanying texts describe these boundary temperatures qualitatively, with an example in Figure 1. Training uses $6.4k$ samples. **Results:** After training the autoencoder E, D, and the diffusion score network V, we perform inference on 200 test fields with observation noise $\sigma_y = 1$, and different numbers of observations $n_{\text{obs}} \in \{2, 4, 6, 8, 10\}$. Tables 1 and 5 compare all models in terms of accuracy, coverage and cost, with metrics detailed in Appendix F. Generative priors improve substantially from GP and supervised baselines, and each conditional model improves on its unconditional counterpart. Posterior histograms in Figure 1 also show that unconditional models produce wider posteriors. **Cond AE** and **Cond Diffusion** report comparable NMSE and coverages in Table 1 and Figure 2(a), but the bracketed mean absolute differences from target coverages are larger for **Cond Diffusion** (12.0 against 8.2 for $\%C_{95}$ at $n_{\text{obs}} = 4$), and Appendix Figure 11 shows that per-system coverages of **Cond Diffusion** are less concentrated around target coverages, with its high average coming from

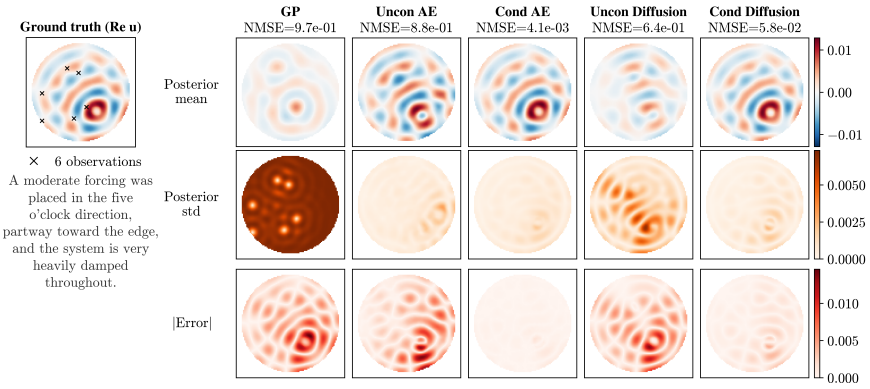

Figure 3: Inference results of one Helmholtz test system (Real part) at $n_{\text{obs}} = 6$.

Table 2: Inference results on 100 Helmholtz test fields, more details in Appendix Tables 9 and 10.

	$n_{\text{obs}} = 6$				$n_{\text{obs}} = 10$				Cost		
Model	NMSE ($\times 10^{-2}$)		$\%C_{68}$	$\%C_{95}$	NMSE ($\times 10^{-2}$)		$\%C_{68}$	$\%C_{95}$	Param(M)	Train(hr)	Pred(s)
GP	86.74 ±	9.20	83.3 (23.0)	94.0 (7.4)	77.12 ±	11.96	82.3 (23.1)	93.6 (7.7)	–	–	0.005
Uncon Supervised	25.93 ±	23.18	–	–	10.17 ±	7.68	–	–	8.22	13.64	0.002
Cond Supervised	19.77 ±	25.27	–	–	10.84 ±	16.48	–	–	8.32	10.91	0.002
Uncon AE	66.61 ±	338.05	45.9 (28.7)	68.6 (28.3)	16.07 ±	30.13	46.1 (26.9)	68.4 (28.5)	19.71	19.37	145.25
Cond AE	7.15 ±	17.18	71.4 (13.6)	94.5 (5.5)	3.16 ±	7.25	69.7 (11.8)	93.8 (5.4)	20.66	15.91	43.90
Uncon Diffusion	22.57 ±	24.20	64.4 (18.0)	93.4 (6.7)	5.36 ±	9.24	67.6 (19.1)	92.0 (8.4)	6.02	17.33	231.47
Cond Diffusion	7.15 ±	12.63	54.5 (23.5)	83.0 (15.3)	3.45 ±	7.54	54.7 (20.7)	83.6 (14.1)	6.31	13.17	232.18

60% of systems reaching $\%C_{95} = 100\%$, where the posterior is over-dispersed. This shows that DPS introduces bias into the posterior approximation due to the estimation in (13). More evaluations such as ablations on different n_{obs} (Figures 12,13) and CRPS (Table 5) are in Appendix G.1.3.

4.2 Helmholtz Equation

Setup: This experiment is a damped Helmholtz equation on a circular disc, with uniform forcing over a small circular region. Each system is a complex field, with two channels for the real and imaginary parts. Accompanying text describes the damping intensity, the amplitude and location of forcing, with an example in Figure 3. The precision of descriptions varies across systems, with details in Appendix G.2.1. $8k$ training samples are generated. **Results:** Inference is performed over 100 test fields with $\sigma_y = 6 \times 10^{-4}$ and $n_{\text{obs}} = \{4, 6, 8, 10, 12\}$. Tables 2 and 9 again show conditional generative priors ahead of other baselines and unconditional alternatives, with **Cond AE** attaining better posterior coverage and faster inference than **Cond Diffusion**. In Figure 2(b), the calibration curve of **Cond AE** sits closer to the diagonal and **Cond Diffusion** sits below the diagonal, indicating overconfident posteriors. Figure 3 plots the posterior mean, standard deviation, and absolute error of five probabilistic models at $n_{\text{obs}} = 6$, with unconditional models having both a larger error and posterior spread. Additional results are provided in Appendix G.2.3.

4.3 UK Weather Reanalysis Data

Table 3: Inference results on 100 UK weather test fields, more details in Appendix Tables 12 and 13.

	$n_{\text{obs}} = 8$				$n_{\text{obs}} = 12$				Cost		
Model	NMSE ($\times 10^{-2}$)		$\%C_{68}$	$\%C_{95}$	NMSE ($\times 10^{-2}$)		$\%C_{68}$	$\%C_{95}$	Params(M)	Train(hr)	Pred.(s)
GP	12.12 ±	16.06	68.3 (8.5)	91.7 (5.0)	10.16 ±	13.80	69.7 (8.1)	92.1 (4.6)	–	–	0.002
Uncon Supervised	7.46 ±	11.38	–	–	6.22 ±	10.47	–	–	8.23	22.34	0.002
Cond Supervised	6.78 ±	9.95	–	–	5.89 ±	8.81	–	–	8.32	14.89	0.002
Uncon AE	8.18 ±	11.55	67.5 (6.5)	93.6 (3.3)	6.90 ±	10.50	65.8 (5.9)	92.3 (3.6)	16.92	14.08	47.02
Cond AE	6.93 ±	9.81	65.9 (5.6)	92.5 (3.5)	6.12 ±	9.58	63.8 (6.6)	90.9 (4.5)	17.71	15.73	47.08
Uncon Diffusion	7.72 ±	11.25	56.2 (12.4)	86.5 (8.7)	6.32 ±	11.66	56.3 (12.2)	86.5 (8.5)	9.19	44.29	283.38
Cond Diffusion	6.81 ±	10.13	56.0 (12.5)	86.0 (9.1)	6.05 ±	11.02	55.7 (12.6)	85.8 (9.2)	10.17	26.73	283.82

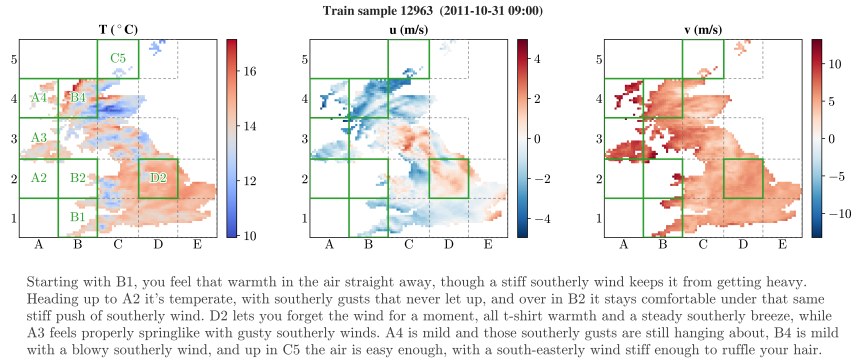

Figure 4: **Example of a 3-channel weather data slice with corresponding text description.** Green boxes mark the regions referenced in the text, and grey dashed boxes mark available land regions.

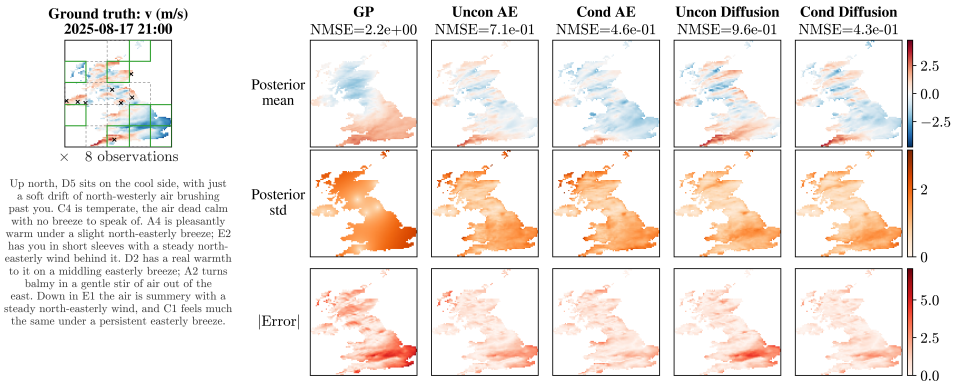

Figure 5: Inference results of one UK weather test system (v) with $n_{\text{obs}} = 8$.

Setup: This experiment uses CERRA reanalysis data (Ridal et al., 2024), taking temperature and two wind velocity components over UK land as three channels $\{T, u, v\}$. A 5×5 grid is laid over the domain and each description covers 8 randomly chosen regions, reporting the temperature and wind of each with an example in Figure 4. Training uses 27,396 fields from 2000-2024. **Results.** Inference is performed over 100 test fields, with $\sigma_y = \{0.5, 0.3, 0.3\}$ and $n_{\text{obs}} = \{4, 6, 8, 10, 12, 14\}$, with results in Tables 3 and 12. Deterministic supervised baselines perform better here, but provide no UQ and are tied to the observation process they were trained on. With a different observation operator at inference in Appendix Table 14, generative models overtake them. Diffusion models provide better NMSE than autoencoders, showing its expressiveness on more irregular fields, but **Cond Diffusion** gives worse calibrated posterior than **Cond AE** in Figure 2(c). Figure 5 plots results for one v field. The gain from text is smaller here than in other experiments, since the text carries less information about the field, but conditional models still reach lower NMSE and show lower error inside the regions that the text describes. More evaluations are provided in Appendix G.3.3.

5 Conclusion

We propose a framework for learning text-conditional generative priors over physical fields, allowing qualitative natural-language descriptions to be combined with sparse quantitative measurements for posterior inference and uncertainty quantification. We develop a complete pipeline from fine-tuning a text encoder, to training generative priors based on autoencoders and diffusion models, and to posterior inference. Comparing the two, autoencoders achieve better-calibrated posteriors while diffusion models offer greater expressiveness on more irregular fields. Across our experiments, incorporating text into the prior consistently improves reconstruction accuracy over unconditional and non-data-driven priors.

AI USE STATEMENT

In this work, we used generative AI tools to help in the retrieval and discovery of related works, and to help write code to implement methods. It was used to a limited extent to polish specific sections of the writing. AI was used in providing some limited ingredients, subsequently used and arranged into the theory, for proving mathematical claims. AI was also used in creating synthetic datasets for experiments. Prompts used to generate synthetic datasets are provided in the Appendix.

REPRODUCIBILITY STATEMENT

Model structures and hyperparameters used in all experiments are provided in Appendix G. Code for reproducing all experiments is provided in the supplementary material.

REFERENCES

O Deniz Akyildiz, Mark Girolami, Andrew M Stuart, and Arnaud Vadeboncoeur. Efficient prior calibration from indirect data. *SIAM Journal on Scientific Computing*, 47(4):C932–C958, 2025.

Mohammad Al-Jarrah, Michele Martino, Marcus Yim, Bamdad Hosseini, and Amirhossein Taghvaei. Conditional sampling via wasserstein autoencoders and triangular transport. *arXiv preprint arXiv:2604.02644*, 2026.

Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta numerica*, 28:1–174, 2019.

Ross N Bannister. A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(703):607–633, 2017.

Ricardo Baptista, Bamdad Hosseini, Nikola B Kovachki, and Youssef M Marzouk. Conditional sampling with monotone gans: from generative models to likelihood-free inference. *SIAM/ASA Journal on Uncertainty Quantification*, 12(3):868–900, 2024.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis M Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6, 2019.

Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure theory*, volume 1. Springer, 2007.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International conference on machine learning*, pp. 537–546, 2017.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and Omer Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Tan Bui-Thanh and Mark Girolami. Solving large-scale pde-constrained bayesian inverse problems with riemann manifold hamiltonian monte carlo. *Inverse Problems*, 30(11):114014, 2014.

Alexander Capstick, Rahul G Krishnan, and Payam Barnaghi. Autoelicit: Using large language models for expert prior elicitation in predictive modelling. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Gabriel Cardoso, Yazid Janati el idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided denoising diffusion models for bayesian linear inverse problems. In *International Conference on Learning Representations (ICLR)*, 2024.

Diego Caviedes-Nozal, Nicolai AB Riis, Franz M Heuchel, Jonas Brunskog, Peter Gerstoft, and Efren Fernandez-Grande. Gaussian processes for sound field reconstruction. *The Journal of the Acoustical Society of America*, 149(2):1107–1119, 2021.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *International Conference on Learning Representations (ICLR)*, 2024a.

Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. In *Proceedings of the 41nd International Conference on Machine Learning*, 2024b.

Simon L Cotter, Masoumeh Dashti, and Andrew M Stuart. Approximation of bayesian inverse problems for pdes. *SIAM journal on numerical analysis*, 48(1):322–345, 2010.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

N Benjamin Erichson, Lionel Mathelin, Zhewei Yao, Steven L Brunton, Michael W Mahoney, and J Nathan Kutz. Shallow neural networks for fluid flow reconstruction with limited sensors. *Proceedings. Mathematical, physical, and engineering sciences*, 476(2238):20200097, 2020.

Eky Febrianto, Liam Butler, Mark Girolami, and Fehmi Cirak. Digital twinning of self-sensing structures using the statistical finite element method. *Data-Centric Engineering*, 3:e31, 2022.

Berthy Feng and Katherine Bouman. Variational bayesian imaging with an efficient surrogate score-based prior. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10486–10497. IEEE, 2023.

Kai Fukami, Romit Maulik, Nesar Ramachandra, Koji Fukagata, and Kunihiko Taira. Global field reconstruction from sparse sensors with voronoi tessellation-assisted deep learning. *Nature Machine Intelligence*, 3(11):945–951, 2021.

Mark Girolami, Eky Febrianto, Ge Yin, and Fehmi Cirak. The statistical finite element method (statfem) for coherent synthesis of observation data and model predictions. *Computer Methods in Applied Mechanics and Engineering*, 375:113533, 2021.

Alex Glyn-Davies, Arnaud Vadeboncoeur, O Deniz Akyildiz, Ieva Kazlauskaite, and Mark Girolami. A primer on variational inference for physics-informed deep generative modelling. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 383(2299):20240324, 2025.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Bamdad Hosseini and Ziqi Huang. Error analysis of bayesian inverse problems with generative priors. *arXiv preprint arXiv:2601.17374*, 2026.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation. *Advances in Neural Information Processing Systems*, 37: 130291–130323, 2024.

Yongchao Huang. Llm-prior: A framework for knowledge-driven prior elicitation and aggregation. *arXiv preprint arXiv:2508.03766*, 2025.

Christian Jacobsen, Yilin Zhuang, and Karthik Duraisamy. Cocogen: Physically consistent and conditioned score-based generative models for forward and inverse problems. *SIAM Journal on Scientific Computing*, 47(2):C399–C425, 2025.

Xinchao Jiang, Xin Wang, Ziming Wen, and Hu Wang. Resolution-independent generative models based on operator learning for physics-constrained bayesian inverse problems. *Computer Methods in Applied Mechanics and Engineering*, 420:116690, 2024.

Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*, volume 1. Cambridge university press Cambridge, 2003.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Jeongsol Kim, Geon Yeong Park, Hyungjin Chung, and Jong Chul Ye. Regularization by texts for latent diffusion inverse solvers. In *International Conference on Learning Representations (ICLR)*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Yuxuan Liu, Zecheng Zhang, and Hayden Schaeffer. Prose: Predicting multiple operators and symbolic expressions using multimodal transformers. *Neural Networks*, 180:106707, 2024b.

Peter Manshausen, Yair Cohen, Peter Harrington, Jaideep Pathak, Mike Pritchard, Piyush Garg, Morteza Mardani, Karthik Kashinath, Simon Byrne, and Noah Brenowitz. Generative data assimilation of sparse weather station observations at kilometer scales. *Journal of Advances in Modeling Earth Systems*, 17(10):e2024MS004505, 2025.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Xuhui Meng, Liu Yang, Zhiping Mao, José del Águila Ferrandis, and George Em Karniadakis. Learning functional priors and posteriors from data and physics. *Journal of Computational Physics*, 457:111073, 2022.

Petrus Mikkola, Osvaldo A Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, et al. Prior knowledge elicitation: The past, present, and future. *Bayesian Analysis*, 19(4):1129–1161, 2024.

Elisa Negrini, Yuxuan Liu, Liu Yang, Stanley J Osher, and Hayden Schaeffer. A multimodal pde foundation model for prediction and scientific text descriptions. *arXiv preprint arXiv:2502.06026*, 2025.

Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*, volume 545. Wiley Chichester, 2006.

Dhruv V Patel and Assad A Oberai. Gan-based priors for quantifying uncertainty in supervised learning. *SIAM/ASA Journal on Uncertainty Quantification*, 9(3):1314–1343, 2021.

Dhruv V Patel, Deep Ray, and Assad A Oberai. Solution of physics-based bayesian inverse problems with deep generative priors. *Computer Methods in Applied Mechanics and Engineering*, 400: 115428, 2022.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Martin Ridal, Eric Bazile, Patrick Le Moigne, Roger Randriamampianina, Semjon Schimanke, Ulf Andrae, Lars Berggren, Pierre Brousseau, Per Dahlgren, Lisette Edvinsson, et al. Cerra, the copernicus european regional reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, 150(763):3385–3411, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

François Rozet and Gilles Louppe. Score-based data assimilation. *Advances in Neural Information Processing Systems*, 36:40521–40541, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Taiga Saito, Yu Otake, Daijiro Mizutani, and Sopheakpolin Mom. What do language priors contribute to darcy-flow inversion? a mechanistic audit. *arXiv preprint arXiv:2606.24967*, 2026.

Javier E Santos, Zachary R Fox, Arvind Mohan, Daniel O'Malley, Hari Viswanathan, and Nicholas Lubbers. Development of the senseiver for efficient field reconstruction from sparse observations. *Nature Machine Intelligence*, 5(11):1317–1325, 2023.

David Selby, Yuichiro Iwashita, Kai Spriestersbach, Mohammad Saad, Dennis Bappert, Archana Warrier, Sumantrak Mukherjee, Koichi Kise, and Sebastian Vollmer. Had enough of experts? quantitative knowledge retrieval from large language models. *Stat*, 14(2):e70054, 2025.

Siming Shan, Min Zhu, Youzuo Lin, and Lu Lu. Regularization by denoising diffusion models for solving inverse pde problems with application to full waveform inversion. *Communications Physics*, 2026.

Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972, 2023.

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José M Hernández-Lobato, Richard E Turner, and Emile Mathieu. On conditional diffusion models for pde simulations. *Advances in Neural Information Processing Systems*, 37:23246–23300, 2024.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations (ICLR)*, 2022.

Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018.

Arnaud Vadeboncoeur, Gregory Duthé, Mark Girolami, and Eleni Chatzi. Geometric autoencoder priors for bayesian inversion: Learn first observe later. In *International Conference on Learning Representations (ICLR)*, 2026a.

Arnaud Vadeboncoeur, Mark Girolami, and Andrew M Stuart. Efficient deconvolution in populational inverse problems. *International Journal for Numerical Methods in Engineering*, 127(9): e70335, 2026b.

Titouan Vayer and Rémi Gribonval. Controlling wasserstein distances by kernel norms with application to compressive statistical learning. *Journal of Machine Learning Research*, 24(149):1–51, 2023.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

Jakob Zech and Youssef Marzouk. Sparse approximation of triangular transports, part i: The finite-dimensional case. *Constructive Approximation*, 55(3):919–986, 2022a.

Jakob Zech and Youssef Marzouk. Sparse approximation of triangular transports, part ii: The infinite-dimensional case. *Constructive Approximation*, 55(3):987–1036, 2022b.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations (ICLR)*, 2023.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary Ross, et al. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations (ICLR)*, 2025.

Anthony Zhou, Zijie Li, Michael Schneier, John R Buchanan Jr, and Amir Barati Farimani. Text2pde: Latent diffusion models for accessible physics simulation. In *International Conference on Learning Representations (ICLR)*, 2025.

A NOTATION

Table 4: Symbols and concepts.

Symbol	Description	Example
$u \in \mathcal{U}$	Quantity of Interest (QoI)	Heat distribution over the physical domain
$\theta \in \Theta$	Human-perceivable quantity related to u	The left boundary condition
$s \in \mathcal{S}$	Qualitative data: description of θ	`"the left boundary is hot"`
$\tau \in \mathcal{T}$	Euclidean description of s	Text encoding of `"the left boundary is hot"`
$y \in \mathbb{R}^{d_y}$	Quantitative data: incomplete/indirect measurements of u	Sparse sensor measurements of heat on domain

In this paper, spaces are written in calligraphic type ($\mathcal{U}, \mathcal{T}, \mathcal{Z}$) and maps between them in sans serif ($\mathsf{D}, \mathsf{E}, \mathsf{G}^{\phi}, \mathsf{V}, \mathsf{H}$). Probability measures are written as $\mathbb{P}, \mathbb{Q}$, with a subscript naming the variable they act on; $\mathbb{P}_{u,\tau}$ is a measure on $(\mathcal{U} \times \mathcal{T}, \mathcal{B}(\mathcal{U} \times \mathcal{T}))$, where $\mathcal{B}(\cdot)$ is the Borel sigma-algebra, and $\mathbb{P}_{u|\tau}$ is the corresponding conditional. p is reserved for densities with respect to Lebesgue measure, as in the score $\nabla_u \log p_t(u|\tau)$.

For a measurable function $g : \mathcal{Z} \to \mathcal{U}$, the pushforward $g_{\#}\mathbb{P}_z$ is defined by $g_{\#}\mathbb{P}_z(A) = \mathbb{P}_z(g^{-1}(A))$ for every measurable set A. A sample $u \sim g_{\#}\mathbb{P}_z$ is obtained by drawing $z \sim \mathbb{P}_z$ and returning $u = g(z)$. A bullet marks an argument left free, so $\mathsf{D}(\bullet, \tau)_{\#}\mathbb{Q}_z$ is the pushforward of $\mathbb{Q}_z$ through the decoder with the text embedding held fixed at τ. Table 4 summarises some symbols and concepts.

B PROOF

B.1 THEOREM 2.1

Proof. For $\mathbb{P}_{u,\tau} - a.e\ u, \tau$, if the first reconstruction error term in (5) is 0, we have

$$\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}} = (\mathsf{Id}_{\mathcal{U}} \times \mathsf{Id}_{\mathcal{T}}).$$

Given the joint distribution $\mathbb{P}_{u,\tau}$, we use notation "$\otimes$", such that for any measurable set A,

$$\mathbb{P}_{u,\tau}(A) = (\mathbb{P}_\tau \otimes \mathbb{P}_{u|\tau})(A) = \int_{\mathcal{T}} \mathbb{P}_{u|\tau}(\{u : (u,\tau) \in A\}) \mathrm{d}\mathbb{P}_\tau(\tau); \tag{14}$$

where, for $\mathbb{P}_\tau$ being the τ-marginal of $\mathbb{P}_{u,\tau}$, $\mathbb{P}_{u|\tau}$ is the $\mathbb{P}_\tau - a.e$ unique system of conditional measure on the sample space of variable u, and is $\mathbb{P}_\tau$-measurable. Now, invoking existence and uniqueness of conditional measures, if the second term in (5) is also 0,

$$\begin{aligned}
&\widetilde{\mathsf{E}}_{\#}(\mathbb{P}_{u,\tau}) = \mathbb{Q}_z \times \mathbb{P}_\tau \\
\implies &(\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}})_{\#}(\mathbb{P}_{u,\tau}) = \widetilde{\mathsf{D}}_{\#}(\mathbb{Q}_z \times \mathbb{P}_\tau) \\
\implies &\mathbb{P}_{u,\tau} = \widetilde{\mathsf{D}}_{\#}(\mathbb{Q}_z \times \mathbb{P}_\tau) \\
\implies &\mathbb{P}_{u,\tau} = (\mathbb{P}_\tau \otimes \mathsf{D}(\bullet, \tau)_{\#}\mathbb{Q}_z) \\
\implies &\mathbb{P}_{u|\tau} = \mathsf{D}(\bullet, \tau)_{\#}\mathbb{Q}_z, \quad \mathbb{P}_\tau - a.e.
\end{aligned}$$

Going from line three to four, involves

$$\begin{aligned}
\widetilde{\mathsf{D}}_{\#}(\mathbb{Q}_z \times \mathbb{P}_\tau)(A) &= \int_{\mathcal{U}\times\mathcal{T}} \mathbb{1}_A(u,\tau)\ \mathrm{d}(\widetilde{\mathsf{D}}_{\#}(\mathbb{Q}_z \times \mathbb{P}_\tau))(u,\tau) \\
&= \int_{\mathcal{Z}\times\mathcal{T}} \mathbb{1}_A(\mathsf{D}(z,\tau),\tau)\ \mathrm{d}(\mathbb{Q}_z \times \mathbb{P}_\tau)(z,\tau) \\
&= \int_{\mathcal{T}} \int_{\mathcal{Z}} \mathbb{1}_A(\mathsf{D}(z,\tau),\tau)\ \mathrm{d}\mathbb{Q}_z(z)\, \mathrm{d}\mathbb{P}_\tau(\tau) \\
&= \int_{\mathcal{T}} \int_{\mathcal{U}} \mathbb{1}_A(u,\tau)\ \mathrm{d}(\mathsf{D}(\bullet,\tau)_{\#}\mathbb{Q}_z)(u)\, \mathrm{d}\mathbb{P}_\tau(\tau) \\
&= \int_{\mathcal{T}} (\mathsf{D}(\bullet,\tau)_{\#}\mathbb{Q}_z)(\{u : (u,\tau) \in A\})\, \mathrm{d}\mathbb{P}_\tau(\tau) \\
&= (\mathbb{P}_\tau \otimes \mathsf{D}(\bullet,\tau)_{\#}\mathbb{Q}_z)(A), \quad \text{for all measurable } A;
\end{aligned}$$

pushing $\widetilde{\mathsf{D}}$ into the integral, and D out of it, invokes Theorem 3.6.1 in Bogachev & Ruas (2007), separation of integrals and exchange of order follows from Tonelli's theorem. □

B.2 THEOREM 2.2

Proof. By squaring the triangle inequality

$$\mathsf{W}_2^2(\mathbb{P}_{u,\tau}, \widetilde{\mathsf{D}}_{\#}(\mathbb{Q}_z \times \mathbb{P}_\tau)) \leq 2\mathsf{W}_2^2(\mathbb{P}_{u,\tau}, (\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}})_{\#}\mathbb{P}_{u,\tau}) + 2\mathsf{W}_2^2((\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}})_{\#}\mathbb{P}_{u,\tau}, \widetilde{\mathsf{D}}_{\#}(\mathbb{Q}_z \times \mathbb{P}_\tau)).$$

We obtain a bound on the first term by selecting $\gamma \in \Pi(\mathbb{P}_{u,\tau}, (\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}})_{\#}\mathbb{P}_{u,\tau})$ to be of the form $\gamma = (\mathsf{Id} \times (\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}}))_{\#}\mathbb{P}_{u,\tau})$, which is a valid coupling, thus

$$\begin{aligned}
\mathsf{W}_2^2(\mathbb{P}_{u,\tau}, (\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}})_{\#}\mathbb{P}_{u,\tau}) &= \inf_{\gamma \in \Pi} \int_{(\mathcal{U}\times\mathcal{T})^2} \|(u,\tau) - (u',\tau')\|^2 \mathrm{d}\gamma \\
&\leq \int_{(\mathcal{U}\times\mathcal{T})^2} \|(u,\tau) - (u',\tau')\|^2 \mathrm{d}((\mathsf{Id}. \times (\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}}))_{\#}\mathbb{P}_{u,\tau}) \\
&= \mathbb{E}_{u,\tau\sim\mathbb{P}_{u,\tau}} \|(u,\tau) - \widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}}(u,\tau)\|^2 \\
&= \mathbb{E}_{u,\tau\sim\mathbb{P}_{u,\tau}} [\|u - \mathsf{D} \circ \widetilde{\mathsf{E}}(u,\tau)\|^2 + \|\tau - \tau\|^2] \\
&= \mathbb{E}_{u,\tau\sim\mathbb{P}_{u,\tau}} \|u - \mathsf{D} \circ \widetilde{\mathsf{E}}(u,\tau)\|^2.
\end{aligned}$$

Bounding the second term involves selecting a coupling $\gamma_\epsilon \in \Pi(\widetilde{\mathsf{E}}_{\#}\mathbb{P}_{u,\tau}, \mathbb{Q}_z \times \mathbb{P}_\tau)$ such that for a chosen $\epsilon > 0$

$$\int_{(\mathcal{Z}\times\mathcal{T})^2} \|(z,\tau) - (z',\tau')\|^2 \mathrm{d}\gamma_\epsilon \leq \mathsf{W}_2^2(\widetilde{\mathsf{E}}_{\#}\mathbb{P}_{u,\tau}, \mathbb{Q}_z \times \mathbb{P}_\tau) + \epsilon.$$

Then, $(\widetilde{\mathsf{D}} \times \widetilde{\mathsf{D}})_{\#}\gamma_\epsilon$ is an admissible coupling such that

$$\begin{aligned}
\mathsf{W}_2^2((\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}})_{\#}\mathbb{P}_{u,\tau}, \widetilde{\mathsf{D}}_{\#}(\mathbb{Q}_z \times \mathbb{P}_\tau)) &\leq \int_{(\mathcal{U}\times\mathcal{T})^2} \|(u,\tau) - (u',\tau')\|^2 \mathrm{d}((\widetilde{\mathsf{D}} \times \widetilde{\mathsf{D}})_{\#}\gamma_\epsilon) \\
&= \int_{(\mathcal{Z}\times\mathcal{T})^2} \|\widetilde{\mathsf{D}}(z,\tau) - \widetilde{\mathsf{D}}(z',\tau')\|^2 \mathrm{d}\gamma_\epsilon \\
&= \int_{(\mathcal{Z}\times\mathcal{T})^2} (\|\mathsf{D}(z,\tau) - \mathsf{D}(z',\tau')\|^2 + \|\tau - \tau'\|^2) \mathrm{d}\gamma_\epsilon \\
&\leq \int_{(\mathcal{Z}\times\mathcal{T})^2} (\|\mathsf{D}\|_{\mathrm{Lip}}^2 \|(z,\tau) - (z',\tau')\|^2 + \|\tau - \tau'\|^2 \\
&\qquad + \|z - z'\|^2) \mathrm{d}\gamma_\epsilon \\
&= \int_{(\mathcal{Z}\times\mathcal{T})^2} (\|\mathsf{D}\|_{\mathrm{Lip}}^2 + 1) \|(z,\tau) - (z',\tau')\|^2 \mathrm{d}\gamma_\epsilon \\
&\leq (\|\mathsf{D}\|_{\mathrm{Lip}}^2 + 1)(\mathsf{W}_2^2(\widetilde{\mathsf{E}}_{\#}\mathbb{P}_{u,\tau}, \mathbb{Q}_z \times \mathbb{P}_\tau) + \epsilon).
\end{aligned}$$

As this is true for any $\epsilon > 0$,

$$\mathsf{W}_2^2((\widetilde{\mathsf{D}} \circ \widetilde{\mathsf{E}})_{\#}\mathbb{P}_{u,\tau}, \widetilde{\mathsf{D}}_{\#}(\mathbb{Q}_z \times \mathbb{P}_\tau)) \leq (\|\mathsf{D}\|_{\mathrm{Lip}}^2 + 1)\mathsf{W}_2^2(\widetilde{\mathsf{E}}_{\#}\mathbb{P}_{u,\tau}, \mathbb{Q}_z \times \mathbb{P}_\tau).$$

Substituting the two bounds into the squared triangle inequality gives Theorem 2.2. □

B.3 COROLLARY 2.3

Proof. This follows directly from Theorem 2.2 and the bound, for distributions on compact supports,

$$\mathsf{W}_2(\mathbb{P}, \mathbb{P}') \leq C\mathsf{MMD}_{\mathrm{RBF}}^{\alpha}(\mathbb{P}, \mathbb{P}')$$

with $C > 0, \alpha \in (0, 1]$ (Vayer & Gribonval, 2023). Note that the compact-support assumption does not hold for $\mathbb{Q}_z = \mathcal{N}(0, \mathrm{I})$ exactly. This corollary is included to connect the MMD term in (5) to the Wasserstein bound in (6) qualitatively. □

C SUPERVISED METHOD

The supervised baseline follows Fukami et al. (2021), a direct map $(y, \tau) \mapsto u$ that can take any number of sensors at arbitrary locations as input. The observations are first expanded to a dense field following Voronoi tessellation, where every pixel takes the value of its nearest sensor, with this field concatenated with a binary mask marking sensor locations. Figure 6 shows ground truth fields and the corresponding Voronoi inputs at $n_{\text{obs}} = \{4, 8\}$. We use U-Net as the base architecture in our paper, and for **Cond Supervised** the text embedding is added through FiLM layers. However, the supervised map is tied to the kind of observation process it was trained under. A different range of sensor numbers or locations, different forward operator, or a different noise level, requires re-training. Generative models keep prior and likelihood separate, so likelihood and the corresponding observation process can be changed after training.

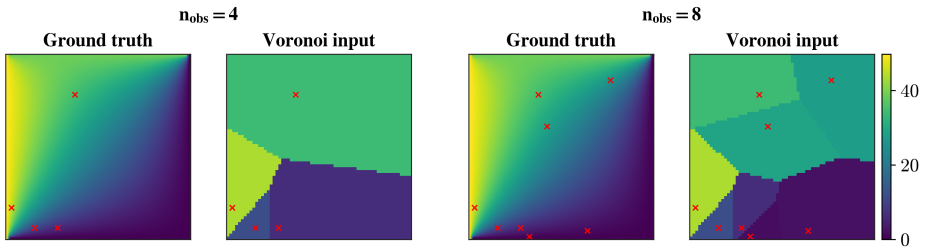

Figure 6: Ground truth fields and Voronoi inputs to the supervised model at $n_{\text{obs}} = \{4, 8\}$.

D MMD IMPLEMENTATION IN AUTOENCODERS

In autoencoders, we use maximum mean discrepancy (MMD) for $\mathsf{d}(\bullet, \bullet)$ in (5) to compute the distance in the joint probability space $\mathcal{Z} \times \mathcal{T}$. In a batch of size B, we have joint samples $\{(z_i, \tau_i)\}_{i=1}^{B} \sim \mathsf{E}_{\#}(\mathbb{P}_{u,\tau})$ that pair each encoder output $z_i = \mathsf{E}(u_i, \tau_i)$ with its own text embedding. We also have reference samples $\{(z_i^{\text{ref}}, \tau_i^{\text{ref}})\}_{i=1}^{B}$ with $z_i^{\text{ref}} \sim \mathcal{N}(0, \mathrm{I}) = \mathbb{Q}_z$, and $\tau_i^{\text{ref}} \sim \mathbb{P}_\tau$ is a text embedding drawn independently from the training pool. We evaluate the MMD between $\{(z_i, \tau_i)\}_{i=1}^{B}$ and $\{(z_i^{\text{ref}}, \tau_i^{\text{ref}})\}_{i=1}^{B}$, with the kernel being a mixture of five RBF kernels with bandwidths $2^m h$, $m = -2, \ldots, 2$, where h is the mean pairwise squared distance of all samples in the batch.

In this setup, z and τ have very different dimensions: $d_\tau = 384$ for the text embedding, and d_z ranges from 16 to 64 across our experiments, so the normalised text block would account for 86–96% of every pairwise distance. Therefore, we downscale normalised τ by a factor approximately equal to $\sqrt{d_\tau / d_z}$, which equalises the two contributions. Figure 7 plots latent histograms across all dimensions over the test set in our three experiments, showing that the marginals of z align closely with the ideal $\mathbb{Q}_z = \mathcal{N}(0, \mathrm{I})$.

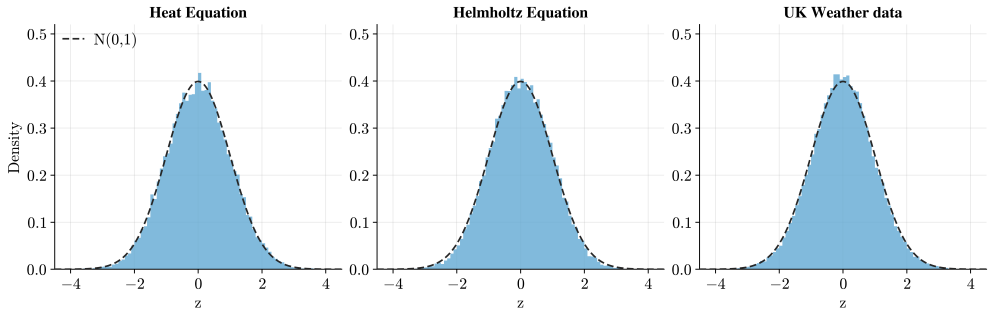

Figure 7: Latent histograms across all dimensions over the test set in three experiments, compared with the ideal $\mathbb{Q}_z = \mathcal{N}(0, \mathrm{I})$.

E MODIFIED DPS ALGORITHM

In Algorithm 3 we provide the modified DPS algorithm, which is an extension of Algorithm 2.

Algorithm 3 Posterior sampling — diffusion (detailed version of Algorithm 2)

Require: trained V, G^{ϕ}; text s, observation y, operator H, noise σ_y; grid $T = t_L > \cdots > t_0 = \varepsilon$, guidance cap c, M posterior draws; write u_i for u_{t_i}

$\tau \leftarrow \mathsf{G}^{\phi}(s)$
for $m = 1$ **to** M **do**
  $u_L \sim \mathcal{N}(0, \mathrm{I})$
  for $i = L$ **to** 1 **do**
    $\beta \leftarrow \beta(t_i)$, $\bar{\alpha} \leftarrow \bar{\alpha}(t_i)$, $\Delta t \leftarrow t_i - t_{i-1}$
    $\hat{v} \leftarrow \mathsf{V}(u_i, t_i, \tau)$ ▷ conditional prior score
    $\hat{u}_0 \leftarrow \big(u_i + (1-\bar{\alpha})\hat{v}\big)/\sqrt{\bar{\alpha}}$ ▷ Tweedie estimate, eq. (12)
    $g \leftarrow -\frac{1}{2\sigma_y^2}\nabla_{u_i}\|y - \mathsf{H}\hat{u}_0\|_2^2$ ▷ likelihood score, eq. (13)
    $\delta \leftarrow \beta\,\Delta t\, g; \quad \delta \leftarrow \delta \cdot \min(1, c/\|\delta\|)$ ▷ guidance, norm capped at c
    $\Delta\bar{w} \sim \mathcal{N}(0, \mathrm{I})$
    $u_{i-1} \leftarrow u_i + \beta\big(\frac{1}{2}u_i + \hat{v}\big)\Delta t + \delta + \sqrt{\beta\Delta t}\,\Delta\bar{w}$ ▷ Euler–Maruyama on eq. (9)
  end for
  $u^{(m)} \leftarrow \big(u_0 + (1-\bar{\alpha}_0)\hat{v}_0\big)/\sqrt{\bar{\alpha}_0}$ ▷ $\hat{v}_0 \leftarrow \mathsf{V}(u_0, t_0, \tau)$, $\bar{\alpha}_0 = \bar{\alpha}(t_0)$
end for
return $\{u^{(m)}\}_{m=1}^{M}$

F PERFORMANCE METRICS

This section defines the performance metrics reported in our experiments.

NMSE. With $u^{\dagger}$ the ground truth field and $\bar{u}$ the posterior mean over M draws, NMSE is computed as $\|u^{\dagger} - \bar{u}\|_2^2/\|u^{\dagger}\|_2^2$, so that it is comparable across systems.

Coverage. For each system, we obtain pixel-wise M posterior samples, and report the percentage of pixels whose ground truth falls inside the central 68%(16th $-$ 84th percentiles) and 95%(2.5th $-$ 97.5th percentiles) posterior intervals, denoting as $\%C_{68}$ and $\%C_{95}$. For a calibrated posterior, the target values should be 68% and 95%. Coverage above the target value indicates an over-dispersed, under-confident posterior; below it, an over-confident one. In the tables, we also report in parentheses the average absolute difference of the per-system coverage from its target values, which measures how far test fields are from being calibrated, which should be 0 when every field is exactly calibrated, and smaller is better.

Calibration Curve. The calibration curves in Figure 2 illustrate posterior calibration at each coverage level. For every pixel in every test field, we record the percentile rank of the ground truth lying within that pixel's posterior ensemble, i.e. the fraction of M posterior samples at that pixel that fall below the ground truth value, a number in $[0, 1]$. At each level q on the horizontal axis, we plot the fraction of pixels whose ground truth falls inside the central q interval of its posterior – equivalently, whose rank lies in $[\frac{1-q}{2}, \frac{1+q}{2}]$ – pooled over the test fields. The diagonal represents perfect calibration, and a curve below the diagonal means the posterior is too narrow and overconfident, while a curve above the diagonal means posterior is too wide and underconfident. Coverages $\%C_{68}$ and $\%C_{95}$ reported in the tables are the values of this curve at $q = 0.68$ and $q = 0.95$.

Continuous ranked probability score (CRPS). CRPS is a proper scoring rule that measures how well a full predictive distribution matches the ground truth, by comparing the empirical cumulative distribution function (CDF) of the predicted values with the step-function CDF of the true value. CRPS is a generalisation of the mean absolute error (MAE) to probabilistic forecasts. Following

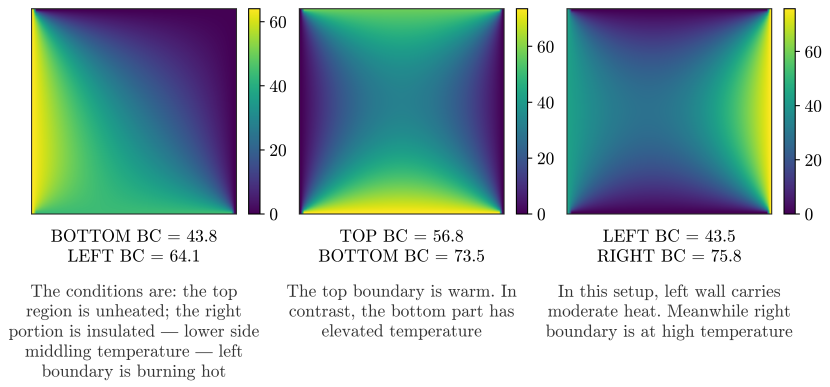

Figure 8: Examples of heat fields with corresponding BC values and text descriptions.

Gneiting & Raftery (2007), CRPS is estimated from M posterior samples as

$$\mathrm{CRPS} = \frac{1}{M}\sum_{m=1}^{M}\left\|\tilde{u}^{(m)} - \tilde{u}^{\dagger}\right\|_1 - \frac{1}{2M(M-1)}\sum_{m,n=1}^{M}\left\|\tilde{u}^{(m)} - \tilde{u}^{(n)}\right\|_1, \tag{15}$$

where $\tilde{u}^{(m)}$ is one normalised posterior sample and $\tilde{u}^{\dagger}$ is the normalised ground truth field. Normalised field is used here with $\tilde{u} = u/\mathrm{RMS}(u^{\dagger})$ as the magnitude of fields varies greatly across systems. Lower CRPS indicates a more accurate and sharper (confident/narrow distributions) posterior ensemble, while higher CRPS indicates ensembles that are either biased or overly diffuse. However, CRPS penalises over-confidence only mildly: as the posterior spread shrinks, the score tends to the MAE of samples, so a posterior whose samples are close to the ground truth attains a low CRPS even if its uncertainty is too narrow to cover the truth.

Train & Predict Time. Reported training time is the total wall-clock time of a run, and each model is trained for enough time for the validation loss to stop improving. Note that in our tables, unconditional models take longer to train because their validation loss converges slower than conditional models. Prediction time is the mean running time to perform inference on a single system, for AE it includes NUTS warmup steps.

G Additional Experimental Details and Results

All experiments are run on a single RTX4090 GPU. This appendix provides more details and evaluations for each of the three experiments including:

1. **Problem Setup:** the governing equations, the solver or data source, further example fields, and details of synthetic text generation;
2. **Experimental Setup:** the choice of $\mathrm{score}_{\mathrm{field}}$ at the fine-tuning stage, network architectures, and hyperparameters used for training and inference;
3. **Additional Results:** more evaluations including samples from the learned priors, per-system coverages, posterior histograms, ablation studies over n_{obs} and how text contributes at different n_{obs}.

G.1 Steady-State Heat Equation

G.1.1 Problem Setup

The PDE used to generate the dataset is a steady-state heat equation on a unit square with Dirichlet boundary conditions,

$$\begin{aligned} \Delta u(x) &= 1, && x \in \Omega = (0,1)^2, \\ u(x) &= g(x), && x \in \partial\Omega, \end{aligned} \tag{16}$$

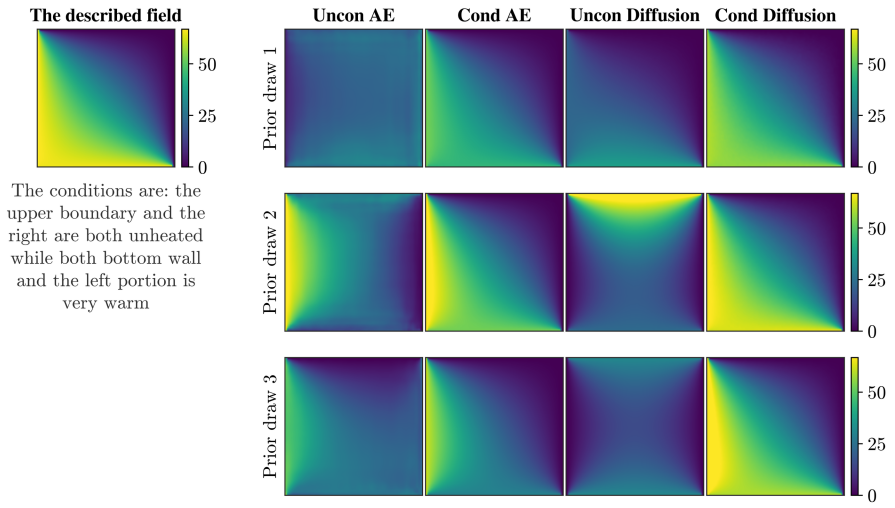

Figure 9: **Prior samples generated given text.** The leftmost panel shows the ground-truth field together with the corresponding text. Each of the four rows shows three prior samples drawn using four models, the text is supplied to conditional models and not provided to unconditional ones.

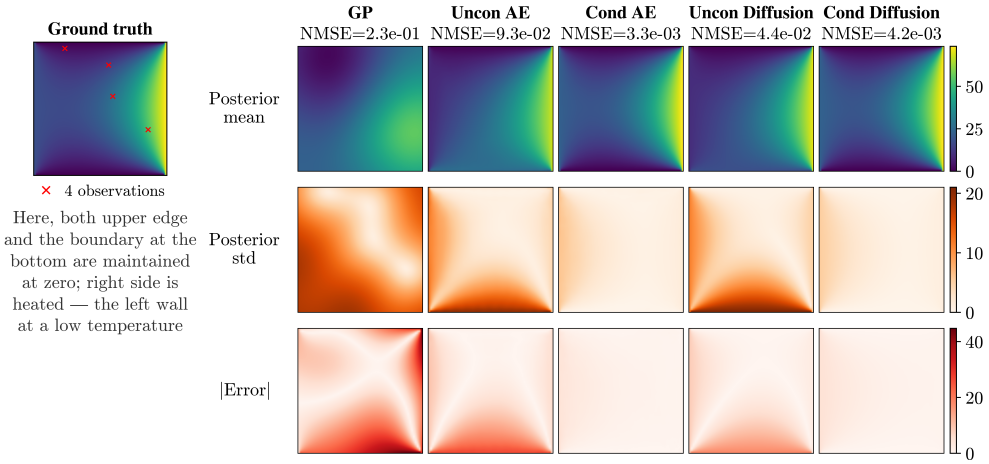

Figure 10: **Posterior reconstructions of one heat equation test system at $n_{obs} = 4$ (same system as in Figure 1).** Each row indicates posterior mean, posterior standard deviation, |error| of the mean respectively. NMSE of posterior mean of each model is reported as well.

where g is piecewise constant on the four sides: a random two of the four sides are held at $g = 0$ and the remaining two at random constant temperatures drawn from Unif$[10, 80]$. We use a finite difference solver on a uniform 64×64 grid.

To generate synthetic texts from given fields, temperatures are split into three overlapping bands: *hot* $[54, 80]$, *warm* $[30, 60]$ and *cold* $[10, 36]$. The bands overlap, and a temperature falling in two of them is described by a random one of the two. In this first and simplest experiment we did not query an API programmatically, instead we prompt *Claude Sonnet 5* directly through a coding agent to write the descriptions. The prompt requires each description to mention boundaries carrying non-zero temperatures, and express it using varied synonyms of *hot*, *warm* and *cold* with different sentence structures, the two unheated boundaries are mentioned in some descriptions but not all. A few examples are provided in Figure 8.

Table 5: **Full inference results on the heat equation test set with n_{obs} varying from 2 to 10.** This table is the extended version of Table 1, and also presents the diffusion inference results obtained via the standard DPS scheme (last two rows). Continuous ranked probability score (CRPS) is also reported, with details in Appendix F.

Model	$n_{obs} = 2$ NMSE ↓ ($\times 10^{-3}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{obs} = 4$ NMSE ↓ ($\times 10^{-3}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{obs} = 6$ NMSE ↓ ($\times 10^{-3}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$
GP	315.79 ± 176.38	30.68 ± 7.54	73.1 (13.0)	95.6 (5.0)	178.87 ± 103.15	21.68 ± 6.17	71.7 (14.8)	93.6 (5.5)	135.62 ± 82.94	18.19 ± 5.63	68.3 (14.8)	91.8 (6.4)
Uncon Supervised	208.16 ± 292.41	–	–	–	54.66 ± 203.76	–	–	–	12.99 ± 45.38	–	–	–
Cond Supervised	17.95 ± 34.60	–	–	–	9.52 ± 17.02	–	–	–	5.82 ± 12.28	–	–	–
Uncon AE	248.75 ± 377.03	21.20 ± 17.44	47.6 (26.5)	80.9 (17.8)	46.45 ± 132.21	6.63 ± 8.83	61.5 (22.4)	88.7 (12.0)	9.75 ± 29.25	3.14 ± 3.76	65.0 (19.3)	92.0 (9.0)
Cond AE	15.42 ± 24.84	4.80 ± 3.76	63.2 (24.5)	88.3 (8.6)	6.13 ± 11.70	2.89 ± 2.52	61.6 (23.4)	88.2 (8.2)	3.14 ± 8.68	1.99 ± 2.00	62.0 (22.5)	87.8 (8.4)
Uncon Diffusion	239.28 ± 291.64	19.93 ± 13.37	62.0 (22.5)	90.1 (10.7)	37.71 ± 107.23	5.02 ± 6.40	64.9 (25.2)	89.2 (12.2)	7.36 ± 22.12	2.30 ± 2.92	59.0 (27.0)	86.5 (14.8)
Cond Diffusion	15.79 ± 37.39	4.69 ± 4.47	60.3 (26.4)	90.4 (11.7)	7.12 ± 17.64	2.88 ± 3.02	62.3 (25.2)	89.8 (12.0)	3.11 ± 8.93	1.91 ± 2.10	63.9 (24.3)	89.5 (12.2)
Uncon Diffusion (DPS)	235.67 ± 310.17	19.50 ± 13.23	59.0 (21.4)	80.9 (16.5)	39.47 ± 118.32	5.34 ± 6.72	37.4 (35.8)	65.2 (31.3)	7.18 ± 20.47	2.55 ± 2.88	29.6 (41.1)	54.8 (41.1)
Cond Diffusion (DPS)	16.46 ± 38.82	5.07 ± 4.91	43.0 (31.9)	71.4 (25.2)	7.05 ± 18.35	3.13 ± 3.33	37.3 (35.5)	67.8 (29.2)	3.23 ± 9.87	2.14 ± 2.34	39.2 (32.9)	67.8 (28.6)

Model	$n_{obs} = 8$ NMSE ↓ ($\times 10^{-3}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{obs} = 10$ NMSE ↓ ($\times 10^{-3}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$
GP	97.03 ± 59.79	14.75 ± 4.77	64.6 (13.8)	89.7 (7.5)	78.78 ± 57.13	12.72 ± 4.44	63.2 (13.5)	88.8 (7.8)
Uncon Supervised	4.87 ± 12.89	–	–	–	2.87 ± 5.10	–	–	–
Cond Supervised	3.54 ± 6.18	–	–	–	2.90 ± 5.41	–	–	–
Uncon AE	4.85 ± 15.95	2.32 ± 2.94	67.6 (17.8)	92.6 (8.5)	4.25 ± 14.05	2.10 ± 2.83	67.7 (17.7)	92.8 (8.3)
Cond AE	1.40 ± 3.01	1.46 ± 1.16	62.5 (20.8)	88.1 (7.9)	1.03 ± 2.20	1.28 ± 0.99	60.4 (21.3)	87.1 (8.7)
Uncon Diffusion	2.45 ± 8.95	1.44 ± 1.62	59.8 (26.0)	87.0 (14.3)	1.09 ± 2.77	1.17 ± 1.06	58.9 (26.7)	85.4 (15.4)
Cond Diffusion	1.44 ± 3.54	1.40 ± 1.29	64.3 (22.7)	90.3 (11.2)	1.16 ± 3.31	1.24 ± 1.23	64.5 (23.0)	90.1 (11.1)
Uncon Diffusion (DPS)	2.74 ± 8.71	1.64 ± 1.57	31.5 (39.3)	55.9 (40.1)	1.40 ± 4.67	1.40 ± 1.25	30.5 (39.0)	54.8 (41.1)
Cond Diffusion (DPS)	1.43 ± 3.44	1.59 ± 1.41	40.2 (31.9)	69.8 (26.8)	1.14 ± 3.05	1.41 ± 1.29	40.5 (31.3)	70.2 (26.2)

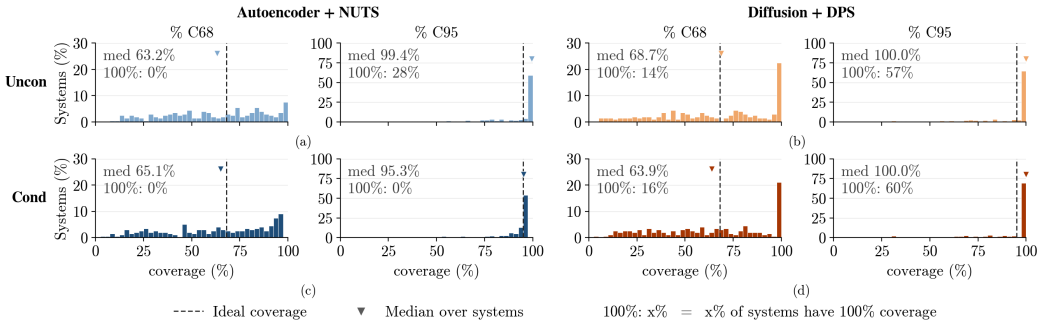

Figure 11: **Per-system coverage over 200 steady-state heat equation test systems at $n_{obs} = 4$.** Each panel shows the distribution of coverages across all systems, i.e. the percentage of systems falling in each coverage bin.

G.1.2 Experimental Setup

Fine-tuning. $\text{score}_{\text{field}}$ in (2) measures similarity between numerical fields, in this experiment we use a Gaussian kernel on the min-max normalised fields $\tilde{u}$,

$$\text{score}_{\text{field}}(\tilde{u}, \tilde{u}') = \exp\Big(-\frac{\|\tilde{u} - \tilde{u}'\|_2^2}{T}\Big), \tag{17}$$

which returns a value in $[0, 1]$. $T = 527.16$ controls how quickly the similarity decays with field distance, which is set so that $\text{score}_{\text{field}} = 0.5$ at the median of $\|\tilde{u} - \tilde{u}'\|_2$ over random pairs of training fields; this makes sure that similarity targets are spread across $[0, 1]$ rather than concentrated near either end. 3 epochs of fine-tuning on `all-MiniLM-L6-v2` are performed on $50,000$ training pairs with learning rate 5×10^{-5}.

Train. Supervised baselines use a U-Net with 64 base channels, **Cond Supervised** is trained for $1k$ epochs, while **Uncon Supervised** for $2k$ epochs with the Adam optimiser (Kingma & Ba, 2014), with a learning rate of 10^{-3} and batch size of 32; AE uses FNO with 2 layers (each for E and D), width 64 and 12 modes, latent dimension is 16, **Cond AE** is trained for $3k$ epochs and **Uncon AE** for $5k$, with a learning rate of 10^{-3} and batch size 200. We modify the original FNO structure into an autoencoder architecture, the encoder first applies FNO blocks to the field and then averages each channel over the spatial grid, and then a MLP maps the resulting vector to the latent; the decoder broadcasts the latent back to a constant field over the grid before its own FNO blocks; the score network is a U-Net with 32 base channels and a 64-dimensional conditioning vector, **Cond**

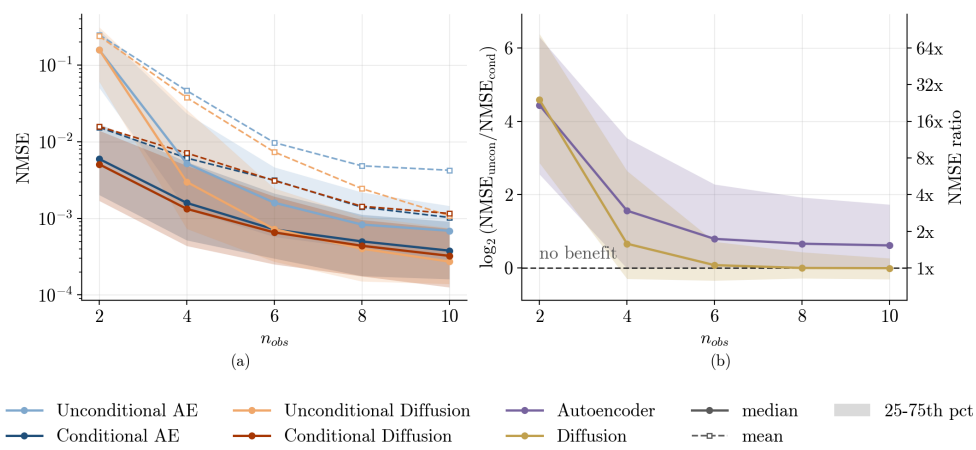

Figure 12: **NMSE performance and text gain at different $\mathbf{n_{obs}}$ for steady-state heat equation.** (a) NMSE across 200 test systems: median (solid), mean (dashed), and 25th–75th percentile bands, see Table 6 for exact values. (b) Per-system relative gain from text conditioning, $\log_2(\mathrm{NMSE}_{\mathrm{uncon}}/\mathrm{NMSE}_{\mathrm{cond}})$, with median and interquartile bands. Zero values indicate that the text adds nothing beyond the observations.

Table 6: **Quartiles of per-system NMSE across 200 heat-equation test systems.** Q_{25}, median and Q_{75} over the same 200 test systems across different n_{obs} are shown.

	$n_{\mathrm{obs}}=2$			$n_{\mathrm{obs}}=4$			$n_{\mathrm{obs}}=6$			$n_{\mathrm{obs}}=8$			$n_{\mathrm{obs}}=10$		
Model NMSE ($\times 10^{-3}$)	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}
Uncon AE	50.64	158.02	287.87	1.31	5.20	23.29	0.58	1.60	4.65	0.41	0.83	2.19	0.29	0.69	1.46
Cond AE	2.02	5.96	17.10	0.52	1.59	5.45	0.30	0.72	2.12	0.18	0.50	1.12	0.16	0.38	0.91
Uncon Diffusion	60.89	159.12	312.58	0.73	2.99	26.18	0.27	0.72	2.46	0.15	0.40	1.06	0.14	0.27	0.75
Cond Diffusion	1.70	5.08	13.97	0.44	1.33	4.87	0.25	0.66	1.93	0.18	0.44	0.96	0.13	0.32	0.74

Diffusion is trained for $3k$ epochs and **Uncon Diffusion** for $5k$, with learning rate of 2×10^{-4} and batch size of 32.

Inference. During inference, different numbers of n_{obs} at random locations for each test system are drawn, with adding Gaussian noise of standard deviation $\sigma_y = 1$. **GP** regression uses RBF kernel with maximum marginal likelihood estimation of the hyperparameters. AE posteriors are sampled with NUTS in the latent space, using a single chain with $M_o = 100$ warmup steps and $M = 200$ retained draws. The maximum tree depth is 5 for **Cond AE** and 8 for **Uncon AE**, as it is harder for **Uncon AE** to converge. Diffusion posteriors are drawn by integrating the guided reverse SDE over $L = 800$ uniform steps and the guidance clip $c = 1.0$. The guidance clip is chosen so that around 10% of initial reverse diffusion steps are clipped, leaving the remainder unaffected and preserving the relative scale of prior and likelihood scores.

G.1.3 Additional Results

Next, we provide some additional results on the first experiment.

Inference Results. Firstly, we present the full table of inference results across $n_{\mathrm{obs}} = \{2, 4, 6, 8, 10\}$ in Table 5. The last two rows report standard DPS, with the guidance strength ζ' tuned over the range recommended by Chung et al. (2023), and we take $\zeta' = 0.75$. Comparing standard DPS with our modified one in Table 5, the two methods achieve similar NMSE but standard DPS attains much lower coverage, meaning that its posteriors are over-confident. Standard DPS has higher CRPS as well. Figure 10 shows, for a single system at $n_{\mathrm{obs}} = 4$, the posterior inference results including the posterior mean, posterior standard deviation and absolute error of the mean of five probabilistic models. It can be seen that unconditional models produce both a larger posterior

Table 7: **Effective sample size (ESS) of NUTS on the heat equation, averaged over the 200 test systems.** This table reports ESS_{min} and ESS_{mean}, the minimum and the mean over the 16-dimensional latent vector of $M = 200$ posterior samples.

	$n_{\text{obs}} = 2$		$n_{\text{obs}} = 4$		$n_{\text{obs}} = 6$		$n_{\text{obs}} = 8$		$n_{\text{obs}} = 10$	
Model	ESS_{min}	ESS_{mean}	ESS_{min}	ESS_{mean}	ESS_{min}	ESS_{mean}	ESS_{min}	ESS_{mean}	ESS_{min}	ESS_{mean}
Uncon AE	53.7	116.2	62.0	124.1	64.6	127.0	65.6	128.4	63.9	124.0
Cond AE	102.9	191.7	100.5	190.2	100.4	188.7	99.7	186.2	100.8	193.4

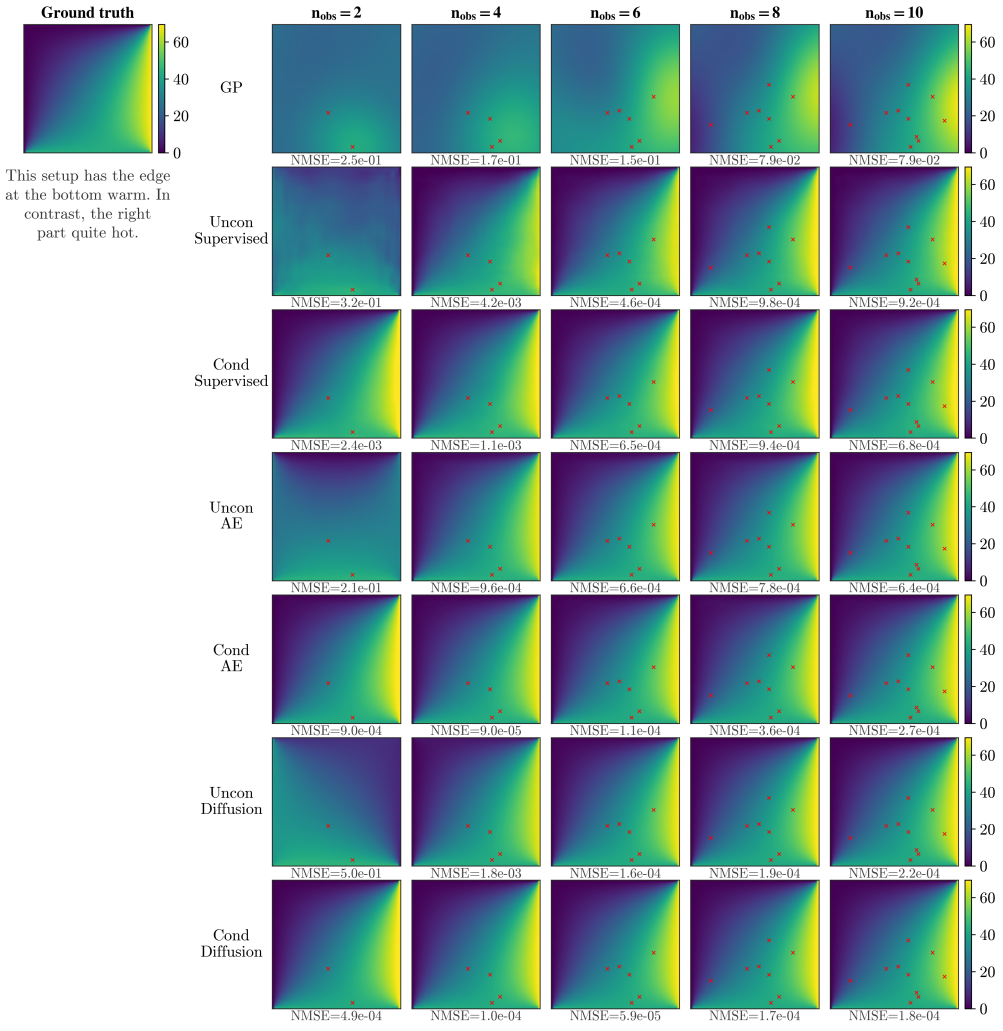

Figure 13: **Posterior inference of one steady-state heat equation test system as the number of observation points increases.** The leftmost panel shows the ground truth field together with the text description. Columns give $n_{obs} = \{2, 4, 6, 8, 10\}$ with nested observation sets. NMSE is also provided at the bottom of each field.

variance and a larger error. Figure 11 plots the distribution of coverages across test systems, showing that coverages for diffusion models are less concentrated around targeted coverage level. Table 7 reports effective sample size (ESS) of the NUTS chains, showing that **Cond AE** has larger ESS than **Uncon AE**.

Prior Generation. Figure 9 provides samples drawn directly from the learned priors, with text provided to conditional models. Priors generated from conditional models are more consistent with the described field, whereas samples from unconditional ones spread over the whole dataset.

Ablation on number of observation points and text contribution. Figure 13 shows the performance of all seven models over a single system with n_{obs} sweeping from 2 to 10; for supervised models the network output is shown, and for probabilistic ones the posterior mean is provided. Unconditional models are visibly inaccurate at small n_{obs}, due to the fact that without additional text information, observations alone cannot determine the field. Figure 12(a) reports the mean, median and interquartile range of the posterior mean NMSE over 200 test systems for the four generative models, with exact values provided in Table 6. The gap between conditional and unconditional models widens as n_{obs} decreases. Figure 12(b) makes this explicit by calculating per-system relative gain from $\log_2(\text{NMSE}_{\text{uncon}}/\text{NMSE}_{\text{cond}})$, showing median and interquartile range across systems. A zero value of text gain indicates that the text adds nothing beyond observations.

G.2 Helmholtz Equation

G.2.1 Problem Setup

The second experiment is based on a damped Helmholtz equation on the unit disk $\Omega = \{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\}$ with a Neumann boundary,

$$\begin{aligned} \big(\Delta + \kappa + i\gamma\kappa\big)u(x) &= f(x), \quad x \in \Omega, \\ \partial_n u(x) &= 0, \qquad x \in \partial\Omega, \end{aligned} \tag{18}$$

forced uniformly over a small disk,

$$f(x) = A\, \mathbb{1}_{\{\|x-c\|_2 \leq R\}}, \qquad R = 0.1. \tag{19}$$

Here $\kappa = 200$ is the squared wavenumber, γ the damping ratio, and A and c the amplitude and the coordinate of the forcing centre. Each system draws c uniformly over the inner disk of radius 0.8, and A and γ log-uniformly from $[1, 10]$ and $[0.02, 0.2]$. The solution is complex, so each system consists of a real and an imaginary field. The Finite element method (FEM) in FEniCSx is used on an unstructured triangular mesh of target size $h = 0.03$. Then the solution is interpolated onto a 96×96 grid over $(-1.1, 1.1)^2$ masked to the disk.

About text generation, descriptions are produced by prompting *DeepSeek-V4-Flash* (Liu et al., 2024a) through its API. Each text consists of three pieces of information in qualitative natural language format: the damping across the whole disk, the magnitude of forcing, and its location, in terms of its angular position and a rough distance from the centre. For forcing magnitude A, damping γ and radial position $\|c\|_2$, numerical parameters are first binned following Table 8. Each bin carries a small pool of anchor words and phrases, one of which is drawn at random and passed to the model, so that numerical values are never shown. Anchor words are different for generating train and test text descriptions. The prompt instructs the model to replace the anchor word with a synonym of the same strength rather than copy it, which preserves the meaning of a description without reusing the same words. The angular position of forcing is reported at one of three different resolutions: a quadrant ($90°$) (eg: in the north-east quadrant), one of eight compass points ($45°$) (eg: in the west), or one of twelve clock directions ($30°$) (eg: one o'clock direction), providing more ambiguity in generated texts.

The following context provides the two halves to be provided to API. The system prompt is fixed across the whole dataset, which states the task and rules the descriptions must obey. The user message carries the content, with one line per system, giving its three binned facts. A few examples of Helmholtz fields and corresponding texts are provided in Figure 14.

System prompt

```
Suppose you are an expert doing experiments on damped Helmholtz equation
on a disc, where you need to state experimental setups, including
magnitude of damping, strength and location of force. You write short,
natural language-style descriptions of experimental setups for each
system, with three facts provided for each configuration:
```

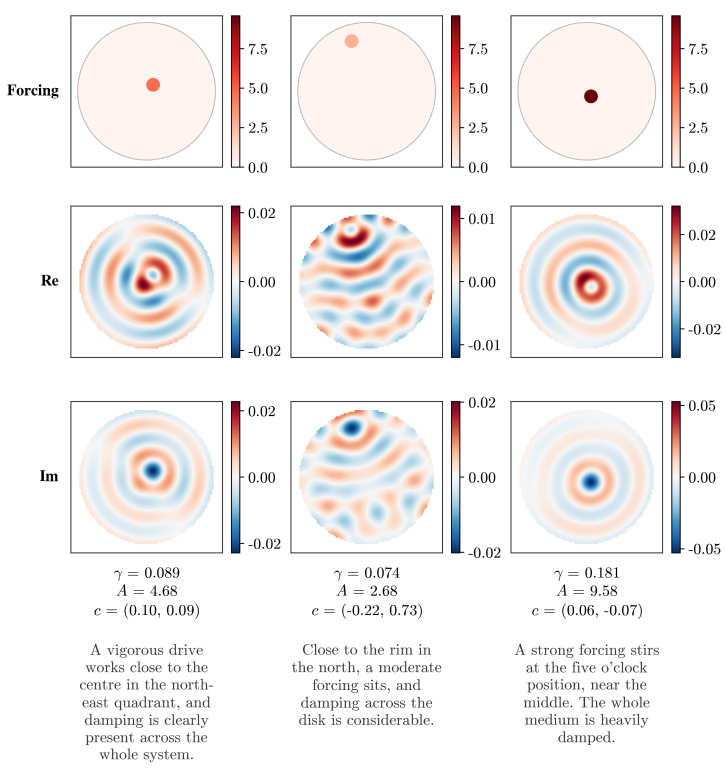

Figure 14: Examples of Helmholtz fields with forcing fields, related parameters and descriptions.

Table 8: **Binning of the Helmholtz parameters for text generation.** Only the bin labels are passed to the language model and the numerical values are never shown.

Forcing magnitude A		Damping γ		Radial position $\|c\|_2$	
Level	Range	Level	Range	Level	Range
faint	1.00–2.15	very light	0.020–0.036	centre	0.00–0.25
moderate	2.15–4.64	light	0.036–0.063	mid	0.25–0.55
strong	4.64–10.0	moderate	0.063–0.113	rim	0.55–0.80
		heavy	0.113–0.200		

```
- forcing position: where the forcing is applied (with angular
  directions at different resolutions: a quadrant, a compass direction,
  or a clock direction, plus a rough distance from the centre)
- forcing magnitude: an anchor word showing how large the forcing is
- damping: an anchor phrase for how strongly damped the WHOLE disk is
  (one global property of the medium, without specific location)

Write ONE description per item. Rules:
- Every description is one or two COMPLETE sentences, each with a subject
  and a verb.
- Include ALL three facts. Although the items are listed in a fixed order
  (position, magnitude, damping) due to input formatting, please do NOT
  write them in the same order. Start with describing forcing as well.
- Do NOT use the exact given anchor word. Represent the same level using
  a different word, but the level must stay unchanged: a low-level anchor
  must never use words sounding strong, nor a high-level anchor mild.
- Restructure the damping phrase as well, keep the level it represents the
  same but do not use the same wording.
- Vary your verbs. Do not reuse "force is applied" or "force is present",
```

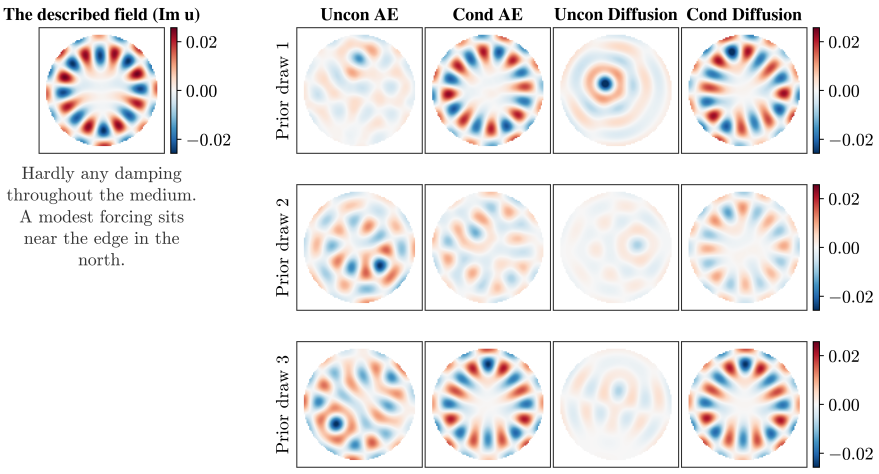

Figure 15: Prior samples generated from four models (with text provided to conditional models).

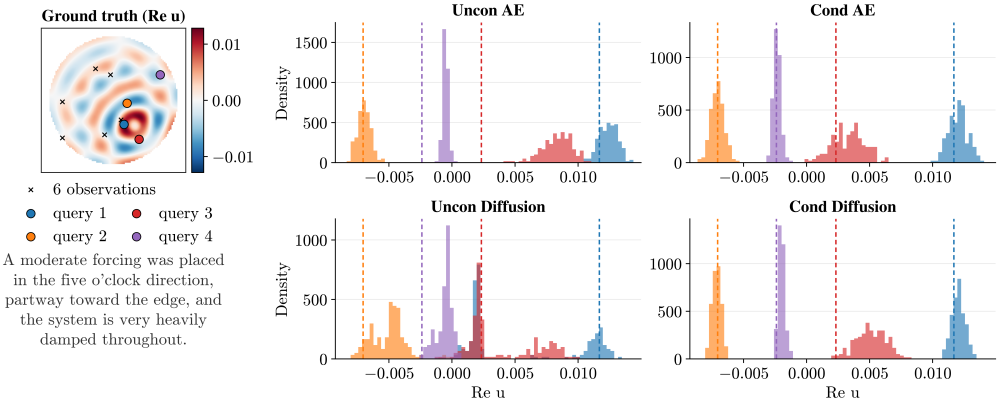

Figure 16: Posterior marginals of one Helmholtz equation test system (Real part) (same system as in Figure 3) at four unobserved query locations when $n_{\text{obs}} = 6$.

```
  can use verbs like drives, pushes, stirs, sits, hums, acts, hits, works
  away, was placed, etc.
- Within one request, do not use the same magnitude word twice, use a
  different synonym instead.
- Damping is GLOBAL, so phrase it as being applied on the whole disk.
- Keep the position EXACTLY as coarse as stated. Never create a more
  precise position, never make it vaguer, never change direction words.
- Never use digits, and say clock hours using words, and ALWAYS add a
  disambiguating word so it reads like a direction, not a time of day.
  (eg: use "the two o'clock direction", "toward two o'clock", "at two
  o'clock on the dial", but "at two o'clock" is WRONG.
```

User message

```
Facts:
1. forcing position: the north-west quadrant, close to the rim |
   forcing magnitude: weak |
   damping: the disk as a whole is barely damped
2. forcing position: the four o'clock direction, halfway towards the edge |
   forcing magnitude: medium-strength |
   damping: the system as a whole is clearly damped ...
```

Table 9: **Full inference table on the Helmholtz equation test set.** This table extends Table 2 with n_{obs} between 4 and 12 over 100 test fields.

Model	$n_{\text{obs}} = 4$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{\text{obs}} = 6$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{\text{obs}} = 8$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$
GP	92.45 ± 8.42	74.49 ± 31.10	81.4 (23.0)	93.1 (8.1)	86.74 ± 9.20	74.66 ± 33.42	83.3 (23.0)	94.0 (7.4)	81.62 ± 10.47	69.13 ± 28.71	81.4 (23.0)	93.0 (8.2)
Uncon Supervised	54.07 ± 28.83	–	–	–	25.93 ± 23.18	–	–	–	14.17 ± 12.89	–	–	–
Cond Supervised	28.39 ± 34.74	–	–	–	19.77 ± 25.27	–	–	–	13.13 ± 17.64	–	–	–
Uncon AE	94.21 ± 186.28	45.20 ± 36.74	37.7 (33.7)	60.4 (35.9)	66.61 ± 338.05	28.47 ± 42.73	45.9 (28.7)	68.6 (28.3)	29.20 ± 64.45	22.53 ± 26.28	46.5 (27.6)	68.4 (28.4)
Cond AE	23.89 ± 46.08	18.68 ± 20.59	67.7 (18.6)	89.6 (10.7)	7.15 ± 17.18	9.62 ± 9.43	71.4 (13.6)	94.5 (5.5)	4.07 ± 8.99	7.79 ± 6.78	70.3 (12.5)	94.1 (5.6)
Uncon Diffusion	48.08 ± 30.47	32.30 ± 14.43	57.5 (19.7)	88.8 (10.5)	22.57 ± 24.20	18.26 ± 13.38	64.4 (18.0)	93.4 (6.7)	9.00 ± 13.21	9.95 ± 8.67	68.7 (18.0)	93.9 (7.0)
Cond Diffusion	20.14 ± 30.17	18.96 ± 16.23	50.5 (26.1)	76.6 (21.2)	7.15 ± 12.63	10.51 ± 9.22	54.5 (23.5)	83.0 (15.3)	4.30 ± 7.95	7.90 ± 7.09	54.8 (21.7)	82.2 (15.9)
Uncon Diffusion (DPS)	47.79 ± 27.75	32.16 ± 13.37	56.6 (19.1)	88.8 (9.7)	24.18 ± 23.35	19.23 ± 12.81	61.5 (17.3)	90.6 (8.0)	10.48 ± 13.98	10.91 ± 8.83	62.7 (18.2)	90.2 (8.7)
Cond Diffusion (DPS)	19.12 ± 28.91	19.21 ± 15.71	45.2 (28.3)	72.6 (24.0)	7.19 ± 13.22	11.40 ± 9.50	47.4 (26.8)	76.9 (19.7)	4.42 ± 8.75	7.91 ± 7.30	46.9 (24.9)	75.2 (20.9)

Model	$n_{\text{obs}} = 10$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{\text{obs}} = 12$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$
GP	77.12 ± 11.96	67.76 ± 29.42	82.3 (23.1)	93.6 (7.7)	72.21 ± 12.94	65.36 ± 28.73	82.3 (23.0)	93.7 (7.6)
Uncon Supervised	10.17 ± 7.68	–	–	–	8.45 ± 8.23	–	–	–
Cond Supervised	10.84 ± 16.48	–	–	–	8.84 ± 13.23	–	–	–
Uncon AE	16.07 ± 30.13	17.42 ± 18.26	46.1 (26.9)	68.4 (28.5)	12.72 ± 29.08	14.82 ± 17.04	46.9 (26.8)	69.0 (27.7)
Cond AE	3.16 ± 7.25	6.77 ± 5.80	69.7 (11.8)	93.8 (5.4)	2.35 ± 4.83	6.00 ± 4.51	68.9 (12.0)	93.1 (5.7)
Uncon Diffusion	5.36 ± 9.24	7.20 ± 6.99	67.6 (19.1)	92.0 (8.4)	3.09 ± 6.62	5.38 ± 5.46	66.8 (17.6)	92.7 (7.6)
Cond Diffusion	3.45 ± 7.54	6.75 ± 6.64	54.7 (20.7)	83.6 (14.1)	2.52 ± 4.92	5.80 ± 5.63	54.5 (20.3)	83.9 (13.8)
Uncon Diffusion (DPS)	5.93 ± 9.74	7.61 ± 7.04	61.6 (18.8)	88.3 (10.1)	3.95 ± 7.94	6.02 ± 5.96	58.2 (18.3)	86.5 (10.5)
Cond Diffusion (DPS)	3.63 ± 8.95	6.79 ± 7.13	47.2 (24.1)	75.8 (20.3)	2.71 ± 5.94	5.85 ± 5.97	47.0 (23.9)	76.4 (19.3)

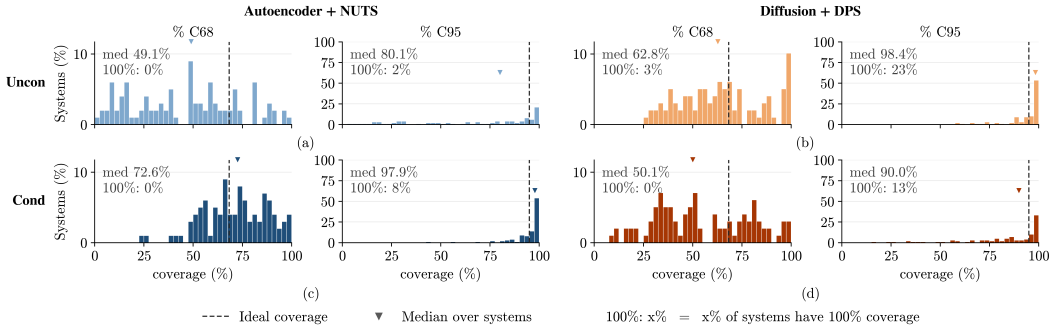

Figure 17: Per-system coverages over 100 Helmholtz test systems at $n_{\text{obs}} = 6$.

G.2.2 Experimental Setup

Fine-tuning. Damped Helmholtz fields oscillate, so an L_2 distance as in (17) is a poor measure of how far apart two fields are. Instead, we compare the pointwise energy $E(x) = |u(x)|^2$ calculated from real and imaginary channels, and use the sliced Wasserstein distance between normalised energy fields $\tilde{E}(x) = E(x)/\|u\|_2^2$, where $\|u\|_2^2 = \sum_x E(x)$ is the total energy. Since the normalisation discards the magnitude, we append the difference in log total energy,

$$\text{score}_{\text{field}}(u, u') = \exp\Big(-\frac{c_1\,\text{SW}_2^2(\tilde{E}, \tilde{E}') + c_2\left|\log\|u\|_2^2 - \log\|u'\|_2^2\right|^2}{T}\Big), \tag{20}$$

where $\text{SW}_2^2(\bullet, \bullet)$ measures the squared sliced Wasserstein distance between two normalised energy fields, c_1 and c_2 scale the two terms to similar magnitudes, and $T = 340.02$ is the bandwidth. 3 epochs of fine-tuning `all-MiniLM-L6-v2` are performed on $50,000$ training pairs with learning rate 5×10^{-5}.

Train. Supervised baselines use a U-Net with 96 base channels, **Cond Supervised** is trained for $2k$ epochs, while **Uncon Supervised** for $2.5k$ epochs, with a learning rate of 10^{-3} and batch size of 32; AE uses FNO with 4 layers (each for E and D), width 56 and 14 modes, latent dimension is 16, **Cond AE** is trained for $4k$ epochs and **Uncon AE** for $6k$, with a learning rate of 10^{-3} and batch size 200; the score network is a U-Net with 80 base channels and a 128-dimensional conditioning vector, **Cond Diffusion** is trained for $3k$ epochs and **Uncon Diffusion** for $4k$, with learning rate of 2×10^{-4} and batch size of 32.

Inference. At inference, observations are drawn randomly from both real and imaginary fields, with the same noise standard deviation $\sigma_y = 6 \times 10^{-4}$. Following Caviedes-Nozal et al. (2021), **GP**

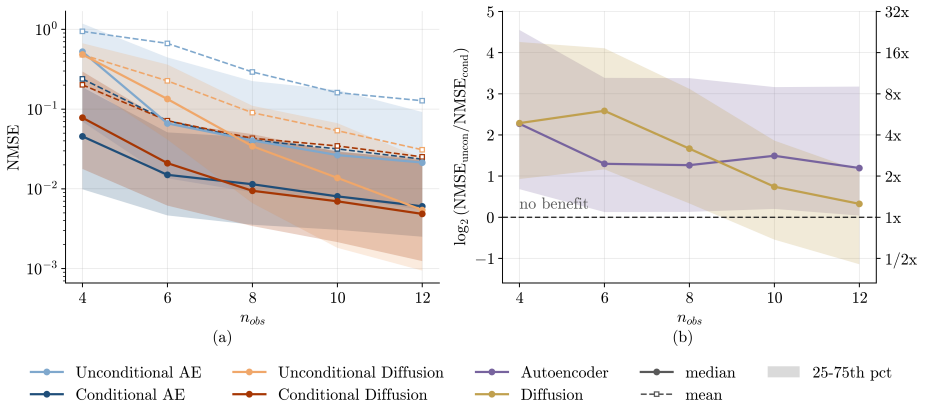

Figure 18: **NMSE performance and text gain at different $\mathbf{n_{obs}}$ for Helmholtz equation.** (a) NMSE across 100 test systems: median (solid), mean (dashed), and 25th–75th percentile bands, see Table 10 for exact values. (b) Per-system relative gain from text conditioning, $\log_2(\mathrm{NMSE}_{\mathrm{uncon}}/\mathrm{NMSE}_{\mathrm{cond}})$, with median and interquartile bands.

Table 10: Quartiles of per-system NMSE across 100 Helmholtz test systems at different n_{obs}.

	$n_{\mathrm{obs}}=4$			$n_{\mathrm{obs}}=6$			$n_{\mathrm{obs}}=8$			$n_{\mathrm{obs}}=10$			$n_{\mathrm{obs}}=12$		
Model NMSE ($\times 10^{-2}$)	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}
Uncon AE	6.98	52.42	118.06	1.36	6.62	44.52	0.91	4.04	22.46	0.78	2.65	17.39	0.55	2.14	9.19
Cond AE	0.99	4.55	18.83	0.46	1.50	5.11	0.35	1.14	4.41	0.31	0.80	2.99	0.25	0.60	2.26
Uncon Diffusion	21.18	48.36	66.97	4.16	13.37	36.19	0.66	3.42	11.01	0.18	1.37	6.66	0.09	0.54	2.40
Cond Diffusion	1.78	7.79	29.61	0.62	2.10	6.79	0.34	0.94	4.86	0.21	0.70	2.91	0.12	0.48	2.39

uses a modified Bessel kernel

$$k(x,x') = \sigma_f^2 \exp\Big(-\frac{\|x-x'\|_2^2}{2\ell^2}\Big) J_0\big(k_0\|x-x'\|_2\big), \tag{21}$$

where $\sigma_f^2 J_0\big(k_0\|x-x'\|_2\big)$ is the Bessel kernel, modulated by an exponential envelope that accounts for the damping in our case, where $k_0 = \sqrt{\kappa} = \sqrt{200}$, and σ_f and ℓ are hyperparameters to be learned using maximum likelihood estimation. AE posteriors are sampled using a single NUTS chain with $M_o = 100$ warmup steps and $M = 200$ retained draws. The maximum tree depth is 10 for both **Cond AE** and **Uncon AE**. Diffusion posteriors are drawn by integrating the guided reverse SDE over $L = 800$ uniform steps and the guidance clip $c = 1.0$.

G.2.3 Additional Results

Inference Results. Table 9 gives the full inference results across a variety of n_{obs}. Again, the last two rows report standard DPS, using a tuned guidance strength $\zeta' = 0.3$. Standard DPS reaches an NMSE comparable to the modified sampler but provides worse posterior coverage. From both Table 9 and posterior histograms in Figure 16, conditional generative models provide more accurate inference results. Figure 17 shows coverage distributions over 100 test systems. **Cond AE** is the best calibrated overall, with the most systems concentrated near the nominal levels (median $\%C_{68} = 72.6\%$ and $\%C_{95} = 97.9\%$). **Uncon Diffusion** reports average coverages close to nominal values in Table 9, but Figure 17(b) shows this comes from 23% of systems reaching $\%C_{95} = 100\%$, i.e. from posteriors that are far too dispersed rather than from calibration. **Uncon Diffusion** also has higher CRPS values than **Cond Diffusion**.

Prior Generation. Figure 15 provides a few prior draws from the four generative models. With the text supplied, conditional priors concentrate on fields consistent with the description, showing the low damping and the forcing near the northern rim, while unconditional draws vary freely over the whole training distribution.

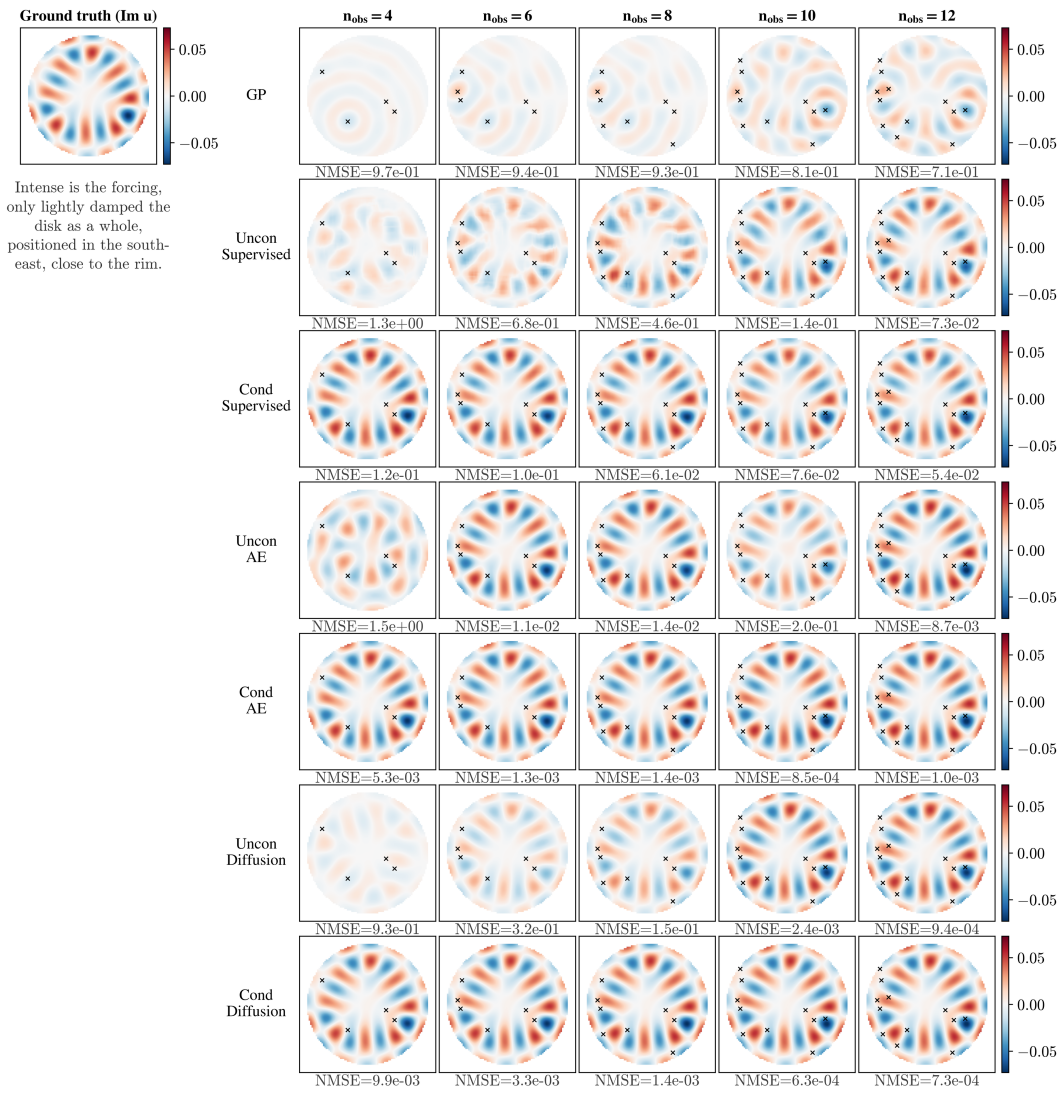

Figure 19: Posterior inference of one Helmholtz equation test system (Imaginary part) as the number of observation points increases.

Ablation on number of observation points and text contribution. Figures 19 and 20 visualise inference results as n_{obs} progresses. **Cond Supervised**, **Cond AE** and **Cond Diffusion** all perform better than unconditional ones, and the margin is largest when observations alone are least informative. Figure 18 plots NMSE means and quartiles, and per-system text gain across different n_{obs}, with the corresponding NMSE values in Table 10. The gain shrinks as observation data accumulate but does not vanish: even at $n_{\text{obs}} = 12$ the median gain is about $2.3\times$ for the autoencoder and $1.3\times$ for diffusion, so the information from text still contributes.

G.3 UK Weather Data

G.3.1 Problem Setup

We download dataset from CERRA (Ridal et al., 2024), which is a regional reanalysis dataset for a domain covering all Europe from 1984. The original dataset is at a horizontal resolution of $5.5km$ on a Lambert conformal grid. We restrict it to UK land by intersecting the CERRA land-sea mask with the Natural Earth country boundary, and then resample the retained points onto a regular 96×96

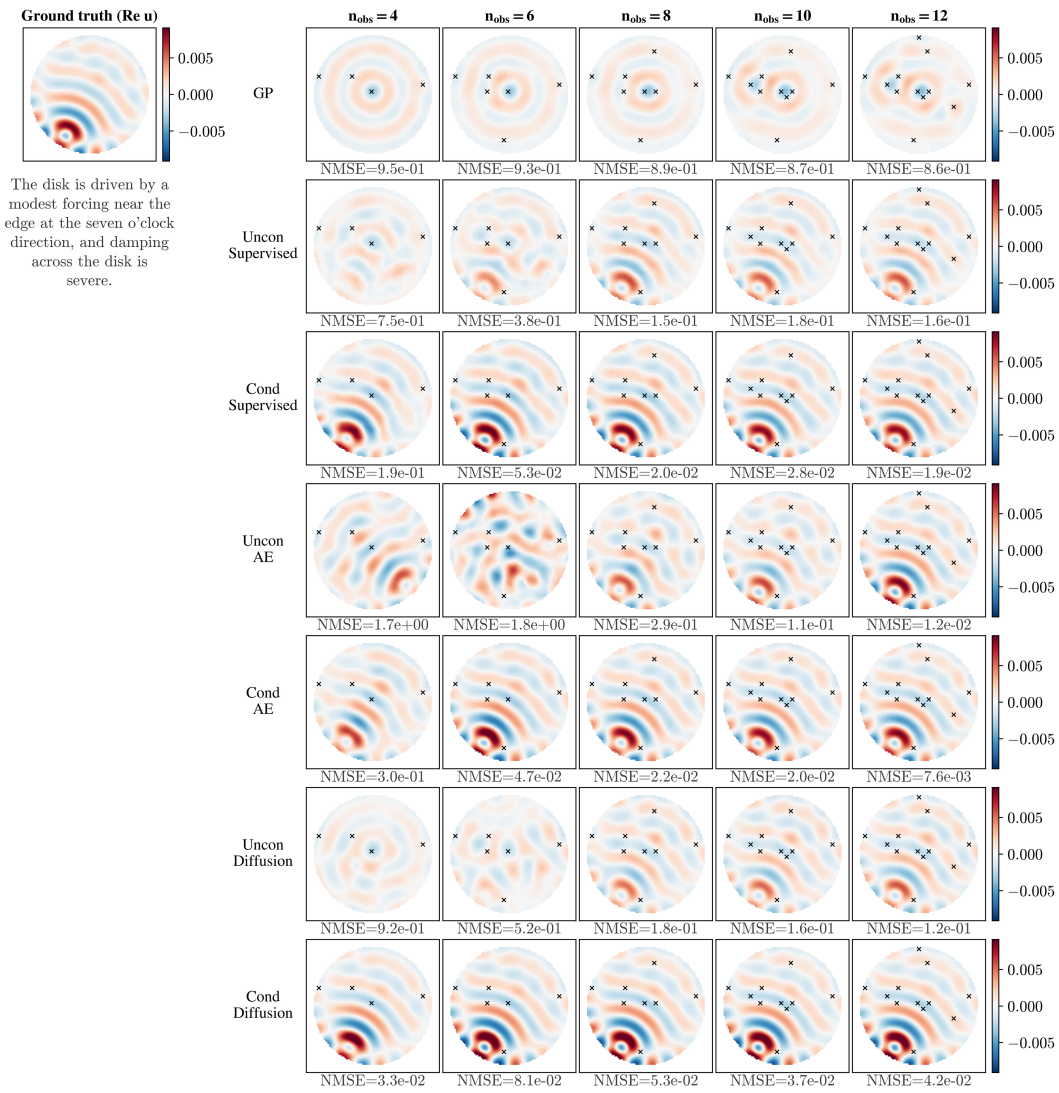

Figure 20: Posterior inference of one Helmholtz equation test system (Real part) as the number of observation points increases.

latitude-longitude grid spanning $49.99°$–$60.75°$ N and $8.08°$ W–$1.74°$ E. Of the 9216 grid points, 3135 fall on land, and the remainder are set to zero and excluded from every loss and metric.

Three channels are obtained from CERRA, *2m temperature*, *10m wind speed* and *10m wind direction*, with the latter two converted to horizontal and vertical wind velocities, making the final three channels $\{T, u, v\}$. We take three daily analysis times 09:00, 15:00 and 21:00 from 2000 to 2025, giving $28{,}491$ total snapshots, of which the $27{,}396$ from 2000-2024 are used for training.

To generate text for this dataset, we first divide the grid into 25 regions, A1-E5, of which 18 contain enough land to be described, as in Figure 4. The text corresponding to each field describes the temperature, wind speed and direction in 8 randomly chosen regions, mimicking a setting in which much of the domain carries no sensors but people can report what the weather feels like. As in the Helmholtz experiment, the regional averages are first binned into levels shown in Table 11. For wind speed, averages are calculated as the mean of pointwise speed $\sqrt{u^2 + v^2}$. Its direction is calculated from the mean wind velocity $(\bar{u}, \bar{v})$ and reported as one of eight compass points (eg: north-west).

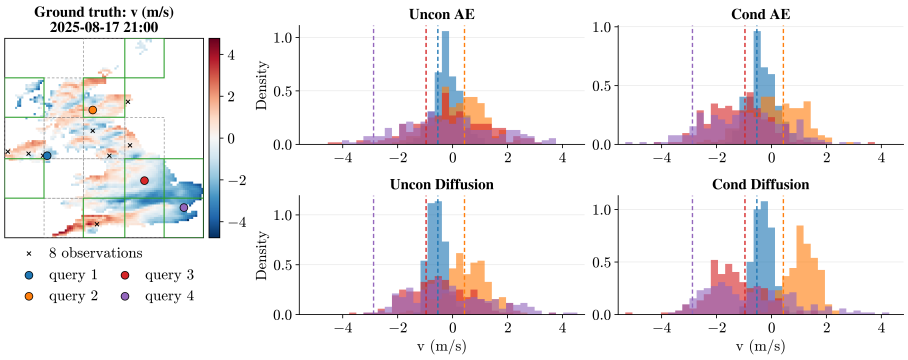

Figure 21: Posterior marginals of one UK weather test system (v) (same system as in Figure 5) at four unobserved query locations when $n_{\text{obs}} = 8$.

Table 11: **Binning of the regional averages for text generation in UK weather dataset.** Bands overlap at their edges: a value falling in two bands is described by one of them, chosen at random.

Temperature (°C)		Wind speed ($\mathrm{m\,s^{-1}}$)	
Level	Range	Level	Range
freezing	< 3	calm	< 2.2
cold	2–6	light	1.8–4.0
chilly	5–9	moderate	3.5–5.5
cool	8–12	breezy	5.0–7.5
mild	11–15	strong	> 7.0
warm	14–19		
very warm	18–24		
hot	> 24		

We again query *DeepSeek-V4-Flash* API to generate text, with the system prompt and user message provided below. Same as in Helmholtz experiment, each bin carries a pool of anchor words, one of which is drawn and passed to the model, and the prompt asks for a synonym.

System prompt

```
You write short weather descriptions as a PERSON describing how current
weather feels like in the United Kingdom, and use relaxed, everyday
language in flowing sentences. NOT a forecast/bulletin: avoid
jargon ('highs/lows of', 'mph', 'outbreaks', 'winds veering').

The UK is divided into a grid of areas, each with a short LABEL: the
letter runs from west to east and the number from south to north. Each
system below gives a DIFFERENT set of areas, listed in a consistent
geographic order. For each you get its TEMPERATURE band,
WIND-STRENGTH band, and WIND DIRECTION.

Full ordered band scale (your reference for meaning + neighbours):
  TEMPERATURE bands (approx degC, overlap at edges), coldest -> hottest:
    freezing (below 3C)  <  cold (2-6C)  <  chilly (5-9C)  < ...
  WIND-SPEED bands (approx m/s, overlap at edges):
    calm (below 2.2 m/s)  <  light (1.8-4 m/s)  < ...

For each variant write ONE description covering three facts of EVERY
listed area. Rules:
  - Describe the areas IN THE ORDER GIVEN in user message so the
    description flows geographically. NAME each area by its grid label.
  - ONE anchor phrase is provided for each region. The anchor fixes the
    INTENSITY. But please change using synonym, using words which person
```

Table 12: Full inference results on 100 UK weather test systems, which is an extension of Table 3.

Model	$n_{\text{obs}} = 4$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{\text{obs}} = 6$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{\text{obs}} = 8$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$
GP	15.05 ± 19.33	18.16 ± 11.44	70.0 (9.5)	92.6 (4.7)	13.10 ± 16.83	16.71 ± 10.37	71.0 (9.4)	93.0 (4.3)	12.12 ± 16.06	15.91 ± 10.12	68.3 (8.5)	91.7 (5.0)
Uncon Supervised	10.81 ± 14.02	–	–	–	8.76 ± 11.89	–	–	–	7.46 ± 11.38	–	–	–
Cond Supervised	8.96 ± 12.98	–	–	–	7.51 ± 10.54	–	–	–	6.78 ± 9.95	–	–	–
Uncon AE	11.30 ± 15.16	15.21 ± 9.38	69.5 (8.0)	94.6 (3.5)	9.28 ± 12.43	13.57 ± 8.50	68.6 (7.3)	94.3 (3.2)	8.18 ± 11.55	12.58 ± 8.02	67.5 (6.5)	93.6 (3.3)
Cond AE	9.27 ± 13.45	13.70 ± 8.67	66.3 (6.7)	92.6 (3.9)	7.68 ± 10.44	12.45 ± 7.64	66.4 (6.0)	92.8 (3.4)	6.93 ± 9.81	11.68 ± 7.37	65.9 (5.6)	92.5 (3.5)
Uncon Diffusion	10.49 ± 12.87	14.79 ± 8.85	55.6 (13.5)	85.9 (9.3)	8.81 ± 12.16	13.14 ± 8.53	56.1 (12.8)	86.4 (8.7)	7.72 ± 11.25	12.07 ± 8.06	56.2 (12.4)	86.5 (8.7)
Cond Diffusion	9.31 ± 15.15	13.55 ± 9.37	55.8 (12.9)	85.9 (9.2)	7.49 ± 10.55	12.22 ± 7.86	56.0 (12.5)	86.0 (9.0)	6.81 ± 10.13	11.45 ± 7.65	56.0 (12.5)	86.0 (9.1)

Model	$n_{\text{obs}} = 10$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{\text{obs}} = 12$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$	$n_{\text{obs}} = 14$ NMSE ↓ ($\times 10^{-2}$)	CRPS ↓ ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$
GP	10.78 ± 14.30	14.98 ± 9.36	69.3 (8.2)	92.1 (4.6)	10.16 ± 13.80	14.43 ± 9.08	69.7 (8.1)	92.1 (4.6)	9.48 ± 13.31	13.90 ± 8.80	69.3 (7.9)	91.9 (4.5)
Uncon Supervised	6.52 ± 10.26	–	–	–	6.22 ± 10.47	–	–	–	5.80 ± 9.93	–	–	–
Cond Supervised	6.26 ± 9.53	–	–	–	5.89 ± 8.81	–	–	–	5.67 ± 8.72	–	–	–
Uncon AE	7.07 ± 10.04	11.66 ± 7.37	66.9 (6.2)	93.2 (3.2)	6.90 ± 10.50	11.38 ± 7.41	65.8 (5.9)	92.3 (3.6)	6.51 ± 10.17	10.99 ± 7.22	64.8 (5.9)	91.6 (4.1)
Cond AE	6.22 ± 8.94	11.07 ± 6.90	65.0 (6.1)	91.8 (3.9)	6.12 ± 9.58	10.82 ± 6.93	63.8 (6.6)	90.9 (4.5)	5.87 ± 9.01	10.56 ± 6.76	62.5 (6.9)	90.3 (5.0)
Uncon Diffusion	6.43 ± 9.54	11.10 ± 7.23	56.2 (12.2)	86.5 (8.5)	6.32 ± 11.66	10.68 ± 7.36	56.3 (12.2)	86.5 (8.5)	5.97 ± 11.83	10.25 ± 7.22	56.3 (12.1)	86.6 (8.5)
Cond Diffusion	6.03 ± 8.88	10.76 ± 7.03	55.8 (12.6)	85.7 (9.3)	6.05 ± 11.02	10.45 ± 7.13	55.7 (12.6)	85.8 (9.2)	5.69 ± 10.45	10.05 ± 6.90	55.9 (12.4)	85.8 (9.2)

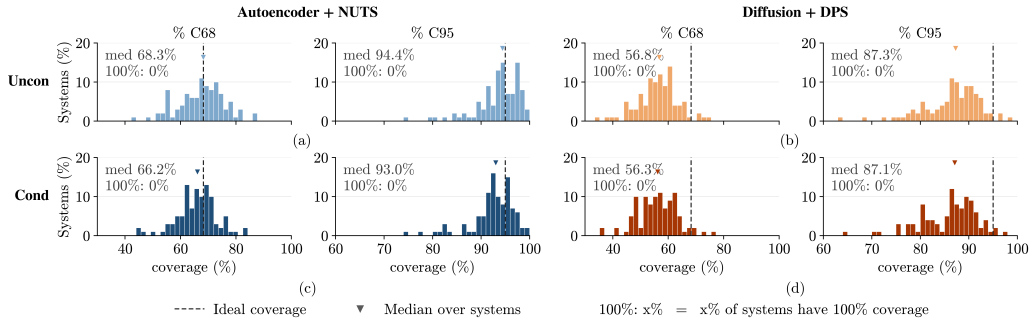

Figure 22: Per-system coverage over 100 Weather test fields with $n_{\text{obs}} = 8$.

```
    would use. Pasting exactly the anchor word is the BORING fallback.
    NEVER shift the intensity to a different band.
  - VARY the sentence structure and the fact order from area to area.
    Never produce a chain of 'X is <word> with a <word> wind'.
  - Wind directions should always end with a NOUN ('wind', 'breeze',
    'gusts', etc.). A 'variable' wind direction indicates a light or
    shifting wind.
  - When several areas share the same direction, describe the shared
    airflow ONCE and then only describe each area by its temperature and
    wind strength.
  - DO NOT use comparatives (cooler, windier, 'more...') and similarity
    references ('just as', 'the same wind', 'likewise') to compare
    between areas.
  - Never create sentences describing rain, sunshine, etc.
```

User message

```
Facts (grid label -- temp band | wind band | direction):
  - B1: agreeable | a persistent breeze | southerly
  - B2: raw | moderate | south-westerly ...
```

G.3.2 Experimental Setup

Fine-tuning. Each text description in this experiment describes a random 8 of 18 regions covering UK land, so L_2 distance between whole fields is not suitable to be matched to distance between text embeddings. Let A and B be the region subsets described by text s_A and s_B, let $\tilde{u}$ be the stack of the three channels $\{T, u, v\}$ after z-score normalisation, and let $\tilde{u}|_A$ represent the field values over

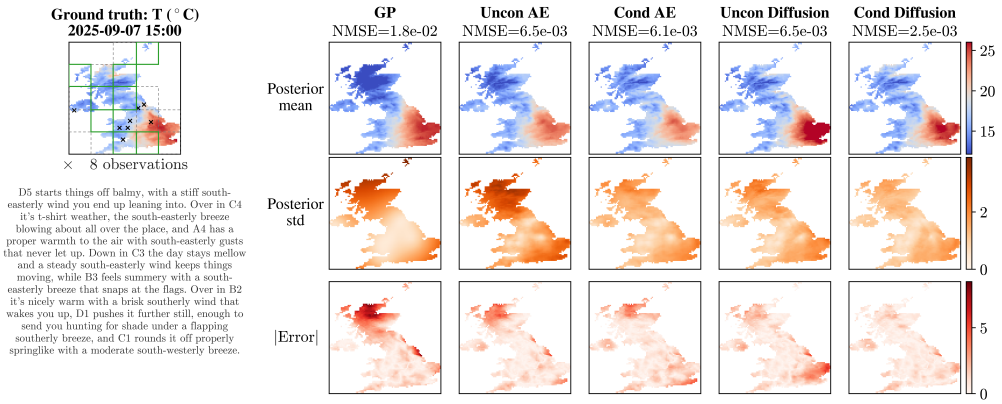

Figure 23: Posterior reconstructions of one UK Weather test system (T) when $n_{\text{obs}} = 8$.

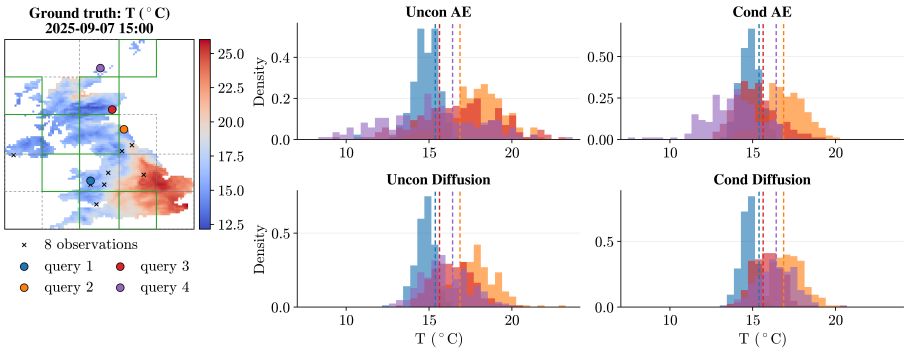

Figure 24: Posterior marginals of one UK weather test system (T) (same system as in Figure 23) at four unobserved query locations when $n_{\text{obs}} = 8$.

regions covered in A. Suppose the field $\tilde{u}$ has corresponding text s_A and $\tilde{u}'$ with s_B. Then

$$\text{score}_{\text{field}}(\tilde{u}, \tilde{u}') = \exp\left(-\frac{d^2}{\beta}\right), \tag{22a}$$

$$d^2 = \tfrac{1}{2}\left(d_A^2 + d_B^2\right), \tag{22b}$$

$$d_A^2 = \left\|\tilde{u}|_A - \tilde{u}'|_A\right\|_2^2, \tag{22c}$$

$$d_B^2 = \left\|\tilde{u}|_B - \tilde{u}'|_B\right\|_2^2, \tag{22d}$$

where d_A calculates the L_2 distance between two fields over regions described by s_A, and d_B over those s_B describes. $\beta = 1.2 \times 10^4$ is the bandwidth. 2 epochs of fine-tuning `all-MiniLM-L6-v2` are performed on $80,000$ training pairs with learning rate 2×10^{-5}.

Train. Supervised baselines use a U-Net with 96 base channels, **Cond Supervised** is trained for $1k$ epochs, while **Uncon Supervised** for $1.5k$ epochs, with a learning rate of 10^{-3} and batch size of 32; AE uses CNN-based autoencoder with 96 base channels, latent dimension is 64, **Cond AE** is trained for $3k$ epochs and **Uncon AE** for $3.5k$, with a learning rate of 10^{-3} and batch size 200; the score network is a U-Net with 96 base channels and a 256-dimensional conditioning vector, **Cond Diffusion** is trained for $1.5k$ epochs and **Uncon Diffusion** for $2.5k$, with learning rate of 2×10^{-4} and batch size of 32.

Inference. At inference, all three channels are observed at a common set of land locations, with $\sigma_y = \{0.5, 0.3, 0.3\}$ for $\{T, u, v\}$. More observations are obtained in regions not covered by text.

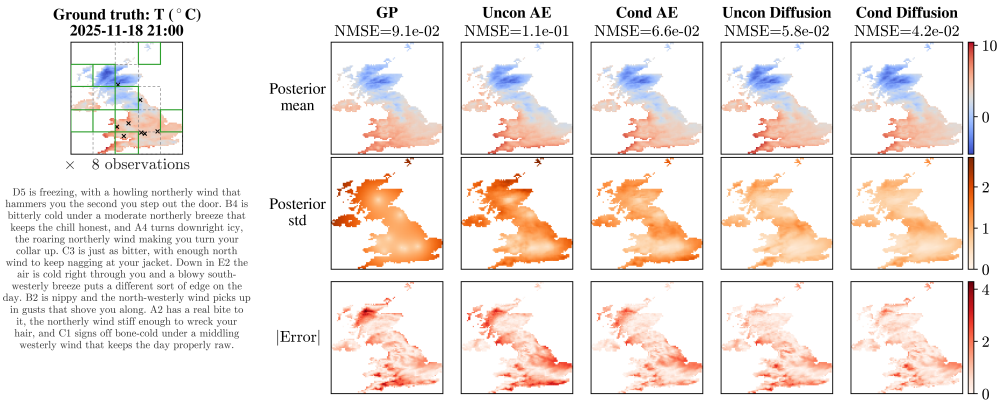

Figure 25: Posterior reconstructions of one UK Weather test system (T) when $n_{\text{obs}} = 8$.

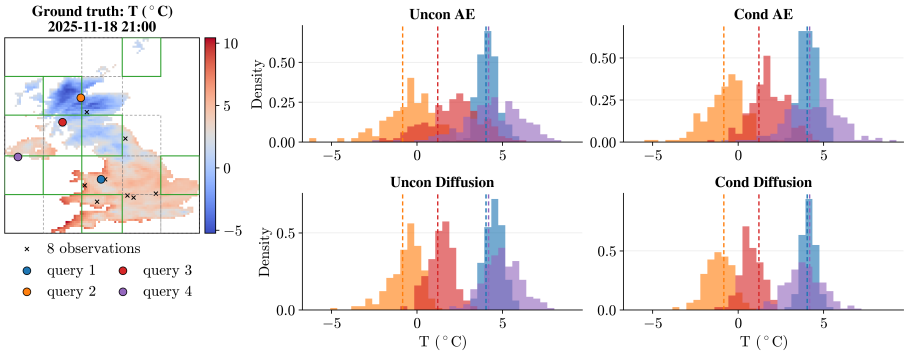

Figure 26: Posterior marginals of one UK weather system (T) (same system as in Figure 25) at four unobserved query locations when $n_{\text{obs}} = 8$.

For the **GP** baseline we use the exponential kernel, i.e. Matérn with $\nu = 1/2$,

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{\|x - x'\|_2}{\ell}\right), \tag{23}$$

where σ_f and ℓ are hyperparameters learned separately for three channels. AE posteriors are sampled using a single NUTS chain with $M_o = 100$ warmup steps and $M = 200$ retained draws. The maximum tree depth is 10 for both **Cond AE** and **Uncon AE**. Diffusion posteriors are drawn by integrating the guided reverse SDE over $L = 1000$ uniform steps and the guidance clip $c = 1.0$.

G.3.3 Additional Results

Table 12 provides full inference results. Figure 22 shows distribution of coverages, which are again better for autoencoders, with medians closer to nominal (68.3% and 94.4% for **Uncon AE**, 66.2% and 93.0% for **Cond AE**), whereas diffusion models yield systematically low coverages at both levels ($\sim$ 56% and $\sim$ 87%), indicating that inference is overconfident. Figures 21, 23, 24, 25, 26 show posterior reconstructions and posterior histograms for several systems. Figure 27 with Table 13 compares NMSE and text gain across various n_{obs} as before. The gain is smaller than in the other two PDE experiments, but the median stays above $1\times$ throughout, roughly $1.1\times$ for the autoencoder at $n_{\text{obs}} = 14$, so the text still contributes when the observations are dense. Figure 28 separates text gain by region, which is larger inside the regions the text describes (solid) than outside them (dashed).

To test whether supervised models are constrained to the observation process they were trained on, we further design an ablation study which uses a different observation operator at inference. In our

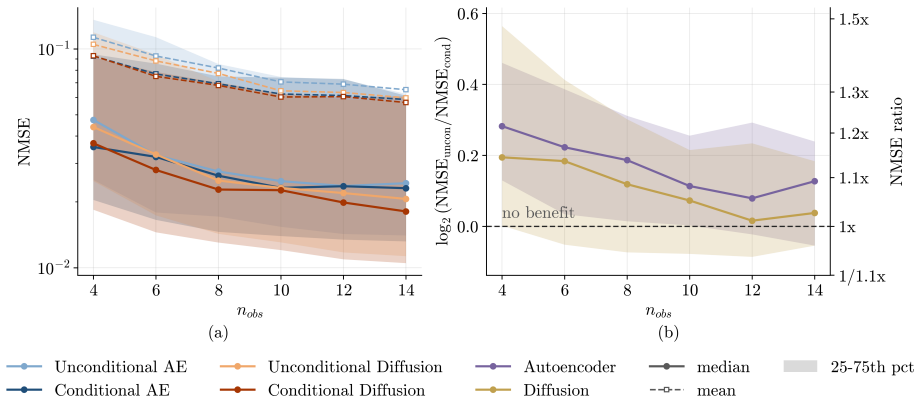

Figure 27: **NMSE performance at different $\mathbf{n_{obs}}$ for UK Weather data.** (a) NMSE across 100 test systems: median (solid), mean (dashed), and 25th–75th percentile bands, see Table 13 for exact values. (b) Per-system relative gain from text conditioning, $\log_2(\mathrm{NMSE}_{\mathrm{uncon}}/\mathrm{NMSE}_{\mathrm{cond}})$, with median and interquartile bands.

Table 13: Quartiles of per-system NMSE across 100 UK weather slices.

	$n_{\mathrm{obs}}=4$			$n_{\mathrm{obs}}=6$			$n_{\mathrm{obs}}=8$			$n_{\mathrm{obs}}=10$			$n_{\mathrm{obs}}=12$			$n_{\mathrm{obs}}=14$		
Model NMSE ($\times 10^{-2}$)	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}	Q_{25}	Median	Q_{75}
Uncon AE	2.52	4.74	13.58	1.79	3.27	11.32	1.71	2.75	8.53	1.54	2.49	7.45	1.43	2.36	7.25	1.41	2.43	6.23
Cond AE	2.04	3.56	9.45	1.66	3.21	8.57	1.46	2.63	7.49	1.40	2.32	7.40	1.34	2.36	7.29	1.32	2.31	6.11
Uncon Diffusion	2.49	4.39	11.91	1.74	3.29	9.42	1.43	2.50	7.09	1.31	2.32	6.20	1.18	2.19	5.88	1.13	2.06	5.77
Cond Diffusion	1.84	3.70	9.15	1.45	2.80	7.85	1.30	2.28	6.99	1.20	2.26	6.25	1.09	1.99	6.03	1.05	1.81	6.07

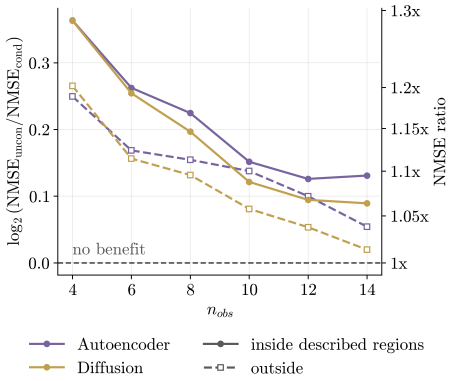

Figure 28: **Mean of per-system text gain from conditioning for UK weather test set**. The figure plots the text gain $\log_2(\mathrm{NMSE}_{\mathrm{uncon}}/\mathrm{NMSE}_{\mathrm{cond}})$, averaged over 100 test systems, inside the text-described regions (solid) and outside them (dashed).

main experiments, three channels $\{T, u, v\}$ share the same observation locations in both training and inference, which vary from system to system but are common to all channels. In the ablation, T is observed at 8 locations and u, v at 8 different ones, so the number of observations is unchanged and only the observation operator differs. Table 14 reports results over 20 test fields, where both conditional generative models overtake the supervised baselines.

Table 14: **Ablation on a different observation operator of the UK weather experiment.** Temperature is measured at 8 locations and the two wind components at 8 different locations, which is a sensor layout the supervised models were not trained on. Results are over 20 test systems.

Model	NMSE ($\times 10^{-2}$)	$\%C_{68}$	$\%C_{95}$
Uncon Supervised	5.95 ± 5.31	–	–
Cond Supervised	5.11 ± 4.36	–	–
Uncon AE	5.22 ± 4.86	68.9 (4.6)	94.0 (2.3)
Cond AE	4.41 ± 4.01	65.9 (5.6)	92.3 (3.6)
Uncon Diffusion	5.04 ± 4.72	56.7 (11.6)	86.3 (9.0)
Cond Diffusion	4.20 ± 3.70	57.2 (11.5)	86.7 (8.5)