OBSERVATION-ALIGNED MASK PRIORS FOR LEARNING PHYSICAL FIELDS FROM AUTHENTIC OCCLUSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning physical fields directly from incomplete observations is challenging because authentic occlusions are structured, sample-dependent, and often missing not at random, whereas existing methods typically rely on heuristic masking rules or predefined mask distributions. We propose Observation-Aligned Mask Priors, a framework that learns the distribution of authentic observation masks and uses it to construct context-query partitions for training from incomplete data. Specifically, we pretrain a Bayesian Flow Network (BFN) on binary observation masks to capture real occlusion topologies, then guide BFN sampling with a globally normalized cross-entropy objective to generate sample-specific masks aligned with each sparse observation. The intersection between the guided mask and the observed mask defines the context, and the remaining observed entries become query targets for a diffusion-based reconstruction model. We show that this intersection-based partitioning gives every valid observed dimension a strictly positive probability of being queried, preventing zero-query dead zones and local generative collapse. Experiments on three real-world oceanographic datasets with authentic satellite occlusions, across resolutions up to 256×256, show consistent improvements over strong diffusion baselines in MSE and PSNR. These results demonstrate that learning mask priors from authentic occlusions is an effective alternative to heuristic masking for learning from incomplete physical observations without access to fully observed fields.

1 INTRODUCTION

Learning physical fields from observational data is a central goal of scientific machine learning, with broad impact on ocean monitoring (Barth et al., 2024; Martin et al., 2025), weather analysis (Bi et al., 2023; Lam et al., 2023), and environmental forecasting (Hsu et al., 2025; Ashfaq et al., 2025). In these applications, however, the underlying physical field is almost never observed on a complete spatial grid (Chen et al., 2025). Measurements arrive through authentic occlusions caused by cloud cover, sensor swath gaps, orbital trajectories, and geographic boundaries (Letraon et al., 2017), yielding observation masks that are structured, sample-dependent, and often missing not at random (MNAR). These masks are not generic corruption that can be approximated by random dropout; they are part of the sensing process itself. A central challenge is therefore to learn physical fields directly from incomplete observations, without requiring access to fully observed fields, while respecting the topology of real-world occlusions.

Existing Bottleneck. Recently, a promising direction is to train directly on incomplete data through *context-query partitioning* (Zhou et al., 2025). Instead of supervising a model with complete ground truth, one splits the observed region into a visible context and a withheld query region, and trains the model to reconstruct the query from the context. This removes the unrealistic requirement of fully observed supervision, but it introduces a new bottleneck: the success of learning now depends critically on how the context-query partition is constructed. In fact, the seminal study (Zhou et al., 2025) shows that a spatial dimension can only be learned if it has a strictly positive probability of being assigned to the query region; otherwise its expected gradient vanishes and the model may output arbitrary values there. Existing approaches typically instantiate this partition using heuristic masking rules (Zhou et al., 2025; Daras et al., 2023; Majid et al., 2026), such as independent pixel dropout, regular block masking, or hand-designed saliency heuristics. While convenient, these

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.

strategies implicitly assume a predefined mask distribution and are poorly matched to authentic physical occlusions, whose geometry is shaped by clouds, coastlines, and sensing constraints. When missingness is structured and potentially MNAR, the mismatch between heuristic masking and real observational topology becomes a fundamental obstacle.

Our Solution. The key idea is to learn, rather than assume, the mask distribution underlying authentic observations. We propose **Observation-Aligned Mask Priors**, which learns a generative prior over real observation masks and uses it to construct sample specific context–query partitions (see Fig. 1). Specifically, we pretrain a Bayesian Flow Network (BFN) (Graves et al., 2025) on binary observation masks to capture authentic occlusion topology, and guide its sampling toward the current sparse observation using a globally normalized cross-entropy objective. The intersection between the guided mask and the observation mask defines the context, while the remaining observed entries form the query region. This yields data-driven partitions that preserve realistic occlusion structure while adapting to each individual sample.

Theoretical & Empirical Value. Theoretically, we show that intersection-based partitioning gives every valid observed dimension a strictly positive probability of being queried. We first establish this property under independent sampling from the learned prior and then show that it remains valid under observation-aligned guided partitioning, thereby avoiding zero-query dead zones that can lead to local generative collapse and structural information deficiency. Empirically, across three real-world oceanographic datasets (Letraon et al., 2017) with genuine satellite occlusions—Black Sea CHL, Baltic Sea NANO, and Global Ocean SSS—and across resolutions up to 256×256, our method consistently outperforms strong diffusion-based baselines in MSE and PSNR.

More related works are discussed in Appendix A. We present our contributions as follows:

- **From heuristic masks to learned occlusion priors.** We introduce *Observation-Aligned Mask Priors*, a new framework that replaces handcrafted context-query partitioning rules with a learned generative prior over authentic observation masks, and aligns this prior to each input sample through guided mask generation.
- **Strict positivity, not just empirical coverage.** We provide theoretical guarantees showing that our intersection-based partitioning strategy assigns every valid observed dimension a strictly positive query probability, both under mask-prior sampling and under observation-aligned guidance, thereby preventing zero-query dead zones and local collapse.
- **Real gains under authentic occlusions.** We validate the framework on *three challenging real-world oceanographic datasets* with genuine structured missingness and show consistent improvements over diffusion baselines across multiple spatial resolutions, without ever requiring fully observed training fields.

2 PRELIMINARIES

In realistic physical science applications, training datasets typically only consist of partial observations $\boldsymbol{u}_{\text{obs}} = \boldsymbol{M} \odot \boldsymbol{u}_0$, where $\boldsymbol{M}$ is a binary observation mask. To learn complete physical fields directly from incomplete data, previous work (Zhou et al., 2025) introduced a context-query partitioning framework.

Learning Fields via Context-Query Partitioning. During training, each incomplete sample is strategically partitioned into an observable context mask $\boldsymbol{M}_{\text{ctx}} \subset \boldsymbol{M}$ and a non-request mask $\boldsymbol{M}_{\text{qry}} \subseteq \boldsymbol{M}$. The imputation model $\boldsymbol{u}_{\boldsymbol{\phi}}$ is trained to reconstruct the query region conditioned solely on the context region:

$$\mathcal{L}(\boldsymbol{u}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}, \boldsymbol{M}_{\text{qry}}) = ||\boldsymbol{M}_{\text{qry}} \odot (\boldsymbol{u}_{\boldsymbol{\phi}}(\boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{u}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}) - \boldsymbol{u}_{\text{obs}})||^2. \tag{1}$$

The context and query regions are constructed to be disjoint so that learning is restricted to values withheld from the model input, making it easier for the model to learn. Since context values are already observed and can be directly copied during reconstruction, the objective focuses supervision on the query region, encouraging the model to infer missing structure rather than rely on local pixel memorization.

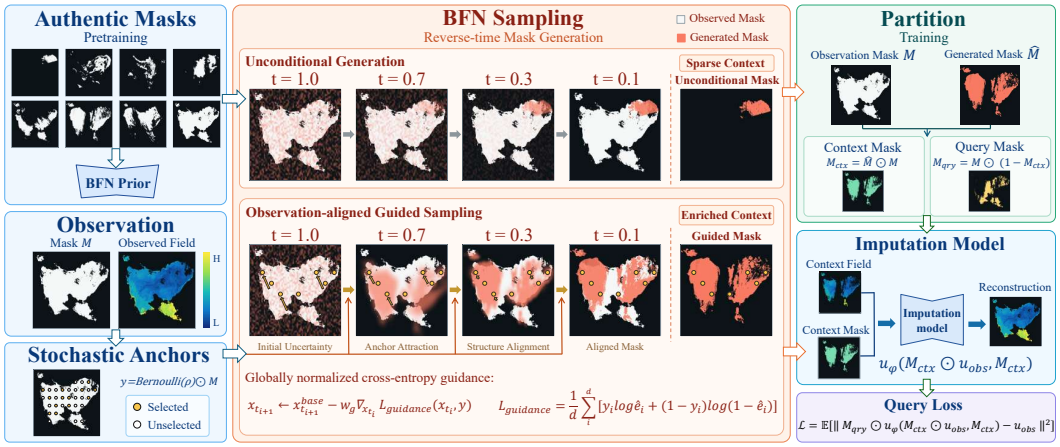


Figure 1: **Overview of our framework.** *Left*: A Bayesian Flow Network (BFN) is first pretrained on authentic observation masks to learn a prior over real occlusion topologies. *Middle*: Given an incomplete sample, stochastic anchors are drawn from the observed mask and used to guide BFN sampling through a globally normalized cross-entropy objective, yielding an observation-aligned generated mask. *Right*: The intersection between the generated mask and the observed mask defines the context region, and the remaining observed entries form the query region.The imputation model is then trained to reconstruct the query from the context.

Theoretical Foundation of Global Prediction. The success of this framework fundamentally depends on the context-query splitting strategy. Theoretical analysis establishes that minimizing Eq.(1) leads to an optimal prediction $(\boldsymbol{u}_\phi)_i$ for a specific feature dimension i given by:

$$\left(\boldsymbol{u}_{\boldsymbol{\phi}}\left(\boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{u}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}\right)\right)_i = \begin{cases} \mathbb{E}\left[\left(\boldsymbol{u}_0\right)_i \mid \boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{u}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}\right], & P\left(\left(\boldsymbol{M}_{\text{qry}}\right)_i = 1 \mid \boldsymbol{M}_{\text{ctx}}\right) > 0 \\ \text{an arbitrary value}, & P\left(\left(\boldsymbol{M}_{\text{qry}}\right)_i = 1 \mid \boldsymbol{M}_{\text{ctx}}\right) = 0 \end{cases} \tag{2}$$

This reveals that the model can only learn a meaningful conditional expectation for a specific spatial dimension if its probability of being queried is strictly positive. This condition ensures non-vanishing expected gradients, as detailed in Appendix D.

Bottleneck of Heuristic Design. To ensure that the model learns a complete physical field, every originally unobserved dimension must have a strictly positive probability of being queried during training. Although previous work (Zhou et al., 2025; Majid et al., 2026) identified this necessity, it relied on manually designing $\boldsymbol{M}_{\text{ctx}}$ selection strategies tailored to each individual observation pattern, such as independent pixel dropout and block-wise occlusion. Such customized methods inevitably introduce distribution-specific hyperparameters and fail to integrate the design of various mask topologies into a unified framework. More importantly, relying on simplistic handcrafted rules makes it mathematically arduous to universally guarantee this strict positivity across complex real-world spatial dependencies. Therefore, overcoming the bottleneck requires moving beyond heuristic splitting rules to establish a unified mechanism capable of providing robust theoretical guarantee for any valid spatial topology.

3 Method

3.1 Overview: The Observation-Aligned Imputation Framework

In this work, we propose a novel observation-aligned imputation framework to reconstruct complete physical fields from incomplete observations. Our method builds upon the context-query learning paradigm, where a prediction model learns to infer missing query regions $\boldsymbol{M}_{\text{qry}}$ conditioned on observed context regions $\boldsymbol{M}_{\text{ctx}}$, but fundamentally transforms how this partitioning is constructed. Rather than relying on simplistic, random splitting strategies that fail to capture the structured nature of real-world missingness, our framework intelligently drives the data partitioning process using a generative prior.

Specifically, our framework pipeline consists of three interconnected modules, which will be detailed in the following subsections:

- **Mask Distribution Modeling (Sec. 3.3):** We first introduce a Bayesian Flow Network (BFN) (Graves et al., 2025) to explicitly model the authentic spatial distribution of the discrete binary observation masks.
- **Observation-Aligned Mask Generation (Sec. 3.4):** To align the generated masks with the actual sparse observations of a given sample, we propose a classifier guidance mechanism during the BFN sampling phase. This ensures the generated context masks respect both the learned physical occlusion priors and the available data points.
- **Guided Context-Query Training (Sec. 3.5):** Finally, these intelligently generated masks are used to construct $\boldsymbol{M}_{\text{ctx}}$ and $\boldsymbol{M}_{\text{qry}}$, dynamically dictating the training of the prediction model.

By replacing heuristic random masking with this generative, observation-aligned selection mechanism, our framework actively concentrates the prediction model's learning capacity on the most critical missing regions, significantly bridging the gap between unstructured data matching and structured physical realities.

3.2 Context-Query Partitioning via Generative Priors

As established in Sec. 2, relying on heuristic masking strategies introduces severe limitations, as handcrafted rules cannot encapsulate the complex, structured nature of authentic occlusions. To establish a universally robust framework, we shift paradigm from manually partitioning observations to explicitly modeling the realistic underlying mask distribution $P(\boldsymbol{M})$ using a data-driven generative prior. By driving the context-query selection through this learned prior, we establish a unified mechanism that guarantees non-zero query probabilities for any valid spatial topology. Specifically, we theoretically establish that intersecting two independent samples from $P(\boldsymbol{M})$ naturally satisfies this strict positivity requirement, thus eliminating the need for heuristic designs.

Theorem 1 (Strict Positivity of Query Probabilities via Mask Intersection)**.** *Let $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \{0,1\}^d$ be independently sampled binary masks from the same underlying distribution $p(\boldsymbol{M})$. Define the context mask as $\boldsymbol{M}_{\text{ctx}} = \boldsymbol{M}_1 \odot \boldsymbol{M}_2$ and the query mask as $\boldsymbol{M}_{\text{qry}} = \boldsymbol{M}_1 \odot (\mathbf{1} - \boldsymbol{M}_{\text{ctx}})$. Under Assumption 1, for any valid context mask $\boldsymbol{m}$ and any unobserved dimension i (where $\boldsymbol{m}_i = 0$), the conditional probability of dimension i being selected as a query is strictly positive:*

$$P((\boldsymbol{M}_{\text{qry}})_i = 1 \mid \boldsymbol{M}_{\text{ctx}} = \boldsymbol{m}) > 0. \tag{3}$$

The proof can be found in Appendix C.3. Theorem 1 provides a rigorous mathematical guarantee for our partitioning strategy. However, realizing this guarantee in practice introduces a new prerequisite: we must successfully model and sample from the complex, discrete distribution $p(\boldsymbol{M})$. This theoretical requirement directly motivates the need for a powerful discrete generative prior.

3.3 Modeling Mask Distributions via Bayesian Flow Networks (BFN)

To fulfill the prerequisite established by Theorem 1, we require a generative model capable of capturing the intricate spatial dependencies inherent in authentic missing patterns. Our mask generation module is fundamentally built upon Bayesian Flow Networks (BFN) (Graves et al., 2025), a framework naturally suited for modeling discrete binary variables. To facilitate gradient-based interventions for observation-aligned conditioning (Sec. 3.4), we use a BFN-inspired continuous latent formulation with a discrete data prediction objective. This formulation is motivated by the connection between BFN regression and diffusion score matching (Xue et al., 2024). Exact equivalence to the original BFN further requires matching the latent parameterization and noise schedule (Appendix B).

Let $C = 2$ denote the number of mask classes, d the number of pixels, and $K > 0$ an independent logit scaling factor. For one pixel, $c \in \{1, \dots, C\}$ has one-hot encoding $\boldsymbol{e}_c \in \mathbb{R}^C$ and continuous target $\boldsymbol{x}_0 = K\boldsymbol{e}_c$. Scaling only affects continuous embedding, while the classification target remains the exact one-hot vector $\boldsymbol{e}_c$. We corrupt the continuous embedding according to

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{4}$$

For a full mask, the latent stacks d blocks of C logits in $\mathbb{R}^{d\times C}$, with independent standard Gaussian noise on all coordinates. Softmax is applied separately over the C classes at each pixel, and the squared loss below is summed over pixels. Under this Gaussian observation model, adding a common scalar to the logits within any pixel leaves the joint class posterior unchanged. Consequently, the optimal data predictor $\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$ is shift invariant and can be represented using only $\mathrm{softmax}(\boldsymbol{x}_t)$, without assuming independent mask pixels (Appendix C.1). The full score and noise predictors are not shift invariant: they retain an explicit linear dependence on $\boldsymbol{x}_t$. We therefore parameterize the data predictor as $K\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \mathrm{softmax}(\boldsymbol{x}_t))$ and obtain a reweighted noise matching objective through the corresponding noise predictor (Appendix C.2) (Zheng et al., 2023):

$$\mathcal{L}_{\text{DM-discrete}} = \mathbb{E}_{t,c,\boldsymbol{\epsilon}}\left[w(t)\|\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \mathrm{softmax}(\boldsymbol{x}_t)) - \boldsymbol{e}_c\|^2\right], \tag{5}$$

where $w(t)$ is a time-dependent weight. Exact equality with noise matching weighted by $w_{\mathrm{NM}}(t)$ requires $w(t) = w_{\mathrm{NM}}(t)\alpha_t^2 K^2/\sigma_t^2$ at nondegenerate training times. Tweedie's formula motivates the score estimate

$$\boldsymbol{s}_{\boldsymbol{\theta}}(t, \boldsymbol{x}_t) = \frac{\alpha_t K\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \mathrm{softmax}(\boldsymbol{x}_t)) - \boldsymbol{x}_t}{\sigma_t^2}. \tag{6}$$

This equals the true score when the network predicts the exact class posterior. We use this estimate in the probability flow ODE, with the Gaussian terminal approximation $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d\times C})$. At each pixel i, we decode $\hat{c}_i = \arg\max_{j\in\{1,\ldots,C\}}(\boldsymbol{x}_0)_{i,j}$ and set $\hat{M}_i = \mathbb{1}[\hat{c}_i = 2]$, where classes 1 and 2 denote missing and observed entries, respectively.

Limitation of Unconditional Intersection. While sampling unconditionally from the learned marginal distribution $p(\boldsymbol{M})$ strictly satisfies the positivity guarantee of Theorem 1, directly applying this unconditional intersection presents a critical practical bottleneck. Because unconditional generation operates entirely independent of the specific valid observations of a given data sample, the spatial overlap between the authentic observation mask and the generated mask can be exceedingly small. Such a minimal intersection produces a highly sparse context region $\boldsymbol{M}_{\text{ctx}}$, which inherently exacerbates the structural information gap identified in prior work (Zhou et al., 2025), leaving the imputation model without sufficient spatial anchors to reconstruct complex physical fields. To overcome this limitation and ensure a sufficiently informative context partition, it is imperative to transition from unconditional sampling to observation-aligned conditional generation. In the following section, we introduce our specific observation-aligned mechanism to achieve this conditional generation.

3.4 Observation-Aligned Context Mask Generation

Before detailing the practical guidance mechanism, a fundamental theoretical question must be addressed: does restricting the mask generation to align with a specific observation break the strict positivity requirement established in Theorem 1? Fortunately, Theorem 2 provides the theoretical assurance that under a well-defined intersection constraint, conditional generation strictly preserves this property.

Theorem 2 (Strict Positivity of Query Probabilities under Ratio-Guided Partitioning)**.** *Let $\boldsymbol{M} \in \{0,1\}^d$ be a fixed observation mask. Let the conditionally generated mask $\hat{\boldsymbol{M}} \in \{0,1\}^d$ be sampled from the guided distribution $P(\hat{\boldsymbol{M}} \mid C_k)$ satisfying the intersection constraint C_k as defined in Assumption 2. Define the context mask as $\boldsymbol{M}_{ctx} = \hat{\boldsymbol{M}} \odot \boldsymbol{M}$ and the query mask as $\boldsymbol{M}_{qry} = \boldsymbol{M} \odot (1 - \boldsymbol{M}_{ctx})$. Under Assumption 2, for any valid observed dimension i (where $\boldsymbol{M}_i = 1$), the marginal probability of dimension i being assigned as a query region is strictly positive: $P((\boldsymbol{M}_{qry})_i = 1 \mid C_k) > 0$.* [1]

The proof can be found in Appendix C.4. Empowered by this theoretical guarantee, we formulate a practical classifier guidance mechanism during the BFN sampling phase to dynamically anchor the generation process to the valid observations. Our method consists of three key steps:

[1]The condition shifts from $\boldsymbol{M}_{\text{ctx}}$ (Theorem 1) to C_k to reflect the algorithmic transition from random partitioning to observation-anchored generation. Guaranteeing the marginal probability $P((\boldsymbol{M}_{\text{qry}})_i = 1 \mid C_k) > 0$ is mathematically sufficient to ensure a strictly positive query frequency ($p_i > 0$ in Eq.(41)) across epochs. See Remark 1 in Appendix for a detailed discussion.

Stochastic anchor construction. First, we construct a stochastic binary anchor $\boldsymbol{y} \in \{0,1\}^d$ by applying Bernoulli sampling to the valid observation points (where $\boldsymbol{M}_i = 1$, representing the i-th pixel of $\boldsymbol{M}$) with a predefined retention ratio $\rho \in (0,1)$:

$$\boldsymbol{y}_i = \mathbb{1}[\boldsymbol{r}_i < \rho] \cdot \boldsymbol{M}_i, \quad \boldsymbol{r}_i \sim \text{Uniform}(0,1). \tag{7}$$

If we directly guide the BFN using all observation points, the generation tends to collapse into a single deterministic output (see Appendix E.9 for details). This lacks the diversity needed for robust downstream training. By randomly dropping a fraction of these observation points, we inject essential stochasticity. It forces the BFN to dynamically connect these scattered anchor points using its learned physical priors. Consequently, the model can generate diverse, physically plausible sub-masks for the exact same observation field.

Globally normalized guidance objective. Let $\boldsymbol{x}_t \in \mathbb{R}^{d\times C}$ be the stacked latent state at a given BFN sampling time step t, with softmax applied over the $C = 2$ classes at each pixel. Denote the predicted observed-class probabilities by $\hat{\boldsymbol{e}} \in [0,1]^d$, where $\hat{e}_i = [\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \text{softmax}(\boldsymbol{x}_t))]_{i,2}$. We compute a cross-entropy loss over all d pixels between these probabilities and the anchor $\boldsymbol{y}$:

$$\mathcal{L}_{\text{guidance}}(\boldsymbol{x}_t, \boldsymbol{y}) = -\frac{1}{d}\sum_{i=1}^{d} \left[\boldsymbol{y}_i \log \hat{\boldsymbol{e}}_i + (1-\boldsymbol{y}_i)\log(1-\hat{\boldsymbol{e}}_i)\right]. \tag{8}$$

Global normalization ensures that the gradient magnitude of the guidance signal remains highly stable across samples with drastically different sparsity levels, preventing gradient explosion or vanishing.

Latent space intervention. Finally, we inject this guidance signal during the reverse sampling process of the BFN. Let t_i and t_{i+1} denote consecutive discrete time steps in the reverse time schedule. Denoting $x^{\text{base}}_{t_{i+1}}$ as the unconditional latent state update computed from $\boldsymbol{x}_{t_i}$, we intervene by applying a gradient descent step using our guidance loss:

$$\boldsymbol{x}_{t_{i+1}} \leftarrow \boldsymbol{x}^{\text{base}}_{t_{i+1}} - w_g \nabla_{\boldsymbol{x}_{t_i}} \mathcal{L}_{\text{guidance}}(\boldsymbol{x}_{t_i}, \boldsymbol{y}), \tag{9}$$

where w_g is the guidance scale that acts as a balancing knob. It guarantees that the final generated mask respects the learned occlusion priors while strictly anchoring to the selected sparse observations.

3.5 Imputation Training with Guided Partitioning

Strategic partitioning via generative intersection. Building upon the theoretical foundations established in Theorem 1 and the validity guarantees of Theorem 2, we now formalize our final context-query partitioning strategy. To learn from incomplete observations $\boldsymbol{u}_{\text{obs}} = \boldsymbol{M} \odot \boldsymbol{u}_0$, we leverage the observation-aligned binary mask $\hat{\boldsymbol{M}} \in \{0,1\}^d$ generated by our guided BFN to execute the intersection-based partition. Specifically, we define context region $\boldsymbol{M}_{\text{ctx}}$ as the intersection between actual observed fields $\boldsymbol{M}$ and BFN conditional generated fields $\hat{\boldsymbol{M}}$. The query region $\boldsymbol{M}_{\text{qry}}$ is the remaining portion after excluding the observable area of the context region. Specifically,

$$\boldsymbol{M}_{\text{ctx}} = \hat{\boldsymbol{M}} \odot \boldsymbol{M}, \quad \boldsymbol{M}_{\text{qry}} = \boldsymbol{M} \odot (\boldsymbol{1} - \boldsymbol{M}_{\text{ctx}}). \tag{10}$$

Training objective and information withholding. During training of the main imputation model, the network $\boldsymbol{u}_{\boldsymbol{\phi}}$ is fed exclusively with the context elements $\boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{u}_{\text{obs}}$ and the context mask $\boldsymbol{M}_{\text{ctx}}$. The optimization objective is to minimize the Mean Squared Error (MSE) between the network's prediction and the clean observations, but *only* within the query regions:

$$\mathcal{L} = \mathbb{E}\left[\|\boldsymbol{M}_{\text{qry}} \odot (\boldsymbol{u}_{\boldsymbol{\phi}}(\boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{u}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}) - \boldsymbol{u}_{\text{obs}})\|^2\right]. \tag{11}$$

By deliberately withholding the query information $\boldsymbol{M}_{\text{qry}}$ during the forward pass, we force the model to abandon local pixel memorization and instead learn the spatial correlations between different physical regions. The model must "fill in the blanks" by understanding the underlying spatial structure of the physical field. Because the query region is selected by the conditional BFN to match realistic physical occlusions, the model focuses its learning capacity on the most challenging and meaningful spatial transitions.

Algorithm 1 Guided BFN Mask Generation

Require: $\boldsymbol{M}$, pre-trained BFN $\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}$, w_g, n, ρ
Ensure: generated mask $\hat{\boldsymbol{M}}$
anchor $\boldsymbol{y}_i = \mathbb{1}[\boldsymbol{r}_i < \rho] \cdot \boldsymbol{M}_i$, $\boldsymbol{r}_i \sim \text{Uni.}(0,1)$
time schedule $\{t_i\}_{i=0}^{n}$ $(t_0 = 1, t_n = 0)$
sample $\boldsymbol{x}_{t_0} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d\times C})$
for $i = 0, \ldots, n-1$ **do**
  $\hat{\boldsymbol{x}}_0 \leftarrow K\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t_i, \text{softmax}(\boldsymbol{x}_{t_i}))$
  $\boldsymbol{x}^{\text{base}}_{t_{i+1}} \leftarrow \alpha_{t_{i+1}}\hat{\boldsymbol{x}}_0 + \frac{\sigma_{t_{i+1}}}{\sigma_{t_i}}(\boldsymbol{x}_{t_i} - \alpha_{t_i}\hat{\boldsymbol{x}}_0)$
  $\boldsymbol{x}_{t_{i+1}} \leftarrow \boldsymbol{x}^{\text{base}}_{t_{i+1}} - w_g \nabla_{\boldsymbol{x}_{t_i}} \mathcal{L}_{\text{guid}}(\boldsymbol{x}_{t_i}, y)$
end for
$\hat{c}_i \leftarrow \arg\max_{j=1,2}(\boldsymbol{x}_0)_{i,j}$ for each pixel i
$\hat{\boldsymbol{M}}_i \leftarrow \mathbb{1}[\hat{c}_i = 2]$ for each pixel i
return $\hat{\boldsymbol{M}}$

Algorithm 2 Train Imputation with Partition

Require: dataset $\mathcal{D} = \{(\boldsymbol{u}^{(i)}_{\text{obs}}, \boldsymbol{M}^{(i)})\}_{i=1}^{N}$
Ensure: imputation model parameters $\boldsymbol{\phi}$
while not converged **do**
  for each batch $(\boldsymbol{u}_{\text{obs}}, \boldsymbol{M}) \in \mathcal{D}$ **do**
    $\hat{\boldsymbol{M}} \leftarrow$ Algorithm 1$(\boldsymbol{u}_{\text{obs}}, \boldsymbol{M}, \ldots)$
    $\boldsymbol{M}_{\text{ctx}} \leftarrow \hat{\boldsymbol{M}} \odot \boldsymbol{M}$
    $\boldsymbol{M}_{\text{qry}} \leftarrow \boldsymbol{M} \odot (1 - \boldsymbol{M}_{\text{ctx}})$
    $\hat{\boldsymbol{u}} \leftarrow \boldsymbol{u}_{\boldsymbol{\phi}}(\boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{u}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}})$
    $\mathcal{L} \leftarrow ||\boldsymbol{M}_{\text{qry}} \odot (\hat{\boldsymbol{u}} - \boldsymbol{u}_{\text{obs}})||^2$
    update $\boldsymbol{\phi}$ using gradient descent on $\mathcal{L}$
  end for
end while

Convergence to conditional expectation. Under this prediction-driven training objective, the network naturally learns to approximate the conditional expectation $\mathbb{E}[\boldsymbol{u}_0 | \boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{u}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}]$. Crucially, as mathematically guaranteed by Theorem 2, our guided partitioning inherently satisfies the strict positivity requirement for every valid spatial dimension. By eliminating the heuristic bottleneck identified in Sec. 2, this ensures that no part of the observation space is ignored. Over many iterations, the model develops a robust ability to reconstruct missing spatial domains. Ultimately, this enables the framework to transfer its reconstructive capabilities from synthetically masked training subsets to the inherently missing regions of real-world data, successfully bridging the gap without ever requiring access to any fully complete observation.

4 Experiments

4.1 Baselines

To evaluate the effectiveness of our framework, we compare it against the following established diffusion-based reconstruction approaches. **DINDiff** (Barth et al., 2024) utilizes a denoising diffusion model with hierarchical masking to reconstruct incomplete satellite observations via ensemble averaging. **Ambient Diffusion** (Daras et al., 2023) reconstructs uncorrupted distributions directly from highly corrupted training data by incorporating known degradation processes into the forward pass. **MissDiff** (Ouyang et al., 2023) handles incomplete datasets by optimizing a unified diffusion objective that effectively marginalizes over unobserved variables. In addition to these evaluated methods, we also explored other baselines in Appendix E.2. For **our method**, we first pre-train a dataset-specific BFN model on the observation masks, freezing its parameters during the primary imputation training to maintain stable generative priors. Detailed settings are provided in Appendix E.3.

4.2 Datasets

We conduct comprehensive experiments on three real-world high-resolution datasets with authentic satellite occlusions (Letraon et al., 2017): **(i) Black Sea CHL** (Black Sea Chlorophyll-a Dataset) offers multi-sensor observations of the Black Sea at a 1-km resolution, measuring the chlorophyll-a concentration. **(ii) Baltic Sea NANO** (Baltic Sea Nano-Phytoplankton Dataset) provides the observations of Baltic Sea nano-phytoplankton concentration at a 1-km resolution. **(iii) Global Ocean SSS** (Global Sea Surface Salinity Dataset) measures global sea surface salinity from satellite at a 25-km resolution. Regarding the data format, we processed the datasets into partially observed gridded fields and observation mask data pairs. Detailed information is provided in Appendix E.1.

4.3 Experimental Results

Evaluation. We evaluate the model performance by overlaying another mask over the existing observations. Specifically, we extract another observation mask $\tilde{\boldsymbol{M}}$ from the remaining dataset. Overlay it onto the existing observations $\boldsymbol{M}_{\text{eval}}$, and use the intersection of $\tilde{\boldsymbol{M}} \odot \boldsymbol{M}_{\text{eval}}$ as the new observation mask $\boldsymbol{M}$. The model will then predict the remaining portion given $\boldsymbol{M}$ and $\boldsymbol{M} \odot \boldsymbol{u}_{\text{obs}}$.

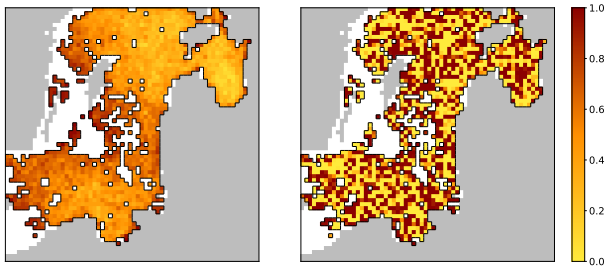
Table 1: Performance comparison on physical fields imputation tasks with the resolution of 64×64 across three real-world datasets.

Method	Black Sea CHL		Baltic Sea NANO		Global Ocean SSS	
	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑
AmbientDiff	0.228	17.717	0.365	15.466	0.286	18.022
DINDiff	0.306	15.355	0.322	15.788	0.408	16.373
MissDiff	0.606	13.575	0.579	13.672	0.413	16.644
Ours	**0.193**	**17.825**	**0.294**	**16.101**	**0.276**	**18.192**

The evaluation will be conducted in the region $\boldsymbol{M}_{\text{eval}} \odot (1 - \boldsymbol{M})$ where real values naturally exist as evaluation criteria. To assess reconstruction quality, we use three physically meaningful metrics: **MSE Loss** measures the mean squared error on masked observed regions using random real-mask overlays. **PSNR** assesses global signal fidelity.

Results. Tab. 1 presents the quantitative evaluation of our framework against established diffusion-based imputation baselines in three real-world oceanographic datasets. In addition to the experiments at 64×64 resolution, we also scaled up to 128×128 and 256×256 resolution (Tab. 5 and Tab. 6). Our method consistently achieves superior performance, yielding the lowest MSE Loss and the highest PSNR across all evaluated scenarios. While baselines such as AmbientDiff suffer from severe structural information gaps due to reliance on random spatial masking, and approaches like DINDiff and MissDiff struggle to maintain global fidelity under severe sparsity, our framework robustly reconstructs complex physical fields. By leveraging our proposed context-query partitioning strategy, the model efficiently concentrates its learning capacity on critical missing regions, successfully bridging deterministic observed anchors with highly coherent, physically realistic generated structures.

4.4 Uncertainty Quantification of Query Distribution



(a) Ours (b) Random Select

Figure 2: Query probability distributions of our method and naive random selection strategy.

Table 2: Comparison of different context-query partitioning strategies. Here the baseline "Unconditional" refers to BFN unconditional generation which reflects the result of Theorem 1.

Methods	MSE ↓	PSNR ↑
Pixel-level	0.263	17.10
Block-wise	0.251	16.43
Saliency-driven	0.317	16.28
Empirical	0.225	16.62
Unconditional	0.216	16.78
Ours	**0.193**	**17.83**

To validate that our observation-aligned guided partitioning strategy strictly preserves the positivity of query probabilities as established in Theorem 2, we design an empirical study to quantify the spatial distribution of query assignments.

Experiment design. We first train a binary classifier to learn to predict the probability of each pixel being assigned to the query region given only the context mask $\boldsymbol{M}_{\text{ctx}}$. Then we repeatedly sample from our conditionally guided Bayesian Flow Network (BFN) under the intersection constraint C_k (regulated by our retention ratio ρ). By generating an ensemble of context-query partitions for the same observation field, we compute the empirical query probability, denoted as $P((\boldsymbol{M}_{\text{qry}})_i = 1 | C_k)$, across the entire spatial domain.

Verification of strictly positive query probabilities. To explicitly demonstrate the effectiveness of our partitioning mechanism, we compare the mean $P(qry)$ spatial distribution of our method against a naive random selection baseline in Fig. 2. As shown, the naive strategy yields a highly fragmented

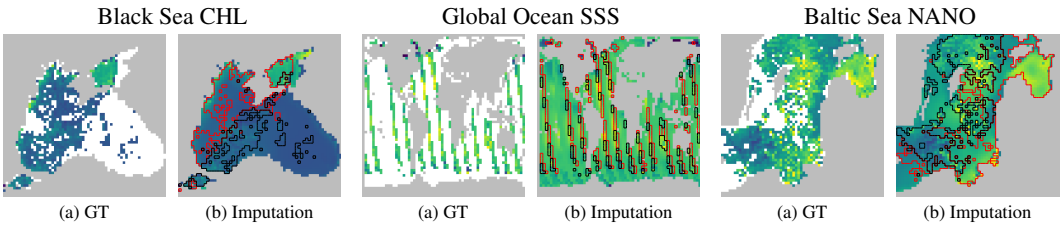


Figure 3: Visualization of data imputation results on Black Sea CHL, Global Ocean SSS, and Baltic Sea NANO with the resolution of 64×64 (grouped from left to right). For each dataset, the left panel shows the ground truth (GT) and the right panel displays our imputation result. The gray areas represent land, and the white areas represent the missing regions. The area outlined in black is the visible input region for the model, while the areas framed in red indicate the region used to evaluate the reconstruction results.

distribution plagued by "zero-query" regions (where $P(qry)$ approaches 0), which inevitably leads to localized generative collapse. In contrast, it demonstrates that the query probability distribution of our method is strictly positive throughout the spatial domain. This empirical evidence directly corroborates Theorem 2: by dynamically anchoring mask generation to the learned physical prior without perfectly memorizing the deterministic constraints, our framework mathematically and practically guarantees that no valid spatial dimension is systematically ignored during the partitioning process. An extended comparative analysis across additional datasets is provided in Appendix E.8. Extensive experiments show that the induced positive query exposure consistently maintains non-negligible magnitudes for each dimension in practice, rather than merely non-zero in expectation.

Heuristic reliability indicator. In Appendix D, we detail the relationship between query probability and gradient updates, illustrating that the $P(qry)$ distribution can provide a heuristic indication of the reconstruction reliability during inference.

4.5 Ablation Study

Context mask selection strategy comparison. Proper selection of $\boldsymbol{M}_{\text{ctx}}$ is crucial to prevent information gaps. Tab. 2 reveals that existing baselines face a critical trade-off: heuristic partitions (Pixel/Block-wise) lose physical coherence, non-adaptive methods ignore spatial topology, and saliency-driven approaches sacrifice large-scale structural anchors, worsening global PSNR. Our strategy resolves trade-offs via cross-entropy guidance, which adaptively optimizes context proportion and placement to balance global anchors with local structure for superior reconstruction (Appendix E.6). Additionally, we perform controlled comparisons to further isolate the contribution of the partitioning mechanism itself and examine the robustness across different backbones (Appendix E.5).

Sampling method comparison. To reconstruct complete ocean fields from sparse observations, the sampling strategy must effectively bridge the deterministic observed regions and the stochastic generated regions. We evaluate our approach against several sampling methods in Appendix E.7.

5 Conclusion

This paper introduces Observation-Aligned Mask Priors, a novel framework for learning physical fields directly from incomplete data with authentic, structured occlusions. Instead of relying on predefined or heuristic masking, our approach uses BFN to learn the distribution of real observation masks, guiding the generation to align with specific sparse observations to construct dynamic context-query partitions. We mathematically prove that this intersection-based partitioning guarantees a strictly positive query probability for every valid observed dimension, effectively eliminating zero-query dead zones and preventing local generative collapse. Extensive experiments on three real-world oceanographic datasets across multiple resolutions demonstrate that our method consistently outperforms strong diffusion baselines in reconstruction accuracy and global fidelity. Ultimately, this work establishes that learning data-driven occlusion priors is a highly effective alternative to heuristic masking for reconstructing complex physical fields without access to fully observed training data.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing of the paper. We have not used generative AI tools for coming up idea, and auto research are not applicable to this work. Additionally, we used generative AI tools for checking and correcting errors. We have reviewed all AI-assisted work. LLM-generated sentence was verified by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The sources of the datasets have been provided in Appendix E.1, and detailed parameters have been provided in Appendix E.3 to make the results reproducible.

REFERENCES

Muhammad Ashfaq, Imran Khan, Dilawar Shah, Shujaat Ali, and Muhammad Tahir. Predicting wheat yield using deep learning and multi-source environmental data. *Scientific Reports*, 15(1): 26446, 2025.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2023. URL https://arxiv.org/abs/2107.03006.

Alexander Barth, Julien Brajard, Aida Alvera-Azcárate, Bayoumy Mohamed, Charles Troupin, and Jean-Marie Beckers. Ensemble reconstruction of missing satellite data using a denoising diffusion model: application to chlorophyll a concentration in the black sea. *Ocean Science*, 20(6): 1567–1584, 2024.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis M. Kochmann. Physics-informed diffusion models, 2025. URL https://arxiv.org/abs/2403.14404.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Ashish Bora, Eric Price, and Alexandros G. Dimakis. Ambientgan: Generative models from lossy measurements. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:3481010.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting, 2023. URL https://arxiv.org/abs/2306.01984.

Yuluo Chen, Qiang Chen, Han Ma, Shuailong Chen, and Qingguo Fei. Transfer machine learning framework for efficient full-field temperature response reconstruction of thermal protection structures with limited measurement data. *International Journal of Heat and Mass Transfer*, 242: 126785, 2025.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alexandros G. Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data, 2023. URL https://arxiv.org/abs/2305.19256.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Huseyin Tuna Erdinc, Rafael Orozco, and Felix J. Herrmann. Generative geostatistical modeling from incomplete well and imaged seismic observations with diffusion models, 2024. URL https://arxiv.org/abs/2406.05136.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching, 2024. URL https://arxiv.org/abs/2407.15595.

Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks, 2025. URL https://arxiv.org/abs/2308.07037.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions, 2021. URL https://arxiv.org/abs/2102.05379.

Wan-Lu Hsu, Ying-Lei Lin, Jung-Pin Lai, Yu-Hui Liu, and Ping-Feng Pai. Forecasting corporate financial performance using deep learning with environmental, social, and governance data. *Electronics*, 14(3):417, 2025.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation, 2024. URL https://arxiv.org/abs/2406.17763.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models, 2024. URL https://arxiv.org/abs/2312.02696.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Py Letraon, A. Ali, E. Alvarez Fanjul, L. Aouf, L. Axell, R. Aznar, M. Ballarotta, A. Behrens, M. Benkiran, A. Bentamy, L. Bertino, P. Bowyer, V. Brando, L. A. Breivik, B. Buongiorno Nardelli, S. Cailleau, S. A. Ciliberti, S. Colella, N. Mc Connell, G. Coppini, G. Cossarini, T. Dabrowski, M. de Alfonso Alonsomuñoyerro, E. O'dea, C. Desportes, F. Dinessen, M. Drevillon, Y. Drillet, M. Drudi, R. Dussurget, Y. Faugère, V. Forneris, C. Fratianni, O. Le Galloudec, I. Garcia Hermosa, M. García Sotillo, P. Garnesson, G. Garric, I. Golbeck, J. Gourrion, M. L. Grégoire, S. Guinehut, E. Gutknecht, C. Harris, F. Hernandez, V. Huess, J.A. Johannessen, S. Kay, R. Killick, R. King, J. De Kloe, G. Korres, P. Lagemaa, R. Lecci, J.F. Legeais, J. M. Lellouche, B. Levier, P. Lorente, A. Mangin, M. Martin, A. Melet, J. Murawski, E. Özsoy, A. Palazov, S. Pardo, L. Parent, A. Pascual, J. Paul, E. Peneva, C. Perruche, D. Peterson, L. Petit de La Villeon, N. Pinardi, S. Pouliquen, M. I. Pujol, R. Rainaud, Pierre Rampal, G. Reffray, C. Regnier, A. Reppucci, A. Ryan, S. Salon, A. Samuelsen, R. Santoleri, Andrew Saulter, J. She, C. Solidoro, E. Stanev, J. Staneva, A. Stoffelen, A. Storto, P. Sykes, T. Szekely, G. Taburet, B. Taylor, J. Tintore, C. Toledano, M. Tonani, L. Tuomi, G. Volpe, H. Wedhe, T. Williams, L. Vandendbulcke, D. Van Zanten, K. Von Schuckmann, J. Xie, A. Zacharioudaki, and H. Zuo. THE COPERNICUS MARINE ENVIRONMENTAL MONITORING SERVICE: MAIN SCIENTIFIC ACHIEVEMENTS AND FUTURE PROSPECTS. Research Report 56, MERCATOR OCEAN, September 2017. URL https://hal.univ-grenoble-alpes.fr/hal-03405376.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1lDV3RcKm.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations, 2021. URL https://arxiv.org/abs/2010.08895.

Guojun Liang, Najmeh Abiri, Atiye Sadat Hashemi, Jens Lundström, Stefan Byttner, and Prayag Tiwari. Latent space score-based diffusion model for probabilistic multivariate time series imputation, 2024. URL https://arxiv.org/abs/2409.08917.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers, 2023. URL https://arxiv.org/abs/2308.05732.

Dengyi Liu, Honggang Wang, and Hua Fang. Impute-macfm: Imputation based on mask-aware flow matching, 2025. URL https://arxiv.org/abs/2509.23126.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471, 2022.

Harris Abdul Majid, Giannis Daras, Francesco Tudisco, and Steven McDonagh. Ambient physics: Training neural pde solvers with partial observations, 2026. URL https://arxiv.org/abs/2602.13873.

Scott A Martin, Georgy E Manucharyan, and Patrice Klein. Generative data assimilation for surface ocean state estimation from multi-modal satellite observations. *Journal of Advances in Modeling Earth Systems*, 17(8):e2025MS005063, 2025.

Yidong Ouyang, Liyan Xie, Chongxuan Li, and Guang Cheng. Missdiff: Training diffusion models on tabular data with missing values, 2023. URL https://arxiv.org/abs/2307.00467.

Chicago Y. Park, Shirin Shoushtari, Hongyu An, and Ulugbek S. Kamilov. Measurement score-based diffusion model, 2025. URL https://arxiv.org/abs/2505.11853.

Keyue Qiu, Yuxuan Song, Jie Yu, Hongbo Ma, Ziyao Cao, Zhilong Zhang, Yushuai Wu, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Empower structure-based molecule optimization with gradient guided bayesian flow networks. *arXiv preprint arXiv:2411.13280*, 2024.

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José Miguel Hernández-Lobato, Richard E. Turner, and Emile Mathieu. On conditional diffusion models for pde simulations, 2024. URL https://arxiv.org/abs/2410.16415.

Kaiwen Xue, Yuhao Zhou, Shen Nie, Xu Min, Xiaolu Zhang, Jun Zhou, and Chongxuan Li. Unifying bayesian flow networks and diffusion models through stochastic differential equations. *arXiv preprint arXiv:2404.15766*, 2024.

Jiahao Yu, Qizhen Ying, Leyang Wang, Ziyue Jiang, and Song Liu. Missing data imputation by reducing mutual information with rectified flows, 2025. URL https://arxiv.org/abs/2505.11749.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Improved techniques for maximum likelihood estimation for diffusion odes. In *International Conference on Machine Learning*, pp. 42363–42389. PMLR, 2023.

Zihan Zhou, Chenguang Wang, Hongyi Ye, Yongtao Guan, and Tianshu Yu. Incomplete data, complete dynamics: A diffusion approach. *arXiv preprint arXiv:2509.20098*, 2025.

A Related Work

A.1 Learning from Incomplete Data

Recent generative approaches for incomplete data face critical limitations when applied to complex physical fields. Diffusion methods like MissDiff (Ouyang et al., 2023) and Generative Geostatistical Modeling (Erdinc et al., 2024) rely on heuristic masking strategies that implicitly assume Missing At Random (MAR) conditions. Lacking a physics-guided context-query partitioning mechanism, they struggle to capture essential long-range spatial correlations. Similarly, continuous-time mask-aware paradigms such as Impute-MACFM (Liu et al., 2025) are tailored for tabular data, missing the structural inductive biases required for global fluid continuity. Furthermore, theoretically complex frameworks encounter hard boundaries in physical scenarios. For example, MIRI (Yu et al., 2025) relies on a strict MAR assumption that fundamentally fails in non-random (MNAR) oceanographic occlusions, while incurring prohibitive computational costs. In addition, latent-space diffusion models, such as LSSDM (Liang et al., 2024), depend entirely on pre-trained Variational Autoencoders, which currently cannot losslessly encode high-dimensional ocean states without violating governing physical conservation laws. Furthermore, (Zhou et al., 2025) propose a fundamental framework for learning physical systems from incomplete observations by dividing valid regions into context and query components. A more detailed technical overview of this framework is provided in Sec. 2.

A.2 Discrete Generation via Diffusion and Flow Models

Classic generative models for categorical spaces often struggle to incorporate smooth, gradient-based guidance due to their hard discrete transitions. To address discrete generation, several advanced diffusion and flow frameworks have recently emerged. Multinomial diffusion (Hoogeboom et al., 2021) extends the standard continuous diffusion framework to categorical distributions by iteratively adding categorical noise through uniform transition matrices. Generalizing this approach, structured discrete diffusion models (D3PM) (Austin et al., 2023) introduce specialized transition matrices, such as absorbing and discretized Gaussian states, allowing the forward process to explicitly respect domain priors and ordinal relationships between categories. Concurrently, discrete flow matching (Gat et al., 2024) introduces a probability path framework utilizing continuous-time Markov chains (CTMCs) to define a probability velocity field for categorical data.

Despite their theoretical elegance, adapting these explicitly discrete frameworks for zero-shot physical imputation is restricted. These discrete frameworks fundamentally rely on classifier-free guidance which strictly necessitates conditional labels during training, or otherwise lack a native, differentiable continuous latent space. Bayesian Flow Networks (BFNs) (Graves et al., 2025) solve this hard discrete transition problem by modeling discrete data with continuous parameters on a probability simplex. Recent works (Xue et al., 2024) further unify BFNs with stochastic differential equations to highlight their continuous dynamics. However, standard BFNs mostly focus on unconditional generation. In physical dynamic imputation tasks, the model must dynamically align with sparse observations. While some gradient-guided BFNs exist (Qiu et al., 2024), they rely on heavy historical trajectory corrections. Therefore, existing discrete generation frameworks lack a mechanism for instantaneously gradient-based guidance, making it difficult to seamlessly guide discrete mask generation with real physical observations.

A.3 Diffusion for PDE Problem

Generative modeling for partial differential equations (PDEs) has shifted from purely data-driven mapping to physics-informed diffusion frameworks. Standard models like Fourier Neural Operator (FNO) (Li et al., 2021) effectively learn solution operators in frequency domains but typically require complete, high-resolution training data to capture the underlying physical dynamics. To improve long-term rollouts and capture high-frequency details, diffusion-based solvers such as Dyffusion (Cachay et al., 2023) and PDE-Refiner (Lippe et al., 2023) leverage score-based generation and multi-step refinement, enhancing the statistical fidelity of complex fluid forecasting. Furthermore, Physics-Informed Diffusion Models (PIDMs) (Bastek et al., 2025) explicitly embed PDE constraints directly into the generative process, reducing residual errors by ensuring that the sampled trajectories strictly adhere to governing physical laws. For inverse problems and partial observations, approaches like DiffusionPDE (Huang et al., 2024) and conditional score-based assimilation models (Shysheya et al., 2024) incorporate physics-informed losses and observation likelihoods to guide the denoising process. Recent frameworks like Ambient Physics (Majid et al., 2026) further attempt to learn the joint distribution of coefficient-solution pairs directly from partial observations by randomly masking known measurements during training.

However, these methods primarily treat sparse inputs as static constraints and often depend on pre-defined, idealistic masking patterns. However, when faced with the complex, non-random occlusions typical of oceanographic data, they lack the adaptive partitioning needed to maintain physical consistency across unknown boundaries. Consequently, bridging the gap between unstructured sparse measurements and structured latent dynamics remains a significant challenge for existing PDE-based diffusion frameworks.

B Methodological Connections and Distinctions

This section contextualizes our proposed mask generation module within the broader landscape of probabilistic generative models, specifically detailing its theoretical connections to and differences from the original BFNs and their recent continuous-time reinterpretation.

B.1 Connections and Differences with Original BFNs

Our mask generation module is fundamentally inspired by BFNs introduced by (Graves et al., 2025).

Theoretical connections.

- **Continuous modeling of discrete data:** Both frameworks address the challenge of modeling discrete variables (such as our binary masks) by operating on a continuous parameter space. In the original BFN, the network inputs for discrete data lie on the probability simplex, making the process natively differentiable. We similarly project our noisy latent variables onto the probability simplex, ensuring the network processes continuous probabilities $\mathrm{softmax}(\boldsymbol{x}_t)$ rather than discrete tokens.
- **Data Prediction Objective:** Original BFNs train the network to output the parameters of a clean distribution based on Bayesian-updated input parameters. Our framework adopts a mathematically analogous data prediction formulation, directly outputting predicted class probabilities $\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}$ to estimate the clean discrete categories.

Key differences.

- **Forward process formulation:** The original BFN framework defines the generative process strictly through Bayesian inference, iteratively updating the parameters of an input distribution based on noisy "sender" samples. In contrast, our framework explicitly recasts these dynamics into a continuous diffusion-style forward process. We treat the discrete categories as continuous targets using a scaled logit representation $\boldsymbol{x}_0 = K\boldsymbol{e}_c$ and directly corrupt this continuous target by injecting Gaussian noise.
- **Guidance and intervention:** BFNs were primarily formulated to optimize data compression via maximum likelihood. Our reinterpretation is specifically designed to facilitate gradient-based intervention in the latent space during sampling. This allows us to introduce a classifier guidance mechanism that dynamically aligns the generated masks with sparse physical observations, a capability not natively explored in the standard BFN formulation.

B.2 Connections and Differences with SDE-Unified BFNs

Our continuous reformulation of BFNs shares conceptual similarities with recent work by (Xue et al., 2024), which unifies BFNs and Diffusion Models (DMs) through SDEs.

Theoretical connections.

- **Latent Variable Perspective:** (Xue et al., 2024) demonstrates that BFNs on discrete data implicitly operate as a linear SDE on a set of continuous latent variables $\boldsymbol{z}(t)$, which the original BFN formulation marginalizes out. Our formulation explicitly adopts this continuous latent perspective, defining a forward diffusion process $\boldsymbol{x}_t$ on scaled logits.
- **Alignment with Diffusion Objectives:** (Xue et al., 2024) formally prove that the continuous-time regression loss of a discrete BFN is a reparameterized Denoising Score Matching (DSM) loss. Our framework embraces this alignment, explicitly reparameterizing the noise matching objective into a direct discrete data matching objective $\mathcal{L}_{\text{DM-discrete}}$.

Key differences.

- **Primary Objective:** The primary goal of (Xue et al., 2024)'s SDE unification is to derive specialized, high-order ODE and SDE solvers (BFN-Solvers) to significantly accelerate the sampling speed of BFNs with a limited number of function evaluations. Conversely, our continuous formulation is not engineered for sampling acceleration, but rather to provide a principled pathway for physics-guided conditioning.
- **Application in the Pipeline:** While (Xue et al., 2024) aims to improve BFNs as standalone generative models for general text and image datasets, our framework utilizes the BFN specifically as a dynamic spatial partitioning engine. The outputs of our guided BFN dictate the context

($\boldsymbol{M}_{\text{ctx}}$) and query ($\boldsymbol{M}_{\text{qry}}$) regions for a secondary continuous diffusion model. This ensures the primary model focuses its learning capacity on reconstructing missing physical field structure rather than simply matching unstructured data.

B.3 Connections and Differences with Gradient-Guided BFNs (MolJO)

Recently, (Qiu et al., 2024) proposed MolJO, which extends BFNs with gradient-based guidance for structure-based molecule optimization (SBMO). While both works leverage BFN's capacity for guided generation, our objectives and architectural choices diverge significantly.

Theoretical connections.

- **Gradient guidance in $\boldsymbol{\theta}$ space:** Both methods capitalize on the insight that BFN's Bayesian posterior parameter $\boldsymbol{\theta}$ constitutes a continuous, differentiable space naturally suited for gradient-based optimization. We similarly exploit this property to apply classifier guidance during sampling, allowing the incorporation of external constraints into the mask generation process.
- **Hybrid continuous-discrete modeling:** MolJO addresses the challenge of jointly optimizing continuous coordinates and discrete atom types by guiding their respective $\boldsymbol{\theta}$ components. Analogously, our framework represents discrete binary masks via continuous probability simplex parameters (logits), enabling seamless gradient flow for discrete structure generation.

Key differences.

- **Guidance objective and domain:** MolJO targets multi-objective molecular property optimization (e.g., binding affinity, synthesizability) within 3D protein pockets. In contrast, our guidance is physics-driven, designed to align spatial mask sampling with sparse physical observations (e.g., scattered ocean sensor measurements) for scientific data imputation, rather than drug-like property enhancement.
- **Temporal dependency and correction:** MolJO introduces a sophisticated *backward correction* strategy with a sliding window mechanism to align gradients across generative steps and preserve SE(3)-equivariance for molecular geometries. Our framework eschews such historical correction; instead, we focus on instantaneous physics-guided conditioning to dynamically partition spatial contexts for downstream reconstruction.
- **System integration:** While MolJO serves as the terminal generative model producing final molecular structures, our guided BFN operates as a *spatial routing module* within a larger pipeline. The generated masks explicitly dictate context ($\boldsymbol{M}_{\text{ctx}}$) and query ($\boldsymbol{M}_{\text{qry}}$) regions for a secondary continuous diffusion model, rather than constituting the end-product themselves.

C Theorems and Proofs

C.1 Representational Sufficiency of the Softmax Projection

We first characterize functions that depend only on softmax probabilities, then show that the optimal data predictor in our Gaussian latent model has this property. Throughout, C denotes the number of classes, $K > 0$ the logit scaling factor, and d the number of pixels. Our implementation uses $C = 2$ and $K = 5.4$, giving a true-class probability $\mathrm{softmax}(K\boldsymbol{e}_c)_c \approx 0.9955$ for the binary case. This is just an embedding choice.

Theorem 3 (Softmax factorization)**.** *Let $h : \mathbb{R}^C \to \mathbb{R}^m$ satisfy $h(\boldsymbol{x} + a\mathbf{1}) = h(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^C$ and $a \in \mathbb{R}$. Define the relative interior of the probability simplex by*

$$\Delta_\circ^{C-1} := \mathrm{ri}(\Delta^{C-1}) = \left\{ \boldsymbol{p} \in \mathbb{R}^C : p_j > 0,\ \sum_{j=1}^{C} p_j = 1 \right\}. \tag{12}$$

There exists a unique function $g : \Delta_\circ^{C-1} \to \mathbb{R}^m$ such that $h(\boldsymbol{x}) = g(\mathrm{softmax}(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathbb{R}^C$.

Proof. For $\boldsymbol{p} \in \Delta_\circ^{C-1}$, set $g(\boldsymbol{p}) = h(\log \boldsymbol{p})$, where the logarithm acts elementwise. Since

$$\log \operatorname{softmax}(\boldsymbol{x}) = \boldsymbol{x} - \log\left(\sum_{j=1}^{C} e^{x_j}\right)\mathbf{1}, \tag{13}$$

shift invariance gives $g(\operatorname{softmax}(\boldsymbol{x})) = h(\boldsymbol{x})$. Moreover, $\operatorname{softmax}(\log \boldsymbol{p}) = \boldsymbol{p}$ for every $\boldsymbol{p} \in \Delta_\circ^{C-1}$, so any such factorization must satisfy $g(\boldsymbol{p}) = h(\log \boldsymbol{p})$. This proves uniqueness on the relative interior. □

Boundary of the simplex. Finite logits never map to the boundary, so the theorem does not determine boundary values. If a continuous extension to the closed simplex exists, it is unique, but continuity of h alone does not guarantee its existence. For example, $h(\boldsymbol{x}) = x_1 - x_2$ is continuous and shift invariant, whereas its factorization $g(\boldsymbol{p}) = \log(p_1/p_2)$ has no finite continuous extension to the entire boundary. For d pixels, the same proof applies blockwise: invariance under an independent shift within each pixel's C logits gives a unique factorization through the product $(\Delta_\circ^{C-1})^d$.

Lemma 1 (Shift invariance of the optimal data predictor)**.** *Let $c \in \{1, \dots, C\}$ have prior probabilities $\pi_j = P(c = j)$, and let $\boldsymbol{x}_0 = K\boldsymbol{e}_c$. At a fixed time t, suppose*

$$\boldsymbol{x}_t = \alpha_t K \boldsymbol{e}_c + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_C), \tag{14}$$

where $\boldsymbol{\epsilon}$ is independent of c and $\sigma_t > 0$. Then the class posterior and the optimal squared-error data predictor $D_t^*(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t = \boldsymbol{x}]$ are invariant under $\boldsymbol{x} \mapsto \boldsymbol{x} + a\mathbf{1}$. The same conclusion holds for a full mask under independent Gaussian noise on the stacked logits, with arbitrary dependencies among mask pixels and independent shifts within each pixel's logits.

Proof. Expanding the Gaussian likelihood and cancelling factors common to all classes gives

$$P(c = j \mid \boldsymbol{x}_t = \boldsymbol{x}) = \frac{\pi_j \exp(\lambda_t x_j)}{\sum_{\ell=1}^{C} \pi_\ell \exp(\lambda_t x_\ell)}, \qquad \lambda_t = \frac{\alpha_t K}{\sigma_t^2}. \tag{15}$$

Replacing every x_j by $x_j + a$ multiplies the numerator and denominator by the same positive factor $\exp(\lambda_t a)$. The posterior therefore remains unchanged, as does

$$D_t^*(\boldsymbol{x}) = K\big(P(c = j \mid \boldsymbol{x}_t = \boldsymbol{x})\big)_{j=1}^{C}. \tag{16}$$

For a full mask, let $\boldsymbol{c} = (c_1, \dots, c_d)$ have any joint prior $\pi(\boldsymbol{c})$, and let $x_{i,j}$ denote the logit for class j at pixel i. The same expansion yields

$$P(\boldsymbol{c} \mid \boldsymbol{x}_t = \boldsymbol{x}) \propto \pi(\boldsymbol{c}) \exp\left(\lambda_t \sum_{i=1}^{d} x_{i,c_i}\right). \tag{17}$$

Under $x_{i,j} \mapsto x_{i,j} + a_i$, every configuration receives the common factor $\exp(\lambda_t \sum_i a_i)$, which cancels upon normalization. Thus the joint posterior, its pixelwise marginals, and the posterior mean of the stacked target are invariant. No factorization of $\pi(\boldsymbol{c})$ is required. □

Combining Lemma 1 with Theorem 3 justifies parameterizing the data predictor as $K\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \operatorname{softmax}(\boldsymbol{x}_t))$. At the population optimum, each output block is the posterior class-probability vector. The input softmax probabilities need not themselves equal this posterior. The representational sufficiency claim concerns the optimal data predictor; the full score and noise predictors require the explicit latent terms derived next.

C.2 Reparameterization of BFN from Noise Matching to Discrete Data Matching

We derive the discrete data matching objective in Eq.(5) using the shift-invariant data predictor established in Appendix C.1. Let C be the number of classes and $K > 0$ an independent logit scale ($C = 2$ and $K = 5.4$ in our implementation). For one pixel, $c \in \{1, \dots, C\}$, $\boldsymbol{e}_c \in \mathbb{R}^C$, and $\boldsymbol{x}_0 = K\boldsymbol{e}_c$. The forward process is

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_C), \tag{18}$$

with $\boldsymbol{\epsilon}$ independent of c. All formulas below apply at nondegenerate training times with $\alpha_t > 0$ and $\sigma_t > 0$; singular endpoints require separate limits. For a full mask, we stack d one-hot vectors, replace $\boldsymbol{I}_C$ by $\boldsymbol{I}_{d \times C}$, and sum squared errors over all pixel-class coordinates.

Population-optimal predictors. For a positive time weight, squared-error noise and data matching have conditional-expectation minimizers (Zheng et al., 2023):

$$\boldsymbol{\epsilon}_t^*(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{\epsilon} \mid \boldsymbol{x}_t = \boldsymbol{x}] = -\sigma_t \boldsymbol{s}_t^*(\boldsymbol{x}), \tag{19a}$$

$$D_t^*(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t = \boldsymbol{x}] = \frac{\boldsymbol{x} + \sigma_t^2 \boldsymbol{s}_t^*(\boldsymbol{x})}{\alpha_t}, \tag{19b}$$

where $\boldsymbol{s}_t^*(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ is the true score. Lemma 1 establishes $D_t^*(\boldsymbol{x} + a\mathbf{1}) = D_t^*(\boldsymbol{x})$. Rearranging Eq.(19b) therefore gives

$$\begin{aligned} \boldsymbol{s}_t^*(\boldsymbol{x} + a\mathbf{1}) &= \boldsymbol{s}_t^*(\boldsymbol{x}) - \frac{a}{\sigma_t^2}\mathbf{1}, \\ \boldsymbol{\epsilon}_t^*(\boldsymbol{x} + a\mathbf{1}) &= \boldsymbol{\epsilon}_t^*(\boldsymbol{x}) + \frac{a}{\sigma_t}\mathbf{1}. \end{aligned} \tag{20}$$

Thus the full score and noise predictors are not shift invariant. A network receiving only $\mathrm{softmax}(\boldsymbol{x})$ cannot represent either of them without an additional explicit dependence on $\boldsymbol{x}$.

Data and noise parameterizations. Theorem 3 and Lemma 1 instead justify the data parameterization

$$D_{\boldsymbol{\theta}}(t, \boldsymbol{x}) = K\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \mathrm{softmax}(\boldsymbol{x})), \tag{21}$$

where the network outputs a class-probability vector in Δ^{C-1} at each pixel. Define the corresponding noise predictor by

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(t, \boldsymbol{x}) \coloneqq \frac{\boldsymbol{x} - \alpha_t D_{\boldsymbol{\theta}}(t, \boldsymbol{x})}{\sigma_t}. \tag{22}$$

The explicit $\boldsymbol{x}/\sigma_t$ term preserves the required shift behavior. This family contains the population-optimal noise predictor whenever the data network represents the posterior mean; the claim does not require an arbitrary noise predictor to be shift invariant.

Exact loss reparameterization. For every training sample and every parameter value, Eqs.(18) and (22) imply

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(t, \boldsymbol{x}_t) - \boldsymbol{\epsilon} = \frac{\alpha_t K}{\sigma_t}\left(\boldsymbol{e}_c - \hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \mathrm{softmax}(\boldsymbol{x}_t))\right). \tag{23}$$

Consequently,

$$\begin{aligned} \mathcal{J}_{\mathrm{noise}}(\boldsymbol{\theta}) &\coloneqq \mathbb{E}\left[w_{\mathrm{NM}}(t)\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(t, \boldsymbol{x}_t) - \boldsymbol{\epsilon}\|^2\right] \\ &= \mathbb{E}\left[w_{\mathrm{NM}}(t)\frac{\alpha_t^2 K^2}{\sigma_t^2}\|\hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \mathrm{softmax}(\boldsymbol{x}_t)) - \boldsymbol{e}_c\|^2\right]. \end{aligned} \tag{24}$$

This is exactly Eq.(5) when

$$w(t) = w_{\mathrm{NM}}(t)\frac{\alpha_t^2 K^2}{\sigma_t^2}. \tag{25}$$

These identities assume finite expectations and the same time-sampling distribution. Other strictly positive time weights have the same pointwise population minimizer when the predictor is unrestricted, but need not give identical objectives or the same optimum within a finite-capacity model.

Score estimation for sampling. The corresponding score estimate is

$$\boldsymbol{s}_{\boldsymbol{\theta}}(t, \boldsymbol{x}) = -\frac{\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(t, \boldsymbol{x})}{\sigma_t} = \frac{\alpha_t K \hat{\boldsymbol{e}}_{\boldsymbol{\theta}}(t, \mathrm{softmax}(\boldsymbol{x})) - \boldsymbol{x}}{\sigma_t^2}. \tag{26}$$

It equals $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ at the population optimum, and otherwise provides an estimate for diffusion sampling. For any network parameters, it obeys the score transformation in Eq.(20) because its learned data-prediction component is shift invariant. This separation of a learned probability predictor and an explicit linear latent term is consistent with the discrete BFN score parameterization of Xue et al. (2024).

C.3 Strict Positivity of Query Probabilities via Mask Intersection

Assumption 1 (Intersection Coverage Completeness)**.** *Let $S = \{\boldsymbol{x} \in \{0,1\}^d \mid P(\boldsymbol{x}) > 0\}$ denote the support of the distribution $P(\boldsymbol{M})$. For any valid context mask $\boldsymbol{m}$ (i.e., $\exists \boldsymbol{x}, \boldsymbol{y} \in S$ such that $\boldsymbol{x} \odot \boldsymbol{y} = \boldsymbol{m}$) and any spatial dimension i where $\boldsymbol{m}_i = 0$, there exists at least one pair of masks $(\boldsymbol{M}_a, \boldsymbol{M}_b) \in S \times S$ such that:*

1. *$\boldsymbol{M}_a \odot \boldsymbol{M}_b = \boldsymbol{m}$ (Intersectability)*
2. *$(\boldsymbol{M}_a)_i = 1$ (Local Coverage)*

Theorem 1 (Strict Positivity of Query Probabilities via Mask Intersection)**.** *Let $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \{0,1\}^d$ be independently sampled binary masks from the same underlying distribution $p(\boldsymbol{M})$. Define the context mask as $\boldsymbol{M}_{\text{ctx}} = \boldsymbol{M}_1 \odot \boldsymbol{M}_2$ and the query mask as $\boldsymbol{M}_{\text{qry}} = \boldsymbol{M}_1 \odot (\boldsymbol{1} - \boldsymbol{M}_{\text{ctx}})$. Under Assumption 1, for any valid context mask $\boldsymbol{m}$ and any unobserved dimension i (where $\boldsymbol{m}_i = 0$), the conditional probability of dimension i being selected as a query is strictly positive:*

$$P((\boldsymbol{M}_{\text{qry}})_i = 1 \mid \boldsymbol{M}_{\text{ctx}} = \boldsymbol{m}) > 0. \tag{3}$$

Proof. By the definition of conditional probability, we can expand the target expression as:

$$P((\boldsymbol{M}_{\text{qry}})_i = 1 \mid \boldsymbol{M}_{\text{ctx}} = \boldsymbol{m}) = \frac{P((\boldsymbol{M}_{\text{qry}})_i = 1 \text{ and } \boldsymbol{M}_{\text{ctx}} = \boldsymbol{m})}{P(\boldsymbol{M}_{\text{ctx}} = \boldsymbol{m})} \tag{27}$$

First, we establish that the denominator is strictly positive. Since $\boldsymbol{m}$ is defined as a valid context mask, there exist $\boldsymbol{x}, \boldsymbol{y} \in S$ such that $\boldsymbol{x} \odot \boldsymbol{y} = \boldsymbol{m}$. By the definition of the support set S, $P(\boldsymbol{x}) > 0$ and $P(\boldsymbol{y}) > 0$. By independence condition, the probability of this specific combination is:

$$P(\boldsymbol{M}_1 = \boldsymbol{x}, \boldsymbol{M}_2 = \boldsymbol{y}) = P(\boldsymbol{x})P(\boldsymbol{y}) > 0 \tag{28}$$

Therefore, the marginal probability of the context mask is bounded from below by this specific combination, implying $P(\boldsymbol{M}_{\text{ctx}} = \boldsymbol{m}) > 0$.

Next, we analyze the numerator, which represents the joint probability of two events occurring simultaneously: $\boldsymbol{M}_1 \odot \boldsymbol{M}_2 = \boldsymbol{m}$ and $(\boldsymbol{M}_{\text{qry}})_i = 1$. We substitute the definition of the query mask into the second condition:

$$(\boldsymbol{M}_{\text{qry}})_i = (\boldsymbol{M}_1)_i \cdot (1 - (\boldsymbol{M}_{\text{ctx}})_i) \tag{29}$$

Given the condition $\boldsymbol{M}_{\text{ctx}} = \boldsymbol{m}$ and the premise that we are evaluating an unobserved dimension where $\boldsymbol{m}_i = 0$, the equation simplifies to:

$$(\boldsymbol{M}_{\text{qry}})_i = (\boldsymbol{M}_1)_i \cdot (1 - 0) = (\boldsymbol{M}_1)_i \tag{30}$$

Thus, the numerator fundamentally computes the probability of the joint event: $\{\boldsymbol{M}_1 \odot \boldsymbol{M}_2 = \boldsymbol{m} \text{ and } (\boldsymbol{M}_1)_i = 1\}$.

By Assumption 1, because $\boldsymbol{m}$ is a valid context mask and $\boldsymbol{m}_i = 0$, there is guaranteed to exist a specific pair of masks $(\boldsymbol{M}_a, \boldsymbol{M}_b) \in S \times S$ satisfying $\boldsymbol{M}_a \odot \boldsymbol{M}_b = \boldsymbol{m}$ and $(\boldsymbol{M}_a)_i = 1$.

Since $\boldsymbol{M}_a \in S$ and $\boldsymbol{M}_b \in S$, their independent marginal probabilities are strictly positive ($P(\boldsymbol{M}_a) > 0$ and $P(\boldsymbol{M}_b) > 0$). Under the independence condition, the probability of sampling this exact pair is:

$$P(\boldsymbol{M}_1 = \boldsymbol{M}_a, \boldsymbol{M}_2 = \boldsymbol{M}_b) = P(\boldsymbol{M}_a)P(\boldsymbol{M}_b) > 0 \tag{31}$$

Because we have found at least one valid combination of masks in the distribution that perfectly satisfies the conditions of the numerator, and the probability of sampling this combination is strictly greater than zero, the total probability of the numerator must also be strictly positive:

$$P((\boldsymbol{M}_{\text{qry}})_i = 1 \text{ and } \boldsymbol{M}_{\text{ctx}} = \boldsymbol{m}) \geq P(\boldsymbol{M}_1 = \boldsymbol{M}_a, \boldsymbol{M}_2 = \boldsymbol{M}_b) > 0 \tag{32}$$

Finally, dividing a strictly positive numerator by a strictly positive denominator yields a strictly positive result:

$$\frac{P((\boldsymbol{M}_{\text{qry}})_i = 1 \text{ and } \boldsymbol{M}_{\text{ctx}} = \boldsymbol{m})}{P(\boldsymbol{M}_{\text{ctx}} = \boldsymbol{m})} > 0 \tag{33}$$

Hence, $P((\boldsymbol{M}_{\text{qry}})_i = 1 \mid \boldsymbol{M}_{\text{ctx}} = \boldsymbol{m}) > 0$, completing the proof. □

C.4 Strict Positivity of Query Probabilities under Ratio-Guided Partitioning

Assumption 2 (Constrained Coverage Completeness). *Let $\boldsymbol{M} \in \{0,1\}^d$ be a fixed incomplete observation mask. Let C_k denote the hard constraint that a conditionally generated mask $\hat{\boldsymbol{M}}$ must share exactly k valid intersecting pixels with the observation, i.e., $|\hat{\boldsymbol{M}} \odot \boldsymbol{M}| = k$, where $0 < k < |\boldsymbol{M}|$. Let $S_{guided} = \{\hat{\boldsymbol{M}} \in \{0,1\}^d \mid P(\hat{\boldsymbol{M}} \mid C_k) > 0\}$ be the support set of the conditionally guided mask distribution. We assume that the physical mask prior does not deterministically collapse on any single spatial coordinate under this volume constraint. Specifically, for any spatial dimension i where $\boldsymbol{M}_i = 1$, there exists at least one mask $\hat{\boldsymbol{M}}_a \in S_{guided}$ such that $(\hat{\boldsymbol{M}}_a)_i = 0$.*

This assumption is physically well-founded. In authentic oceanographic observations, missing patterns, such as those driven by dynamic cloud cover or shifting satellite trajectories, exhibit immense spatial diversity. Constraining the intersection volume k merely restricts the *capacity* of the context region without eliminating the inherent *stochasticity* of the physical prior. A well-trained generative model naturally synthesizes multiple distinct, physically plausible occlusion topologies that satisfy the exact same volume constraint, ensuring that no single spatial dimension is deterministically anchored across all valid generations.

Theorem 2 (Strict Positivity of Query Probabilities under Ratio-Guided Partitioning). *Let $\boldsymbol{M} \in \{0,1\}^d$ be a fixed observation mask. Let the conditionally generated mask $\hat{\boldsymbol{M}} \in \{0,1\}^d$ be sampled from the guided distribution $P(\hat{\boldsymbol{M}} \mid C_k)$ satisfying the intersection constraint C_k as defined in Assumption 2. Define the context mask as $\boldsymbol{M}_{ctx} = \hat{\boldsymbol{M}} \odot \boldsymbol{M}$ and the query mask as $\boldsymbol{M}_{qry} = \boldsymbol{M} \odot (1 - \boldsymbol{M}_{ctx})$. Under Assumption 2, for any valid observed dimension i (where $\boldsymbol{M}_i = 1$), the marginal probability of dimension i being assigned as a query region is strictly positive: $P((\boldsymbol{M}_{qry})_i = 1 \mid C_k) > 0$.* [2]

Proof. By the definitions of the context and query masks, we can explicitly express the query mask $\boldsymbol{M}_{\text{qry}}$ in terms of the fixed observation $\boldsymbol{M}$ and the conditionally generated mask $\hat{\boldsymbol{M}}$:

$$\boldsymbol{M}_{\text{qry}} = \boldsymbol{M} \odot (1 - \hat{\boldsymbol{M}} \odot \boldsymbol{M}) \tag{34}$$

Since $\boldsymbol{M}$ is a binary mask (i.e., $\boldsymbol{M}_i \in \{0,1\}$), the element-wise multiplication yields $\boldsymbol{M} \odot \boldsymbol{M} = \boldsymbol{M}$. Thus, the equation simplifies to:

$$\boldsymbol{M}_{\text{qry}} = \boldsymbol{M} \odot (1 - \hat{\boldsymbol{M}}) \tag{35}$$

We are evaluating a specific spatial dimension i where the authentic observation is valid, meaning $\boldsymbol{M}_i = 1$. Substituting this into the simplified query equation for dimension i, we obtain:

$$(\boldsymbol{M}_{\text{qry}})_i = \boldsymbol{M}_i \cdot (1 - \hat{\boldsymbol{M}}_i) = 1 \cdot (1 - \hat{\boldsymbol{M}}_i) = 1 - \hat{\boldsymbol{M}}_i \tag{36}$$

This deterministic relationship reveals that under a fixed observation $\boldsymbol{M}$, the event $\{(\boldsymbol{M}_{\text{qry}})_i = 1\}$ is mathematically equivalent to the event $\{\hat{\boldsymbol{M}}_i = 0\}$. Consequently, their probabilities under the guided distribution are identical:

$$P((\boldsymbol{M}_{\text{qry}})_i = 1 \mid C_k) = P(\hat{\boldsymbol{M}}_i = 0 \mid C_k) \tag{37}$$

According to Assumption 2, because $0 < k < |\boldsymbol{M}|$, the constrained support set S_{guided} preserves sufficient generative diversity. Therefore, there exists at least one valid mask $\hat{\boldsymbol{M}}_a \in S_{\text{guided}}$ that satisfies $(\hat{\boldsymbol{M}}_a)_i = 0$.

By the definition of the support set S_{guided}, any mask within it has a strictly positive probability of being sampled:

$$P(\hat{\boldsymbol{M}} = \hat{\boldsymbol{M}}_a \mid C_k) > 0 \tag{38}$$

The marginal probability $P(\hat{\boldsymbol{M}}_i = 0 \mid C_k)$ is formulated as the sum of the probabilities of all generated masks in S_{guided} where the i-th dimension equals 0. Since $\hat{\boldsymbol{M}}_a$ is a member of this set and

[2] The condition shifts from $\boldsymbol{M}_{\text{ctx}}$ (Theorem 1) to C_k to reflect the algorithmic transition from random partitioning to observation-anchored generation. Guaranteeing the marginal probability $P((\boldsymbol{M}_{\text{qry}})_i = 1 \mid C_k) > 0$ is mathematically sufficient to ensure a strictly positive query frequency ($p_i > 0$ in Eq.(41)) across epochs. See Remark 1 in Appendix for a detailed discussion.

possesses a strictly positive probability, the entire sum must be strictly bounded away from zero:

$$P(\hat{\boldsymbol{M}}_i = 0 \mid C_k) = \sum_{\hat{\boldsymbol{m}} \in S_{\text{guided}}, \hat{\boldsymbol{m}}_i = 0} P(\hat{\boldsymbol{M}} = \hat{\boldsymbol{m}} \mid C_k) \geq P(\hat{\boldsymbol{M}} = \hat{\boldsymbol{M}}_a \mid C_k) > 0 \tag{39}$$

It trivially follows that:

$$P((\boldsymbol{M}_{\text{qry}})_i = 1 \mid C_k) > 0 \tag{40}$$

This completes the proof. □

Remark 1 (Connection to Expected Gradients and Theorem 1)**.** *Readers may notice a subtle shift in the condition from $\boldsymbol{M}_{ctx}$ in Theorem 1 to C_k in Theorem 2. This reflects the fundamental architectural shift from random bipartite splitting to observation-anchored generation. In the unconditional case (Theorem 1), $\boldsymbol{M}_{ctx}$ is highly stochastic, so we must mathematically guarantee that any valid randomly formed context leaves a non-zero query probability to avoid zero-gradient dead zones. However, under the guided framework, the actual observation $\boldsymbol{M}$ is fixed for a given data sample. Both $\boldsymbol{M}_{ctx}$ and $\boldsymbol{M}_{qry}$ are jointly and deterministically derived from a single generated mask $\hat{\boldsymbol{M}}$ subject to the guidance constraint C_k. Therefore, guaranteeing the marginal probability $P((\boldsymbol{M}_{qry})_i = 1 \mid C_k) > 0$ is mathematically sufficient. It ensures that across multiple training epochs, every valid observation point has a strictly positive frequency of being queried ($p_i > 0$ in Eq.(41), thereby satisfying the prerequisite for non-vanishing expected gradients.*

We also need to clarify that the exact-overlap constraint used in Theorem 2 is an analytical abstraction for characterizing the support condition required for positive query exposure, rather than an exact sampling operation implemented by Algorithm 1. In practice, Algorithm 1 constructs stochastic anchors through Bernoulli retention and applies a soft cross-entropy guidance signal during BFN sampling. For the high-dimensional observation masks considered in our experiments, each sample contains a large number of valid observed coordinates, so Bernoulli retention produces a non-empty stochastic anchor set with overwhelming probability. In particular, the event that all valid observations are simultaneously discarded is vanishingly unlikely under our practical settings; hence, the required condition $k > 0$ is effectively always satisfied in practice rather than constituting an additional restrictive assumption. At the same time, the stochastic anchor set remains non-empty without overwhelmingly high probability of deterministically retaining any individual observed coordinate across all sampling instances. Thus, the practical procedure preserves the stochastic non-degeneracy that motivates the support condition in Theorem 2, rather than enforcing a fixed-cardinality constraint explicitly. The theorem can therefore be interpreted as identifying a sufficient support-level condition under which observation-guided partitioning preserves positive query exposure, while the empirical query-frequency analysis in Appendix E.8 verifies that this condition is maintained by the implemented soft-guidance sampler in practice.

D Learning Dynamics Basis and Analysis

Preliminary property. As illustrated in previous work (Zhou et al., 2025), Eq.(2) is fundamentally dictated by the learning dynamics, where the expected squared gradient magnitude for feature i scales linearly with its query probability:

$$\mathbb{E}\left[\left(\frac{\partial \mathcal{L}}{\partial (\boldsymbol{u}_{\boldsymbol{\phi}})_i}\right)^2\right] = 4p_i \mathbb{E}\left[((\boldsymbol{u}_{\boldsymbol{\phi}})_i - (\boldsymbol{u}_{\text{obs}})_i)^2 \,\middle|\, (\boldsymbol{M}_{\text{qry}})_i = 1\right], \tag{41}$$

where $p_i = P((\boldsymbol{M}_{\text{qry}})_i = 1 | \boldsymbol{M}_{\text{ctx}})$ directly controls the frequency of parameter updates. If $p_i = 0$, the given dimension receives no gradient updates, causing the model to predict arbitrary and meaningless values in that region.

Heuristic reliability indicator. As established in our Eq.(41), the expected gradient update for a specific feature dimension scales linearly with its query probability. This dictates a direct relationship between sampling probability and learning frequency: regions with a higher $P(qry)$ are exposed to the loss function more frequently during training.Because the prediction model is optimized more heavily in these high-frequency regions, the $P(qry)$ distribution effectively captures the model's spatial familiarity with the data. Although not a formally calibrated statistical uncertainty metric, $P(qry)$ acts as a strong proxy for epistemic confidence. During inference, regions that exhibited

high query probabilities in the training phase can generally be interpreted as having higher predictive reliability. Consequently, this distribution provides an accessible and heuristic indicator to assess the relative trustworthiness of the reconstructed physical fields across different spatial domains.

E Supplementary Experiments

E.1 Dataset Settings

All real-world oceanographic datasets utilized in our experiments are sourced from the E.U. Copernicus Marine Environment Monitoring Service (CMEMS). We specifically selected Level-3 (L3) satellite observation products. Unlike Level-4 (L4) products, which are already synthetically interpolated and gap-filled, L3 products contain unadulterated, authentic missingness patterns caused by real-world physical constraints such as cloud occlusions, sensor swath gaps, and orbital trajectories. This makes them ideal testbeds for evaluating the imputation of complex, structured missing fields.

Black Sea CHL (Black Sea Chlorophyll-a Dataset) (Letraon et al., 2017). This dataset is derived from the CMEMS product `OCEANCOLOUR_BLK_BGC_L3_MY_009_153` (`https://doi.org/10.48670/moi-00303`). It provides daily multi-sensor merged objective observations of Chlorophyll-a (CHL) concentration in the Black Sea at a 1-km spatial resolution. Chlorophyll-a distributions exhibit highly complex, non-linear spatial gradients driven by mesoscale eddies and coastal currents. Coupled with the extensive and contiguous cloud occlusions typical of the Black Sea region, this dataset challenges an imputation model's ability to recover fine-grained, localized ecological structures from sparse and fragmented observations. We picked up approximately 10,000 samples for our experiments.

Baltic Sea NANO (Baltic Sea Nano-Phytoplankton Dataset) (Letraon et al., 2017). Sourced from the CMEMS product `OCEANCOLOUR_BAL_BGC_L3_MY_009_133` (`https://doi.org/10.48670/moi-00296`), this dataset provides daily regional ocean color observations, specifically targeting nano-phytoplankton concentrations at a 1-km resolution. The Baltic Sea possesses a highly irregular topology with numerous archipelagos and narrow straits. Consequently, the missingness patterns are governed not only by atmospheric occlusions but also by complex, rigid coastal boundaries. This presents a unique challenge for imputation tasks, requiring models to reconstruct physically plausible fields that strictly respect intricate geographical constraints rather than applying naive spatial smoothing across landmasses. The dataset used in our experiments contains roughly 5,000 samples.

Global Ocean SSS (Global Sea Surface Salinity Dataset) (Letraon et al., 2017). This dataset utilizes the CMEMS global product `MULTIOBS_GLO_PHY_SSS_L3_MYNRT_015_014` (`https://doi.org/10.48670/mds-00368`). It offers daily global Sea Surface Salinity measurements at a coarser 25-km spatial resolution. In contrast to regional datasets, this global dataset involves macroscopic spatial correlations dominated by large-scale ocean circulation and climatic forcing. Furthermore, the missingness typology fundamentally shifts from primarily cloud-induced irregular holes to distinct satellite orbital tracks and systematic inter-swath gaps. This dataset is ideal for validating the scalability of imputation algorithms and their capacity to maintain global physical consistency under highly structured, geometry-driven masking topologies. We selected around 6,000 samples as our dataset.

E.2 Baseline and Method Comparison

Here we also considered other baselines but ultimately omitted them from direct empirical comparisons. We excluded AmbientGAN (Bora et al., 2018) and MisGAN (Li et al., 2019) because purely adversarial training suffers from severe instability, making it exceedingly difficult to converge on highly sparse datasets. Furthermore, we reviewed recent generative frameworks tailored for partial observations, including Ambient Physics (Majid et al., 2026), MSM (Park et al., 2025), and the context-query diffusion approach (Zhou et al., 2025). They fundamentally rely on similar overly simplistic or manually designed mask distributions (e.g., uniform random dropout or regular block masking). These idealistic assumptions usually fail to capture the complex, missing-not-at-random (MNAR) topology of authentic physical occlusions. And we equivalently demonstrate their core idea by comparing different masking strategies in Sec. 4.5. A more direct and comprehensive comparison with the most related methods is shown in Tab. 3.

Table 3: Comparison with the most related methods. We use × to denote that it cannot be achieved or is extremely costly or engineering-difficult. And use ✓ to indicate that this can be achieved.

Methods	MNAR Assumption	Non-heuristic	High Dimensional
ImputeMACFM (Liu et al., 2025)	✓	×	×
MIRI (Yu et al., 2025)	×	×	×
LSSDM (Liang et al., 2024)	×	✓	×
AmbientDiff (Daras et al., 2023)	×	✓	✓
MissDiff (Ouyang et al., 2023)	×	✓	✓
DINDiff (Barth et al., 2024)	✓	✓	✓
Ambient Physics (Majid et al., 2026)	×	×	✓
MSM (Park et al., 2025)	×	×	✓
Ours	✓	✓	✓

E.3 Implementation Details

In all experiments, considering both the stability of the guidance and the diversity of the generated data, we empirically set $\rho = 0.8$. Setting the value of ρ too high will result in insufficient diversity of conditional generation, while setting it too low will make it difficult to align with the guidance target. Regarding BFN sampling steps n, we find that 10 to 20 ODE sampling steps are usually sufficient to achieve the desired effect. As the number of sampling steps increases, the improvement in effect becomes relatively limited. For guidance scale w_g, a moderate interval is between 80 and 200 and usually needs to be adapted to the selected sampling steps. The value of the guidance scale seems larger than that in traditional classifier guidance because we perform global normalization on our guidance objective. This parameter is not particularly sensitive, a reasonable setting is sufficient.

Our training and test sets are completely separated in time with no overlap, and each has been internally shuffled. This ensures that no data leakage occurs between data from adjacent dates. Also, to explore the best performance of the baseline AmbientDiff, we attempt several combinations of context and query ratio in Tab. 4.

Table 4: Exploration of suitable parameters for AmbientDiff on the Black Sea CHL dataset with the resolution of 64×64.

Context Ratio	Query Ratio	Black Sea CHL		Baltic Sea NANO		Global Ocean SSS	
		MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑
10%	10%	0.243	16.775	0.376	15.652	0.294	17.992
30%	30%	**0.228**	**17.717**	**0.365**	**15.466**	**0.286**	**18.022**
50%	50%	0.275	16.926	0.439	14.776	0.292	17.957
70%	70%	0.344	16.171	0.495	14.323	0.323	17.596
90%	90%	0.428	15.439	0.619	13.488	0.344	17.363

E.4 Complete Results

To further demonstrate the scalability and robustness of our framework across different spatial scales, we provide additional experimental results at 128×128 and 256×256 resolutions on all three datasets in Tab. 5 and 6. This demonstrates that our method remains effective and stands out more at high resolutions.

E.5 Supplementary Experiments

To make a more fair and clean comparison, we provide more experiments with a strict complementary context-query setup and similar ratios, aligning them with our method. This ensures, to the greatest extent possible, that the ratio of context and query closely matches that of our approach on different datasets. And there is no overlap between context and query, thereby eliminating the information leakage as a potential confounding factor in the comparison. Based on the real-time curves of the context and query ratio recorded during training, we observed that the context proportion fluctuates around 50% for the Black Sea CHL and Baltic Sea NANO dataset, and around 70% for the Global

Table 5: Performance comparison on physical fields imputation tasks with the resolution of 128×128 across three real-world datasets.

Method	Black Sea CHL		Baltic Sea NANO		Global Ocean SSS	
	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑
AmbientDiff	0.268	20.432	0.466	15.648	0.440	19.165
DINDiff	0.275	18.800	0.385	16.275	0.578	17.967
MissDiff	0.706	16.702	0.753	13.480	0.683	17.342
Ours	**0.210**	**21.020**	**0.333**	**16.730**	**0.379**	**19.799**

Table 6: Performance comparison on physical fields imputation tasks with the resolution of 256×256 across three real-world datasets.

Method	Black Sea CHL		Baltic Sea NANO		Global Ocean SSS	
	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑
AmbientDiff	0.256	20.523	0.487	16.234	0.524	20.578
DINDiff	0.266	21.389	0.449	17.222	0.567	19.610
MissDiff	0.655	16.887	0.818	14.147	0.731	19.187
Ours	**0.160**	**21.889**	**0.320**	**17.616**	**0.441**	**20.801**

Ocean SSS dataset, with the magnitude of fluctuation generally remaining within 10%. And the query is complementary to the context. The results in Tabs. 9, 10, and 11 clearly demonstrate that the effectiveness of our method is not determined by the ratio. The performance gains achieved by adjusting the ratio are far smaller than those yielded by our method based on all experiments. This makes a clean isolation to support that the bottleneck mainly exists in the context-query partition strategy rather than the context-query ratio.

To assess the generalizability of our guided partitioning strategy across different neural architectures, we also evaluate our method using two distinct backbones: Vision Transformer (ViT) (Dosovitskiy et al., 2021) and Karras UNet (Karras et al., 2024). The results presented in Tab. 8 indicate that while both architectures achieve comparable performance on the Black Sea Chlorophyll-a Dataset at a resolution of 64×64, while UNet achieves superior performance, producing a lower MSE and a higher PSNR compared to ViT. This performance gap suggests that the hierarchical inductive bias of UNet is well-suited for capturing the multi-scale spatial dependencies inherent in oceanic physical fields.

E.6 Context Mask Selection Strategy Comparison.

We conduct a comprehensive comparison of our approach with various context mask selection strategies here.

- **Pixel-level Partition:** Each pixel in the observed region is independently sampled using a Bernoulli distribution with a constant probability p (where p is the ratio of required points to total points), producing a spatially uniform subset as $\boldsymbol{M}_{\text{ctx}}$.

Table 7: Performance comparison of different sampling strategies on the 64×64 Black Sea CHL Dataset.

Methods	MSE ↓	PSNR ↑
Direct Projection	0.193	17.825
Proximal Prediction	**0.189**	17.669
Iterative Conditioning	0.199	18.330
Resampling Inpainting	0.199	18.330
Recursive Jump Inpainting	0.199	**18.359**

Table 8: Backbone performance comparison on the Black Sea CHL dataset with the resolution of 64×64.

Methods	MSE ↓	PSNR ↑
Vit	0.225	16.701
Unet	**0.193**	**17.825**

Table 9: Comparison on the Black Sea CHL dataset under controlled context-query ratios.

Method	**Context**	**Query**	**MSE↓**	**PSNR↑**
AmbientDiff	50%	50%	0.263	17.097
Ours	∼50%	∼50%	**0.193**	**17.825**

- **Block-wise Partition:** The observed region is partitioned into a grid of non-overlapping blocks. A subset of these blocks is randomly sampled with a constant probability p (where p is the ratio of selected blocks to total blocks), and their intersection with the ground-truth mask produces a subset as $\boldsymbol{M}_{\text{ctx}}$.
- **Saliency-driven Partition:** This heuristic strategy relies on sorting local gradient magnitudes to allocate the available context points. Sample half of $\boldsymbol{M}_{\text{ctx}}$ from high-gradient regions and the other half from low-gradient regions.
- **Empirical Distribution Partition:** Randomly sample a real observation mask from the remaining dataset, then consider the intersection of this mask and $\boldsymbol{M}_{\text{real}}$ as $\boldsymbol{M}_{\text{ctx}}$.
- **Unconditional Prior Partition:** It employs a mask sampled unconditionally by a BFN which was pre-trained to capture the dataset's observation mask distribution (Sec. 3.3). The intersection of this sampled mask and $\boldsymbol{M}_{\text{real}}$ is used as $\boldsymbol{M}_{\text{ctx}}$.
- **Ours:** Our framework introduces a cross-entropy guidance mechanism during the BFN sampling phase to dynamically anchor the generation process (Sec. 3.4). This produces a tailored mask specific to the current data point, and $\boldsymbol{M}_{\text{ctx}}$ is the intersection of this conditionally generated mask and $\boldsymbol{M}_{\text{real}}$.

To make a fair and clean comparison, we provide experiments with a strict complementary context-query setup and under similar context ratios. Based on the real-time curves of the context and query ratio recorded during training, we observed that the context proportion fluctuates around 50% for the Black Sea CHL ith the magnitude of fluctuation generally remaining within 10%. Thus, for Pixel-level and Block-wise, we use the context ratio to 0.5 for better comparison.

E.7 Sampling Method Comparison.

We conduct a comprehensive comparison of our approach with various sampling methods here.

- **Direct Projection:** Perform a single-step projection to approximate the conditional expectation $\mathbb{E}[\boldsymbol{u}_0|\boldsymbol{u}_{\text{obs}}, \boldsymbol{M}]$. By mapping the uncorrupted partial observations directly onto the data manifold via ensemble averaging, it minimizes global reconstruction error but often produces blurry results in high-frequency regions.
- **Proximal Prediction:** A variant of the single-step approach that applies an infinitesimal noise level before projection. This initial noise injection acts as a proximal constraint that aligns with the diffusion model's prior, while the final step still tightly anchors the solution to the deterministic manifold of the observed data, better mitigating blurriness and ensuring higher fidelity to local pixel values.
- **Iterative Conditioning:** Decomposes the complex conditional expectation into a diffusion expectation capturing the global prior and an imputation expectation enforcing structural constraints. By blending the two components through a monotonically increasing weight schedule over multiple steps, it effectively balances global physical dynamics with local observation consistency.

Table 10: Comparison on the Baltic Sea NANO dataset under controlled context-query ratios.

Method	Context	Query	MSE↓	PSNR↑
AmbientDiff	50%	50%	0.441	15.021
Ours	∼50%	∼50%	**0.294**	**16.101**

Table 11: Comparison on the Global Ocean SSS dataset under controlled context-query ratios.

Method	Context	Query	MSE↓	PSNR↑
AmbientDiff	70%	30%	0.331	17.455
Ours	∼70%	∼30%	**0.276**	**18.192**

- **Resampling Inpainting:** Harmonize observation boundaries by alternating between denoising and a resampling step where noise is added back to the current estimate. This iterative loop allows the model to better align the generated content with the fixed observed regions through multiple stochastic trials.
- **Recursive Jump Inpainting:** An extension of the Resampling Inpainting framework that utilizes profound temporal jumps back to earlier noise levels. By re-injecting significant variance, the model alleviates local boundary inconsistencies and more effectively reconciles the global physical dynamics with the sparse observation constraints.

The quantitative results in Tab. 7 demonstrate a fundamental trade-off between minimizing expectation error and preserving global signal fidelity. The single-step methods achieve the lowest global reconstruction loss as theoretically established in (Zhou et al., 2025). Iterative approaches, such as Iterative Conditioning and Resampling Inpainting, significantly improve PSNR but albeit at a slight cost to the global loss. This improvement aligns with the theoretical foundation of Resampling Inpainting, that is, the iterative forward-backward diffusion process continuously harmonizes the joint distribution of the known context and the generated content. Notably, the Recursive Jump Inpainting achieves the best overall structural fidelity (highest PSNR). By introducing large noise jumps, this profound re-injection strategy prevents the model from settling into local boundary inconsistencies, ultimately yielding the most seamless and physically natural transitions across the mask boundaries. Detailed algorithm implementation for all sampling methods are provided in Appendix G.

E.8 Uncertainty Quantification of Query Distribution

To further validate the efficacy of our partitioning mechanism, we conduct a comparative analysis of the spatial query probability distribution between our proposed method and a naive random selection baseline across all three datasets. We set the number of ensemble samplings to 200 and conduct multiple experiments to make the results more generalizable and reliable. As shown in Fig. 4, a naive random selection strategy yields a highly fragmented query distribution with numerous "dead zones" where the selection probability approaches zero. These "dead zones" receive virtually no updates, often leading to localized generative collapse. In contrast, observation-aligned partitioning strategy maintains consistently positive query frequencies across the whole valid spatial locations. Our extensive practical experiments complements the marginal positivity guarantee of Theorem 2 by showing that the induced positive query exposure consistently maintains non-negligible magnitudes globally in practice, rather than merely non-zero in expectation. Consequently, all valid locations can receive effective supervision. This reliable update frequency effectively broadens the model's learning capacity across the physical field.

E.9 Effectiveness of Stochastic Anchor

To demonstrate the effectiveness of the stochastic anchor construction mechanism, we conduct a comparison analysis on $\rho = 1$ (i.e. do not use this mechanism), $\rho = 0.8$ and $\rho = 0.6$ respectively. As shown in Fig. 5 , we visualize the mean and variance of the mask conditionally generated by the BFN after ensemble sampling 200 times. It is easy to observe that without stochastic anchor

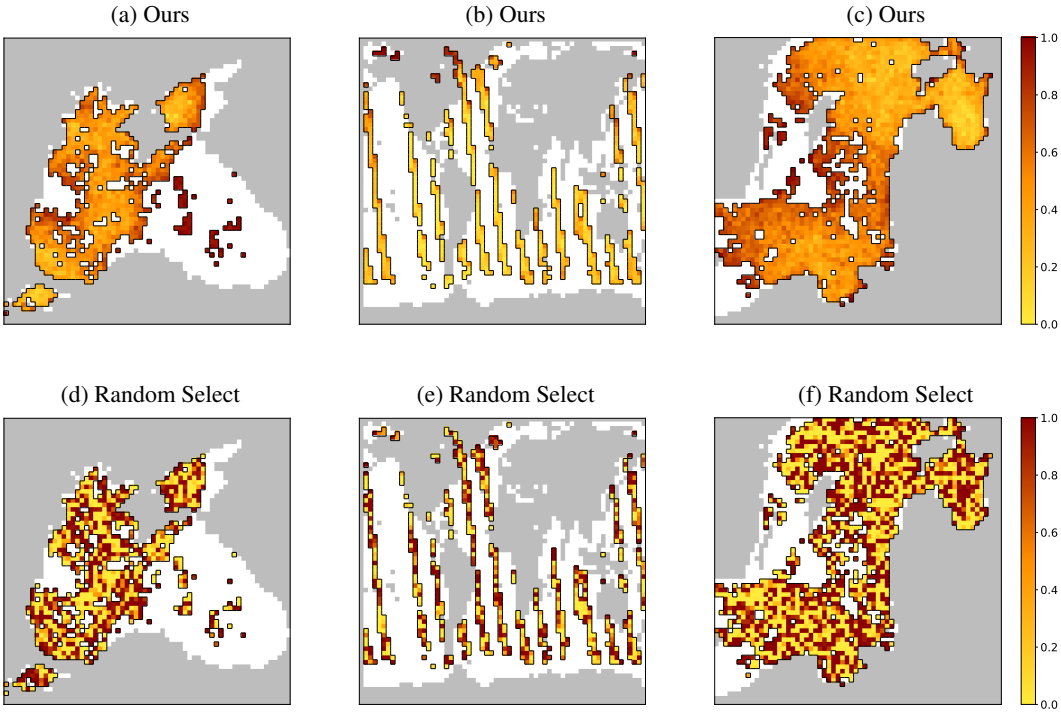


Figure 4: Comparison of spatial query probability distributions between our proposed method and naive random selection strategy across the Black Sea CHL, Global Ocean SSS, and Baltic Sea NANO datasets with the resolution of 64×64 (from left to right).

($p = 1.0$), the sampled masks exhibit minimal variance, which severely limits the spatial diversity of the samples. While, when p is set to 0.6, the proportion of the context region is reduced, resulting in insufficient information to guide the model effectively. Balancing the need for spatial diversity and sufficient context information, $p = 0.8$ provides a better trade-off, ensuring adequate mask variance while preserving enough information.

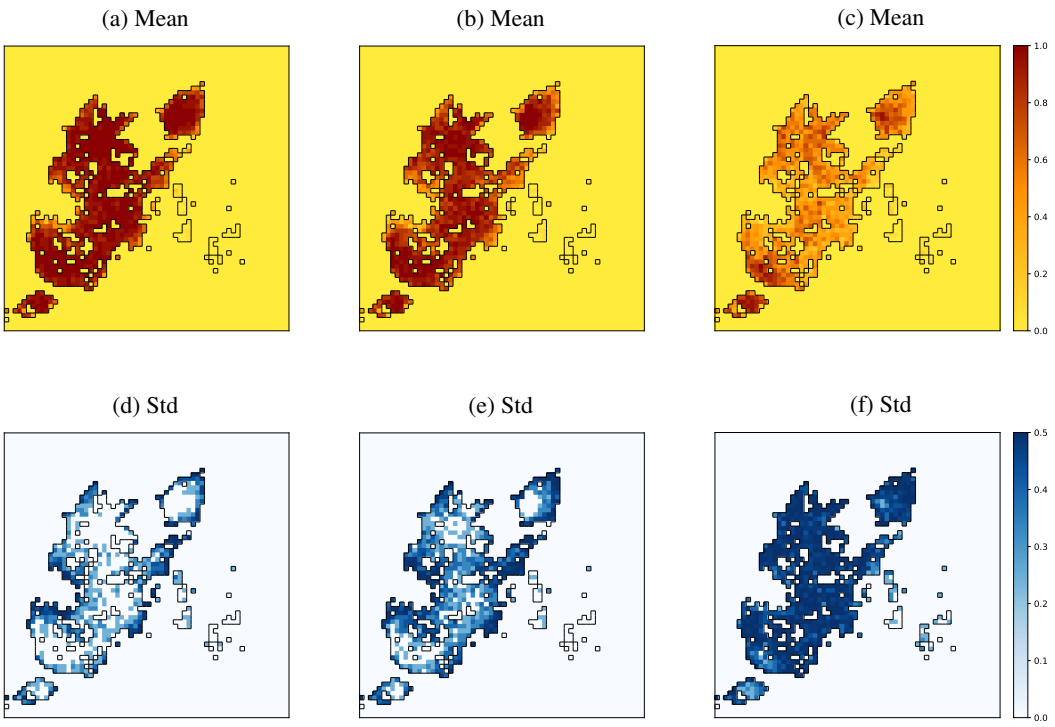


Figure 5: Comparison of different setting of ρ in conditional BFN sampling on the 64×64 Black Sea CHL dataset. The value of ρ is set to be 1, 0.8, 0.6 from left to right.

E.10 Visualization of Generated Samples

By statistically analyzing the masks within the dataset, we defined positions where no valid observations were ever recorded as "land", and represented them in gray during visualization to distinguish them from the unobserved fields, which are depicted in white. The areas enclosed by the red and black wireframes together constitute the observable region of the sample. During evaluation, the area enclosed by the black wireframe is the visible part provided to the model, which performs global reconstruction based on it. The red area is an artificially masked region used to evaluate the model's reconstruction performance. Here we provide visual comparison with the resolution of 64×64, 128×128 and 256×256 respectively in Fig. 6, Fig. 7 and Fig. 8

F Implementation of Context Mask Selection Strategies

We provide concrete algorithmic implementations of those context mask selection strategies mentioned in Sec. 4.5. Algorithm 3 presents the **Pixel-level Partition** strategy, where each point in the observation mask $\boldsymbol{M}$ is independently selected as context or query through Bernoulli sampling with ratios $r_{\text{ctx}} = 0.3$ and $r_{\text{qry}} = 0.3$ here. Algorithm 4 shows the **Block-wise Partition** strategy, where a subset of blocks is randomly sampled with ratios $r_{\text{ctx}} = 0.5$ and $r_{\text{qry}} = 0.5$ here, and their intersection with the observation mask $\boldsymbol{M}$ forms the context and query masks, respectively. Algorithm 5 outlines the **Saliency-driven Partition** strategy, a heuristic approach that allocates context points based on sorting local gradient magnitudes. Algorithm 6 details the **Empirical Distribution Partition** strategy,

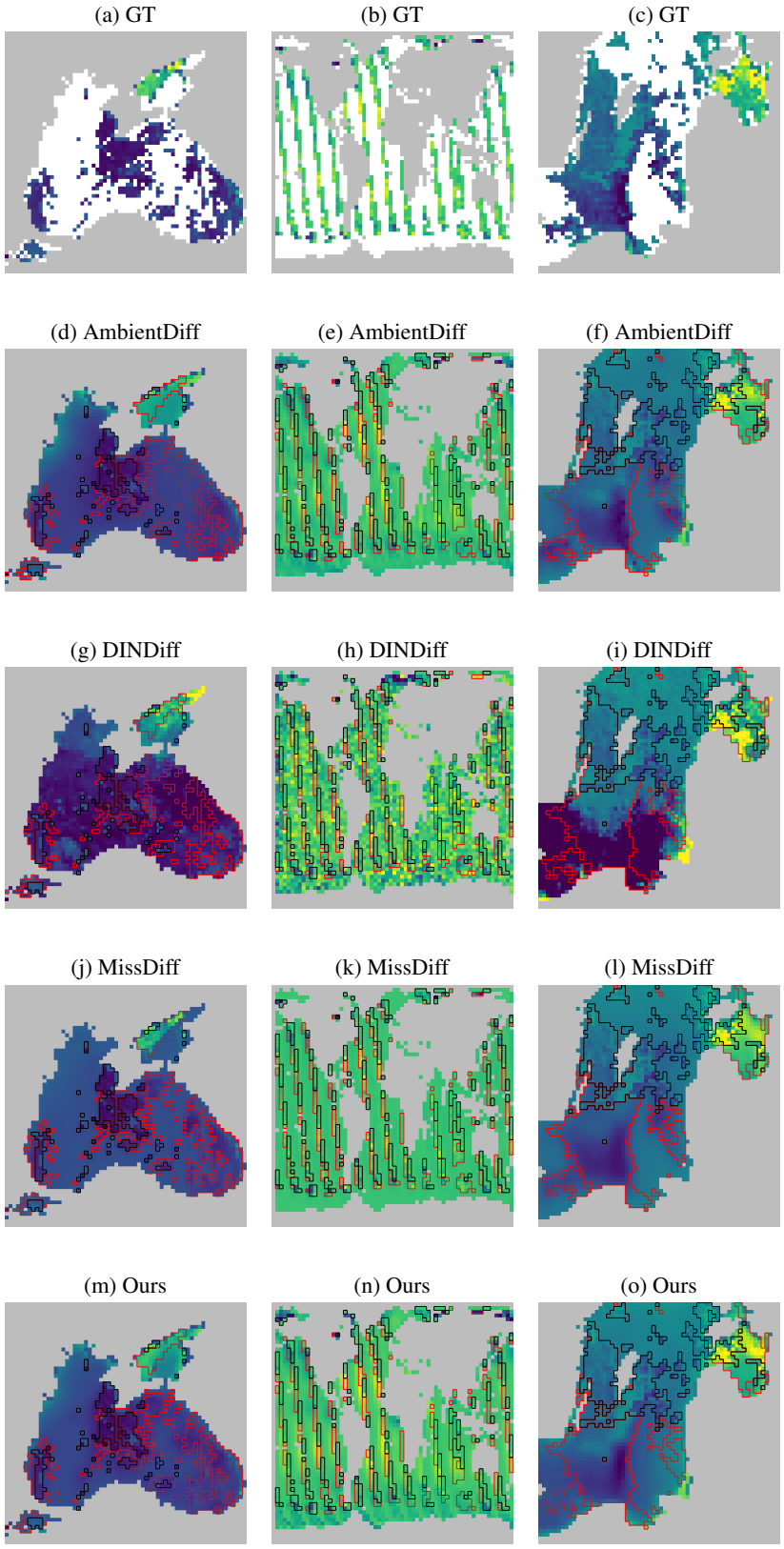


Figure 6: Visualization of imputation results on Black Sea CHL, Global Ocean SSS, and Baltic Sea NANO with the resolution of 64×64 (from left to right).

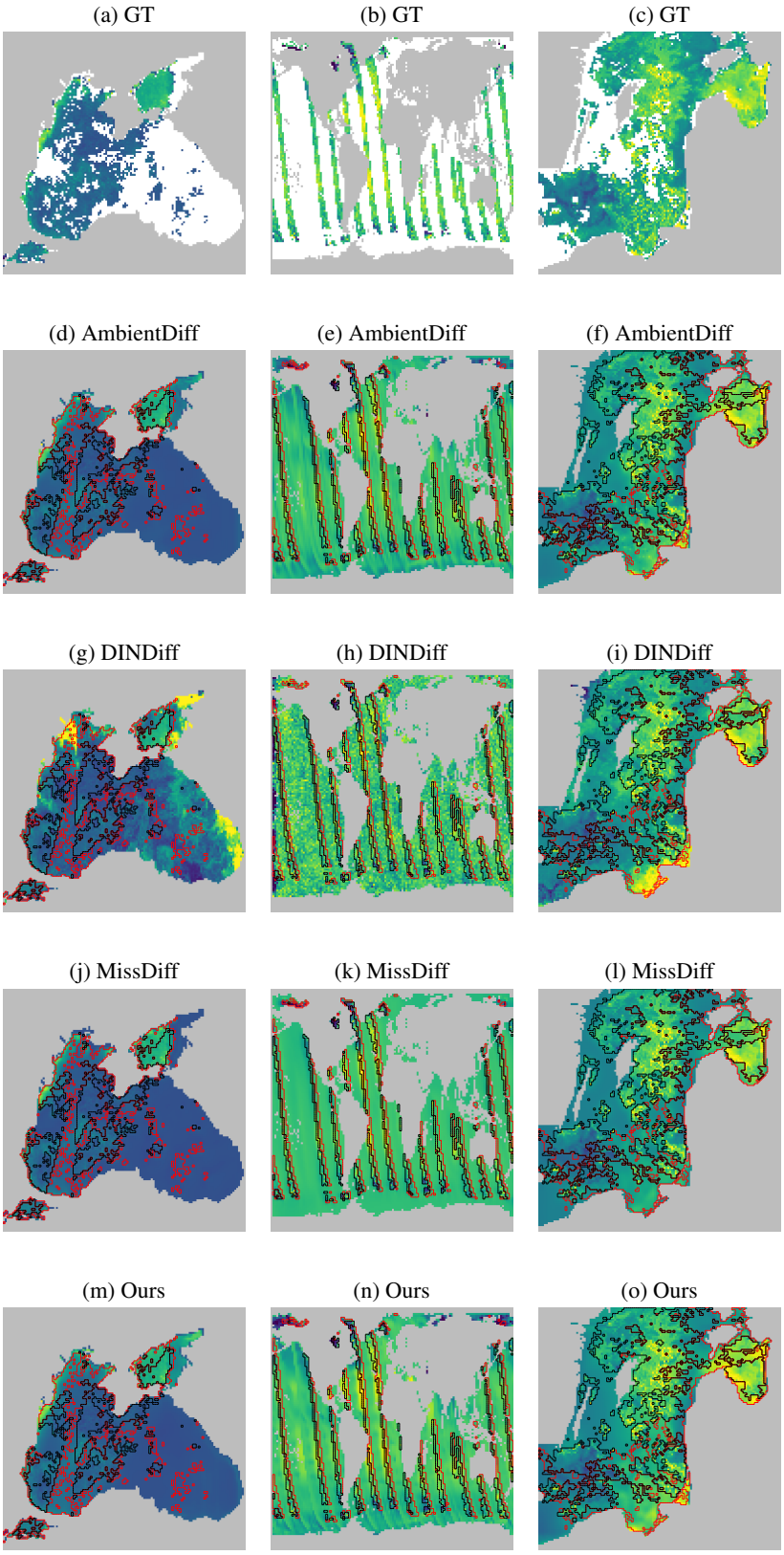


Figure 7: Visualization of imputation results on Black Sea CHL, Global Ocean SSS, and Baltic Sea NANO with the resolution of 128×128 (from left to right).

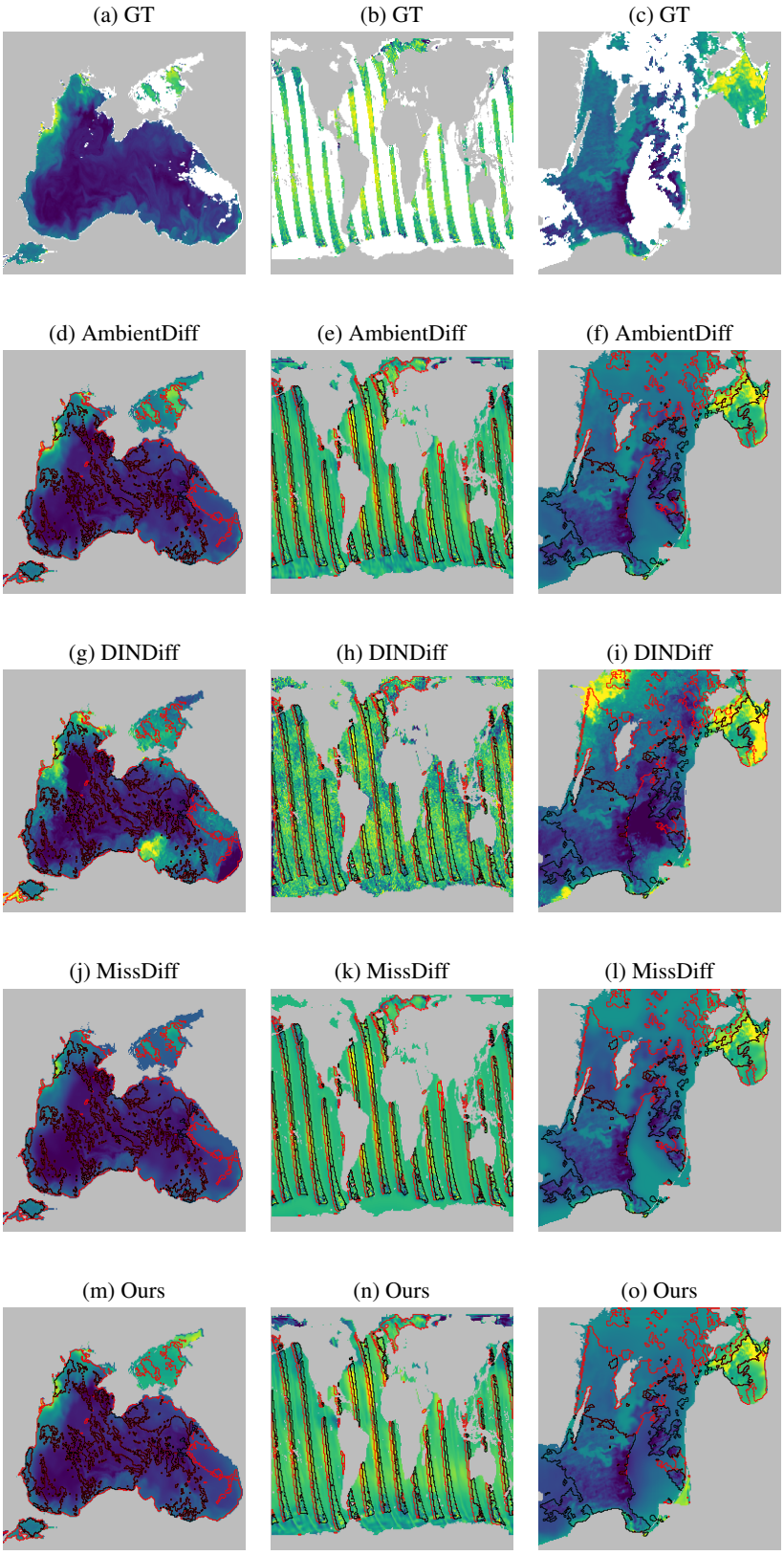


Figure 8: Visualization of imputation results on Black Sea CHL, Global Ocean SSS, and Baltic Sea NANO with the resolution of 256×256 (from left to right).

which samples a real observation mask from the dataset to intersect with the current observation. Algorithm 7 demonstrates the **Unconditional Prior Partition** strategy, employing an unconditionally sampled mask from a pre-trained BFN to intersect with the current observation. As for **Our Proposed Partition** strategy, we have provided the implementation in Algorithm 1 and Algorithm 2.

Algorithm 3 Pixel-Level Partition

Require: Observation mask $M \in \{0,1\}^d$, context ratio $r_{\text{ctx}} \in (0,1)$
Ensure: Context mask M_{ctx}, query mask M_{qry}
Initialize $M_{\text{ctx}} \leftarrow \mathbf{0}$
for each spatial index $i \in \{1, \ldots, d\}$ **do**
  if $M_i = 1$ **then**
    Sample $u \sim \text{Uniform}(0,1)$
    if $u < r_{\text{ctx}}$ **then**
      $(M_{\text{ctx}})_i \leftarrow 1$
    end if
  end if
end for
$M_{\text{qry}} \leftarrow M \odot (1 - M_{\text{ctx}})$
return $M_{\text{ctx}}, M_{\text{qry}}$

Algorithm 4 Block-Wise Partition

Require: Observation mask $M \in \{0,1\}^{H \times W}$, block grid dimension d, context ratio $r_{\text{ctx}} \in (0,1]$
Ensure: Context mask M_{ctx}, query mask M_{qry}
Initialize $M_{\text{ctx}} \leftarrow \mathbf{0}^{H \times W}$
Partition the $H \times W$ domain into a $d \times d$ grid of blocks $\{B_{i,j}\}_{i,j=1}^{d}$
$K \leftarrow d^2$
$N_{\text{ctx}} \leftarrow \lfloor r_{\text{ctx}} K \rfloor$
Uniformly sample a subset S_{ctx} of N_{ctx} blocks:

$$S_{\text{ctx}} \leftarrow \text{Sample}\big(\{B_{i,j}\}_{i,j=1}^{d}, N_{\text{ctx}}\big)$$

for each block $B_{i,j} \in S_{\text{ctx}}$ **do**
  $M_{\text{ctx}}[\text{region}(B_{i,j})] \leftarrow M[\text{region}(B_{i,j})]$
end for
$M_{\text{qry}} \leftarrow M \odot (1 - M_{\text{ctx}})$
return $M_{\text{ctx}}, M_{\text{qry}}$

Algorithm 5 Saliency-driven Partition

Require: observation mask $\boldsymbol{M} \in \{0,1\}^d$, observed image data $\boldsymbol{x}_{\text{obs}}$, context ratio $r_{\text{ctx}} \in (0,1)$
Ensure: context mask $\boldsymbol{M}_{\text{ctx}}$, query mask $\boldsymbol{M}_{\text{qry}}$
initialize $\boldsymbol{M}_{\text{ctx}} \leftarrow \mathbf{0}, \boldsymbol{M}_{\text{qry}} \leftarrow \mathbf{0}$
$\boldsymbol{S}_{\text{obs}} \leftarrow \{i \in \{1,\dots,d\} \mid M_i = 1\}$
$N_{\text{obs}} \leftarrow |\boldsymbol{S}_{\text{obs}}|$
$N_{\text{ctx}} \leftarrow \lfloor r_{\text{ctx}} \cdot N_{\text{obs}} \rfloor$
$N_{\text{half}} \leftarrow \lfloor N_{\text{ctx}}/2 \rfloor$
compute local gradient magnitudes $G_i = \|\nabla x_{\text{obs},i}\|$ for all $i \in \boldsymbol{S}_{\text{obs}}$
sort $\boldsymbol{S}_{\text{obs}}$ in descending order based on G_i into an ordered list L
$\boldsymbol{S}_{\text{high}} \leftarrow$ first half of elements in L
$\boldsymbol{S}_{\text{low}} \leftarrow$ second half of elements in L
randomly sample a subset $\boldsymbol{S}_{\text{ctx_high}} \subseteq \boldsymbol{S}_{\text{high}}$ of size N_{half}
randomly sample a subset $\boldsymbol{S}_{\text{ctx_low}} \subseteq \boldsymbol{S}_{\text{low}}$ of size $(N_{\text{ctx}} - N_{\text{half}})$
$\boldsymbol{S}_{\text{ctx}} \leftarrow \boldsymbol{S}_{\text{ctx_high}} \cup \boldsymbol{S}_{\text{ctx_low}}$
for each spatial index $i \in \{1,\dots,d\}$ **do**
  if $\boldsymbol{M}_i = 1$ **then**
    if $i \in \boldsymbol{S}_{\text{ctx}}$ **then**
      $(\boldsymbol{M}_{\text{ctx}})_i \leftarrow 1$
    else
      $(\boldsymbol{M}_{\text{qry}})_i \leftarrow 1$
    end if
  end if
end for
return $\boldsymbol{M}_{\text{ctx}}, \boldsymbol{M}_{\text{qry}}$

Algorithm 6 Empirical Distribution Partition

Require: observation mask $\boldsymbol{M} \in \{0,1\}^d$, dataset of real observation masks $\mathcal{D}_{\text{mask}}$
Ensure: context mask $\boldsymbol{M}_{\text{ctx}}$, query mask $\boldsymbol{M}_{\text{qry}}$
sample a real observation mask $\boldsymbol{M}_{\text{sample}} \sim \mathcal{D}_{\text{mask}}$
initialize $\boldsymbol{M}_{\text{ctx}} \leftarrow \mathbf{0}, \boldsymbol{M}_{\text{qry}} \leftarrow \mathbf{0}$
for each spatial index $i \in \{1,\dots,d\}$ **do**
  if $\boldsymbol{M}_i = 1$ **then**
    if $(\boldsymbol{M}_{\text{sample}})_i = 1$ **then**
      $(\boldsymbol{M}_{\text{ctx}})_i \leftarrow 1$
    else
      $(\boldsymbol{M}_{\text{qry}})_i \leftarrow 1$
    end if
  end if
end for
return $\boldsymbol{M}_{\text{ctx}}, \boldsymbol{M}_{\text{qry}}$

Algorithm 7 Unconditional Prior Partition

Require: observation mask $\boldsymbol{M} \in \{0,1\}^d$, pre-trained BFN model $\mathcal{B}_{\text{mask}}$
Ensure: context mask $\boldsymbol{M}_{\text{ctx}}$, query mask $\boldsymbol{M}_{\text{qry}}$
unconditionally sample a $M_{\text{sample}} \sim \mathcal{B}_{\text{mask}}$
initialize $\boldsymbol{M}_{\text{ctx}} \leftarrow \mathbf{0}, \boldsymbol{M}_{\text{qry}} \leftarrow \mathbf{0}$
for each spatial index $i \in \{1, \dots, d\}$ **do**
  if $\boldsymbol{M}_i = 1$ **then**
    if $(\boldsymbol{M}_{\text{sample}})_i = 1$ **then**
      $(\boldsymbol{M}_{\text{ctx}})_i \leftarrow 1$
    else
      $(\boldsymbol{M}_{\text{qry}})_i \leftarrow 1$
    end if
  end if
end for
return $\boldsymbol{M}_{\text{ctx}}, \boldsymbol{M}_{\text{qry}}$

G Implementation of Sampling Methods

We provide detailed algorithmic implementations of various sampling methods evaluated in Sec. 4.5. Inspired by the methodologies introduced in (Zhou et al., 2025), we first present three fundamental approaches. Algorithm 8 demonstrates the **Direct Projection** method, which performs a one-step projection to minimize global reconstruction error. Algorithm 9 details **Proximal Prediction** method, introducing a single step denoising process with an infinitesimal noise level to better align with the diffusion prior and mitigate blurriness. Algorithm 10 details the **Iterative Conditioning** method. Notably, to heuristically blend the diffusion expectation and imputation expectation, we employ a simple linear schedule ($\omega_s = s$) as the time-varying weight. Furthermore, drawing inspiration from the resampling mechanisms proposed in RePaint (Lugmayr et al., 2022), we present two advanced sampling strategies. Algorithm 11 describes the implementation of **Resampling Inpainting** method, which harmonizes observation boundaries by alternating between denoising and stochastic resampling steps. Algorithm 12 demonstrates the **Recursive Jump Inpainting** method, which utilizes profound temporal jumps to re-inject significant variance, effectively reconciling global physical dynamics with sparse observation constraints.

Algorithm 8 Direct Projection

Require: partially observed data $\boldsymbol{x}_{\text{obs}}$, mask $\boldsymbol{M}$, trained model $\boldsymbol{u}_{\boldsymbol{\phi}}$, ensemble size K
Ensure: imputed complete data $\boldsymbol{x}_0$
Generate K random context masks $\{\boldsymbol{M}_{\text{ctx}}^{(k)}\}_{k=1}^{K} \subseteq \boldsymbol{M}$
$\boldsymbol{x}^* \leftarrow \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{u}_{\boldsymbol{\phi}}\left(\boldsymbol{M}_{\text{ctx}}^{(k)} \odot \boldsymbol{x}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}^{(k)}\right)$
$\boldsymbol{x}_0 \leftarrow \boldsymbol{M} \odot \boldsymbol{x}_{\text{obs}} + (1 - \boldsymbol{M}) \odot \boldsymbol{x}^*$
return $\boldsymbol{x}_0$

Algorithm 9 Proximal Prediction

Require: partially observed data $\boldsymbol{x}_{\text{obs}}$, mask $\boldsymbol{M}$, trained model $\boldsymbol{u}_{\boldsymbol{\phi}}$, minimal timestep $\delta \ll 1$, ensemble size K
Ensure: imputed complete data $\boldsymbol{x}_0$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
$\boldsymbol{x}_{\text{obs},\delta} \leftarrow \alpha_\delta \boldsymbol{x}_{\text{obs}} + \sigma_\delta \boldsymbol{\epsilon}$
Generate K random context masks $\{\boldsymbol{M}_{\text{ctx}}^{(k)}\}_{k=1}^{K} \subseteq \boldsymbol{M}$
$\boldsymbol{x}_\delta \leftarrow \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{u}_{\boldsymbol{\phi}}\left(\delta, \boldsymbol{M}_{\text{ctx}}^{(k)} \odot \boldsymbol{x}_{\text{obs},\delta}, \boldsymbol{M}_{\text{ctx}}^{(k)}\right)$
$\boldsymbol{x}_0 \leftarrow \boldsymbol{M} \odot \boldsymbol{x}_{\text{obs}} + (1 - \boldsymbol{M}) \odot \boldsymbol{x}_\delta$
return $\boldsymbol{x}_0$

Algorithm 10 Iterative Conditioning

Require: partially observed data $\boldsymbol{x}_{\text{obs}}$, mask $\boldsymbol{M}$, trained model $\boldsymbol{u}_{\boldsymbol{\phi}}$, ensemble size K, weight schedule $\omega_s = s$
Ensure: imputed complete data $\boldsymbol{x}_0$

Generate K random context masks $\{\boldsymbol{M}_{\text{ctx}}^{(k)}\}_{k=1}^{K} \subseteq \boldsymbol{M}$
$E_{\text{imp}} \leftarrow \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{u}_{\boldsymbol{\phi}}\left(\boldsymbol{M}_{\text{ctx}}^{(k)} \odot \boldsymbol{x}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}^{(k)}\right)$
initialize: $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
for diffusion steps from s to t **do**
  sample a $\boldsymbol{M}_{\text{ctx}}$
  $E_{\text{diff}} \leftarrow \boldsymbol{u}_{\phi}\left(s, \boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{x}_s, \boldsymbol{M}_{\text{ctx}}\right)$
  $\hat{\boldsymbol{x}}_0 \leftarrow \omega_s E_{\text{diff}} + (1-\omega_s) E_{\text{imp}}$
  $\boldsymbol{\epsilon}_{\text{unobs}} \leftarrow \frac{\boldsymbol{x}_s - \alpha_s \hat{\boldsymbol{x}}_0}{\sigma_s}$
  $\boldsymbol{\epsilon}_{\text{obs}} \leftarrow \frac{\boldsymbol{x}_s - \alpha_s \boldsymbol{x}_{\text{obs}}}{\sigma_s}$
  $\boldsymbol{\epsilon}_{\text{full}} \leftarrow \boldsymbol{M} \odot \boldsymbol{\epsilon}_{\text{obs}} + (1-\boldsymbol{M}) \odot \boldsymbol{\epsilon}_{\text{unobs}}$
  $\boldsymbol{x}_t \leftarrow \text{DiffusionODE}(s, t, \boldsymbol{x}_s, \boldsymbol{\epsilon}_{\text{full}})$
end for
$\boldsymbol{x}_0 \leftarrow \boldsymbol{M} \odot \boldsymbol{x}_{\text{obs}} + (1-\boldsymbol{M}) \odot \boldsymbol{x}_t$
return $\boldsymbol{x}_0$

Algorithm 11 Resampling Inpainting

Require: partially observed data $\boldsymbol{x}_{\text{obs}}$, mask $\boldsymbol{M}$, trained model $\boldsymbol{u}_{\boldsymbol{\phi}}$, ensemble size K, weight schedule $\omega_s = s$, jump length j, jump frequency f
Ensure: imputed complete data $\boldsymbol{x}_0$

Generate K random context masks $\{\boldsymbol{M}_{\text{ctx}}^{(k)}\}_{k=1}^{K} \subseteq \boldsymbol{M}$
$E_{\text{imp}} \leftarrow \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{u}_{\phi}\left(\boldsymbol{M}_{\text{ctx}}^{(k)} \odot \boldsymbol{x}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}^{(k)}\right)$
initialize: $\boldsymbol{x}_T \sim \mathcal{N}(0, I)$
$s \leftarrow T$
$c \leftarrow 0$
while $s > 0$ **do**
  $t \leftarrow s - 1$
  sample a $\boldsymbol{M}_{\text{ctx}}$
  $E_{\text{diff}} \leftarrow \boldsymbol{u}_{\boldsymbol{\phi}}(s, \boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{x}_s, \boldsymbol{M}_{\text{ctx}})$
  $\hat{x}_0 \leftarrow \omega_s E_{\text{diff}} + (1-\omega_s) E_{\text{imp}}$
  $\boldsymbol{\epsilon}_{\text{unobs}} \leftarrow \frac{\boldsymbol{x}_s - \alpha_s \hat{\boldsymbol{x}}_0}{\sigma_s}$
  $\boldsymbol{\epsilon}_{\text{obs}} \leftarrow \frac{\boldsymbol{x}_s - \alpha_s \boldsymbol{x}_{\text{obs}}}{\sigma_s}$
  $\boldsymbol{\epsilon}_{\text{full}} \leftarrow \boldsymbol{M} \odot \boldsymbol{\epsilon}_{\text{obs}} + (1-\boldsymbol{M}) \odot \boldsymbol{\epsilon}_{\text{unobs}}$
  $\boldsymbol{x}_t \leftarrow \text{DiffusionODE}(s, t, \boldsymbol{x}_s, \boldsymbol{\epsilon}_{\text{full}})$
  $c \leftarrow c + 1$
  if $c \,\%\, f == 0$ **and** $t > 0$ **then**
    $s_{\text{new}} \leftarrow \min(t + j, T)$
    $\boldsymbol{x}_{s_{\text{new}}} \leftarrow \text{ForwardDiffusion}(t, s_{\text{new}}, \boldsymbol{x}_t)$
    $s \leftarrow s_{\text{new}}$
  else
    $s \leftarrow t$
  end if
end while
$\boldsymbol{x}_0 \leftarrow \boldsymbol{M} \odot \boldsymbol{x}_{\text{obs}} + (1-\boldsymbol{M}) \odot \boldsymbol{x}_t$
return $\boldsymbol{x}_0$

Algorithm 12 Recursive Jump Inpainting

Require: partially observed data $\boldsymbol{x}_{\text{obs}}$, mask $\boldsymbol{M}$, trained model $\boldsymbol{u}_{\boldsymbol{\phi}}$, ensemble size K, weight schedule $\omega_s = s$, total jump stages n

Ensure: imputed complete data $\boldsymbol{x}_0$

Generate K random context masks $\{\boldsymbol{M}_{\text{ctx}}^{(k)}\}_{k=1}^{K} \subseteq \boldsymbol{M}$
$E_{\text{imp}} \leftarrow \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{u}_{\boldsymbol{\phi}}\left(\boldsymbol{M}_{\text{ctx}}^{(k)} \odot \boldsymbol{x}_{\text{obs}}, \boldsymbol{M}_{\text{ctx}}^{(k)}\right)$
initialize: $\boldsymbol{x}_T \sim \mathcal{N}(0, I)$
$s \leftarrow T$
$k \leftarrow n - 1$
while $s > 0$ **do**
  $t \leftarrow s - 1$
  sample a $\boldsymbol{M}_{\text{ctx}}$
  $E_{\text{diff}} \leftarrow \boldsymbol{u}_{\boldsymbol{\phi}}(s, \boldsymbol{M}_{\text{ctx}} \odot \boldsymbol{x}_s, \boldsymbol{M}_{\text{ctx}})$
  $\hat{\boldsymbol{x}}_0 \leftarrow \omega_s E_{\text{diff}} + (1 - \omega_s) E_{\text{imp}}$
  $\boldsymbol{\epsilon}_{\text{unobs}} \leftarrow \frac{\boldsymbol{x}_s - \alpha_s \hat{\boldsymbol{x}}_0}{\sigma_s}$
  $\boldsymbol{\epsilon}_{\text{obs}} \leftarrow \frac{\boldsymbol{x}_s - \alpha_s \boldsymbol{x}_{\text{obs}}}{\sigma_s}$
  $\boldsymbol{\epsilon}_{\text{full}} \leftarrow \boldsymbol{M} \odot \boldsymbol{\epsilon}_{\text{obs}} + (1 - \boldsymbol{M}) \odot \boldsymbol{\epsilon}_{\text{unobs}}$
  $\boldsymbol{x}_t \leftarrow \text{DiffusionODE}(s, t, \boldsymbol{x}_s, \boldsymbol{\epsilon}_{\text{full}})$
  if $t == 0$ **and** $k > 0$ **then**
    $T_{\text{jump}} \leftarrow \lfloor \frac{k}{n} T \rfloor$
    $\boldsymbol{x}_{T_{\text{jump}}} \leftarrow \text{ForwardDiffusion}(0, T_{\text{jump}}, \boldsymbol{x}_t)$
    $s \leftarrow T_{\text{jump}}$
    $k \leftarrow k - 1$
  else
    $s \leftarrow t$
  end if
end while
$\boldsymbol{x}_0 \leftarrow \boldsymbol{M} \odot \boldsymbol{x}_{\text{obs}} + (1 - \boldsymbol{M}) \odot \boldsymbol{x}_t$
return $\boldsymbol{x}_0$