PHYDIM: PHYSICS-GUIDED DIFFUSION MODEL FOR SPATIO-TEMPORAL SATELLITE IMAGE RECONSTRUCTION UNDER SPARSE OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing continuous spatio-temporal images from sparse and irregular visual observations is a challenging machine learning problem due to complex spatial dependencies and strong temporal dynamics that must be inferred from a limited number of temporal observations. In this setting, incorporating physical priors is particularly beneficial, as it enables the reconstructed images to respect underlying physical laws and produce more realistic and temporally consistent dynamics. Recent generative approaches, particularly diffusion models, have demonstrated good performance in different reconstruction tasks, but they remain largely data-driven and lack explicit physical structure to guide the generation process. Existing attempts treat physical constraints as external regularization terms in the loss function, which can lead to unstable optimization. In this paper, we introduce PhyDiM, a physics-guided diffusion model that incorporates partial differential equation (PDE) priors directly into the diffusion process. Specifically, we embed the physical constraints directly into both the forward diffusion process and the reverse sampling dynamics. PhyDiM is applied to land surface temperature reconstruction from sparsely sampled satellite observations, where the PDE is instantiated as a heat equation. Experimental results on large-scale and climatically diverse regions covering France, Spain, Germany, Belgium, and the Netherlands demonstrate that PhyDiM consistently outperforms both state-of-the-art reconstruction and diffusion-based methods, particularly at extreme diurnal time steps where existing approaches tend to degrade. Moreover, qualitative analysis further demonstrates that PhyDiM preserves sharper spatial structures and more realistic thermal patterns across diverse conditions.

1 INTRODUCTION

Spatio-temporal image reconstruction is a core problem in machine learning that aims to recover complete representations of a scene over space and time from sparse observations, with applications in video frame interpolation (VFI) (Bao et al., 2019; Shi et al., 2022; Wu et al., 2023), medical imaging (Schlemper et al., 2017; Qin et al., 2018; Baik & Yoo, 2025), and satellite remote sensing (Vandal & Nemani, 2021; Marisca et al., 2022; Shu et al., 2025). The problem is particularly challenging when observations are irregularly and sparsely sampled, making it difficult to recover temporally consistent representations of dynamic processes (Kye et al., 2026). This limitation is especially pronounced in satellite remote sensing, where orbital constraints and sensor revisit times inherently limit temporal coverage (Bouaziz et al., 2026a; Gibson, 2020), despite the need to monitor rapidly evolving phenomena such as urban heat and extreme weather events (Zhao et al., 2025; Xu et al., 2025). For example, the MODIS sensors onboard the Terra and Aqua satellites provide Land Surface Temperature (LST) observations at approximately four fixed times per day.

A wide range of approaches have been explored for spatio-temporal reconstruction. Classical methods rely on linear or spline-based interpolation (Lehmann et al., 2001; Li & Revesz, 2004; Zhou et al., 2016), which assume smooth temporal evolution and often fail to capture complex non-linear dynamics. Frequency-domain approaches use Fourier and spectral representations to model periodic or quasi-periodic evolution (Meyer et al., 2015; Khormizi et al., 2025; Meyer et al., 2018), but remain limited when observations are highly sparse or dynamics deviate from strict periodicity. Physics-based approaches instead leverage partial differential equations (PDEs) to describe con-

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.

tinuous system evolution (Ferziger et al., 2002; Bergman, 2011; Vallis, 2017). They offer strong interpretability by explicitly encoding conservation laws (LeVeque, 2002; Evans, 2022), but rely on restrictive assumptions and careful tuning, limiting their applicability in complex and heterogeneous environments with unknown or partially observed governing equations (Chung, 2010).

Recently, deep learning approaches have advanced spatio-temporal reconstruction, particularly VFI, using convolutional (Ilg et al., 2017; Sun et al., 2018), recurrent (Shen et al., 2020; Schatz et al., 2020), and transformer-based architectures (Lu et al., 2022; Zhang et al., 2025). These methods learn complex transformations directly from data but struggle with large temporal gaps and irregular sampling. Diffusion models further improve generative reconstruction through iterative denoising (Voleti et al., 2022; Jain et al., 2024; Danier et al., 2024), enabling high-fidelity synthesis and improved uncertainty modeling. However, they remain largely data-driven and rely on isotropic Gaussian noise assumptions (Vlassis & Sun, 2023; Li et al., 2023; Lienen et al., 2023), neglecting spatial coupling and structured temporal dynamics (Yuan et al., 2023; Bastek et al., 2024). This can limit their ability to preserve physically coherent dynamics in scientific applications such as remote sensing, where observations evolve jointly across space and time according to underlying physical processes. Recent physics-guided diffusion approaches incorporate such knowledge through additional loss terms enforcing governing PDEs during training (Bastek et al., 2025; Soni et al., 2025). However, these formulations introduce additional optimization objectives that complicate and destabilize training, while requiring careful balancing between data fidelity and physical constraints.

To address these limitations, we propose **PhyDiM**, a physics-informed generative model based on conditional diffusion models for continuous spatio-temporal image reconstruction, applied to hourly LST generation from sparse MODIS Terra and Aqua observations (see Fig. 1). Our approach rethinks how physical knowledge can be injected into diffusion models by reformulating both the forward corruption process and the sampling strategy to better align with the underlying dynamics governing the observed images. Specifically, we introduce a two-stage corruption mechanism, where an initial physics-driven evolution captures spatial interactions through a heat-diffusion-inspired process, followed by a stochastic perturbation stage. In addition, we incorporate physics-guided sampling during inference, where physical consistency is reinforced through a constraint on the generated image. Furthermore, we design a conditional denoising architecture that leverages multiple temporal observations as context to guide the reconstruction of both intermediate states and boundary time steps. The model encodes spatial features and temporal metadata into a unified representation and employs cross-attention mechanisms to dynamically integrate information across different time steps, which facilitates the fusion of local spatial structure with long-range temporal dependencies. Our contributions are summarized as follows:

- We propose a physics-aware diffusion formulation that embeds a heat-diffusion PDE into the forward process, enabling structured physical evolution during corruption.
- We design a conditional diffusion architecture with cross-attention over multi-temporal observations for flexible reconstruction across time.
- We introduce a physics-guided sampling strategy for physically consistent generation.

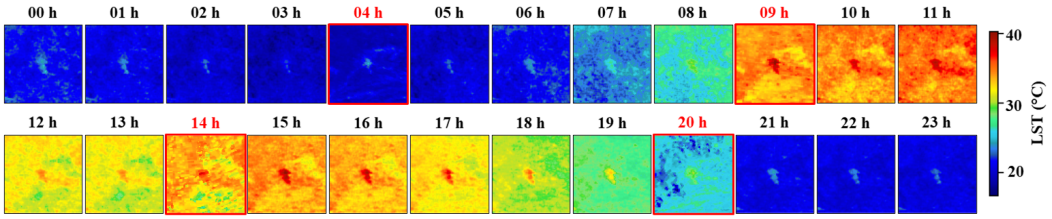


Figure 1: Hourly LST reconstruction over the Orléans metropolitan area, France, on 24 June 2026, the hottest day ever recorded in France (30.0°C national 24 h mean). Red-bordered frames are the four MODIS observations, while the remaining hourly maps are reconstructed by PhyDiM.

2 Related Works

Denoising Diffusion Probabilistic Models (DDPMs) are a class of generative models that learn data distributions through a forward gradual corruption process and its learned reversal (Sohl-Dickstein et al., 2015; Ho et al., 2020; Croitoru et al., 2023). The forward process progressively perturbs a clean sample $\mathbf{x}_0$ by adding Gaussian noise over T timesteps using a predefined variance

schedule $\{\beta_t\}_{t=1}^T$, resulting in a sequence of noisy variables $\{\mathbf{x}_t\}_{t=1}^T$:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{1}$$

here, $\mathbf{I}$ is the identity matrix, and $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes a Gaussian distribution with mean μ and covariance Σ. This recursive process admits a closed-form expression for $\mathbf{x}_t$ given $\mathbf{x}_0$:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^t (1-\beta_i)$. The reverse process learns to recover clean samples by predicting the added noise and iteratively denoising from Gaussian noise. For structured prediction, conditional diffusion models incorporate context $\mathbf{c}$, such as neighboring frames or observations, and model the reverse process as $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})$ (Zhang et al., 2023; Hu et al., 2023; Luo et al., 2024).

Physics-based modeling of spatio-temporal fields relies on PDEs to describe the evolution of continuous physical processes over space and time (Ferziger et al., 2002; Bergman, 2011; Vallis, 2017). The heat equation is a classic example of such PDEs that models diffusion-driven dynamics where each point evolves according to local spatial gradients (Widder, 1976), as follows:

$$\frac{\partial x}{\partial t} = \kappa \nabla^2 x, \tag{3}$$

where κ is the diffusion coefficient and ∇^2 is the Laplacian operator capturing local spatial interactions. This formulation is widely used in geophysical and environmental modeling, including LST dynamics in remote sensing (Asadzadeh & Souza Filho, 2023). For discrete images, the Laplacian is approximated using finite-difference stencils over local pixel neighborhoods, yielding a diffusion operator that iteratively propagates information between neighboring pixels. This produces local smoothing that suppresses high-frequency variations while preserving coarse spatial structure, corresponding to an explicit Euler discretization of the heat equation (LeVeque, 2007).

3 METHODOLOGY

As outlined in the introduction, PhyDiM is designed for hourly LST reconstruction from sparse satellite observations. It integrates physical knowledge into both the forward diffusion and reverse sampling processes, while conditioning the reconstruction on multi-temporal observations. The overall framework is illustrated in Fig. 2, and each component is detailed in the following sections.

3.1 PROBLEM FORMULATION

We consider the task of continuous spatio-temporal LST reconstruction from sparse satellite observations. Specifically, we use MODIS Terra and Aqua measurements, which provide four LST acquisitions per day at fixed and irregularly spaced timestamps $\{\mathbf{x}(t_i)\}_{i=1}^4$ (e.g., around 03:00, 10:00, 14:00, and 21:00), resulting in large temporal gaps. Given a set of observed LST $\{\mathbf{x}(t_i)\}_{i=1}^3$ acquired at known timestamps $\{t_i\}_{i=1}^3$, the objective is to reconstruct the continuous LST ($\mathbf{x}(t^*)$) at an arbitrary target time ($t^* \in [0, 24]$), conditioned on both the observed measurements and the target time. This formulates the task as a conditional spatio-temporal generation problem:

$$p\big(\mathbf{x}(t^*) \mid \{\mathbf{x}(t_i), t_i\}_{i=1}^3, t^*\big). \tag{4}$$

We formulate training as a randomized temporal reconstruction task. For each sample, a target time t^* is randomly selected from $\{t_i\}_{i=1}^4$, and the corresponding observation is reconstructed from the remaining measurements. The target time is fixed across epochs for stable optimization while maintaining temporal diversity across the dataset.

Beyond standard spatio-temporal reconstruction, our formulation supports both temporal interpolation and extrapolation, enabling reconstruction of past observations (e.g., $\mathbf{x}(t_1)$ from $\{\mathbf{x}(t_i)\}_{i=2}^4$) and prediction of future states (e.g., $\mathbf{x}(t_4)$ from $\{\mathbf{x}(t_i)\}_{i=1}^3$). Thus, it extends conventional VFI by unifying backward and forward temporal prediction.

3.2 PHYSICS-INFORMED FORWARD DIFFUSION

Standard diffusion models rely on an isotropic Gaussian corruption process that progressively destroys structure by injecting independent noise at each spatial location (Vlassis & Sun, 2023; Li et al., 2023; Lienen et al., 2023). While effective for natural image generation, this assumption neglects the fact that many spatio-temporal physical fields evolve according to structured dynamical laws governed by spatial interactions (Yuan et al., 2023; Bastek et al., 2024).

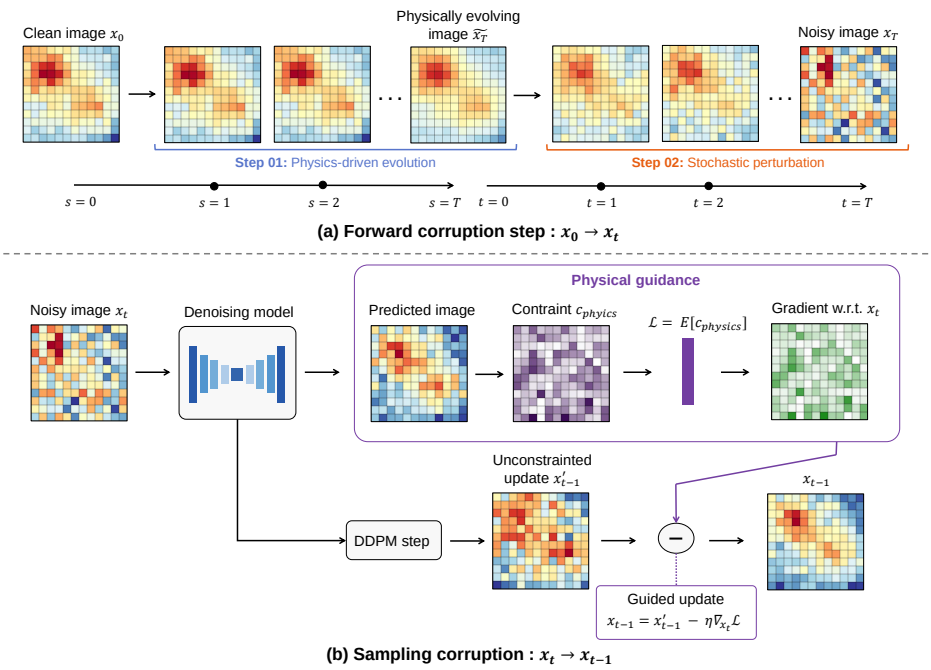


Figure 2: Overview of PhyDiM. The forward process combines physics-driven evolution with stochastic perturbation, while the reverse process uses physics-guided gradient corrections during denoising. Here, t is the diffusion timestep and T the total number of steps.

To address this mismatch, we reformulate the forward diffusion process by introducing a physics-driven evolution before stochastic perturbation. Specifically, the clean field is first evolved through a deterministic drift modeled by a physical process. In our setting, this drift is defined by the heat PDE (Eq. 3), which induces a spatially consistent smoothing effect aligned with thermal propagation dynamics in satellite-derived LST images. This continuous evolution is discretized using an explicit Euler time-stepping scheme, where Δt controls the magnitude of each update (LeVeque, 2007). The resulting state is then perturbed with standard Gaussian noise:

$$\tilde{\mathbf{x}}_t = \mathbf{x}_0 + \sum_{s=1}^{t} \kappa \nabla^2 \mathbf{x}_{s-1} \Delta t, \qquad \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\, \tilde{\mathbf{x}}_t + \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{\epsilon} \tag{5}$$

Moreover, the physics-guided evolution acts as a structure-preserving regularization that suppresses non-physical fluctuations while maintaining large-scale coherent spatial patterns. This is particularly relevant for MODIS LST, which exhibits retrieval errors of approximately 2°C (Duan et al., 2019). Fig. 1(a) illustrates the proposed two-stage forward process. During stochastic perturbation, we use the diffusion process's closed-form solution to directly sample $\mathbf{x}_t$ from the physically evolved field $\tilde{\mathbf{x}}_t$, as described in Section 2. The procedure is summarized in Algorithm 1.

Algorithm 1 Physics-Informed Forward Diffusion Process. The highlighted components denote the physics-driven evolution introduced in PhyDiM compared to standard DDPM.

Input: Clean field $\mathbf{x}_0$, diffusion timestep t, physical parameter κ, time step Δt, noise schedule $\{\bar{\alpha}_t\}$
Output: Corrupted sample $\mathbf{x}_t$
Initialize $\mathbf{x_s} \leftarrow \mathbf{x}_0$
for $s = 1$ to t **do**
  Compute Laplacian: $\mathbf{L} \leftarrow \nabla^2 \mathbf{x}$
  Compute drift: $\mathbf{d} \leftarrow \kappa \mathbf{L}$
  Update field: $\mathbf{x_s} \leftarrow \mathbf{x_s} + \Delta t \cdot \mathbf{d}$
end for
Sample noise: $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$
Construct final sample: $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\, \mathbf{x_s} + \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{\epsilon}$
return $\mathbf{x}_t$

3.3 CONDITIONAL DIFFUSION MODEL

We parameterize the noise prediction network $\boldsymbol{\epsilon}_\theta$ with a U-Net architecture conditioned on both spatial observations and temporal information. The model takes a noisy sample $\mathbf{x}_t$, a set of context images $\{\mathbf{x}(t_i)\}_{i=1}^3$ with their associated timestamps $\{t_i\}_{i=1}^3$, and a target time t^*. Conditioning is incorporated through context encoding and cross-attention.

Context encoding. Each observation $\mathbf{x}(t_i)$ and its acquisition time t_i are encoded into a latent representation. Spatial features are extracted via convolution and global average pooling, while t_i is encoded using a sinusoidal embedding followed by a multilayer perceptron (MLP). The features are concatenated and projected into a shared latent space via a learnable mapping $\phi(\cdot)$, resulting in context tokens, $\mathbf{c}_i = \phi\big(\mathbf{x}(t_i), t_i\big), i \in \{1, 2, 3\}$. The resulting tokens form the context matrix $\mathbf{C} \in \mathbb{R}^{3\times d}$, where d is the latent dimension.

Temporal conditioning. The target time t^* is encoded using a sinusoidal embedding and MLP, then added to each context token for continuous-time prediction. The diffusion timestep t is similarly encoded and injected into each convolutional block via additive feature-wise conditioning.

Cross-attention conditioning. Cross-attention layers are inserted into each convolutional block to incorporate contextual observations. Given feature maps $\mathbf{h} \in \mathbb{R}^{C_f\times H\times W}$, where C_f is the channel dimension and $H \times W$ the spatial resolution, spatial features are flattened and projected to queries $\mathbf{Q} \in \mathbb{R}^{HW\times d}$. The $N = 3$ context tokens $\mathbf{C} \in \mathbb{R}^{N\times d}$ are projected to keys and values $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times d}$. This enables each spatial location to attend to relevant observations across time.

U-Net backbone. The network adopts a multi-scale encoder-decoder architecture with skip connections. The input consists of the concatenation of the noisy sample $\mathbf{x}_t$ and the encoded context representation $\mathbf{C}$. The encoder progressively downsamples spatial features through convolutional residual blocks, each comprising two convolutional layers with Group Normalization, SiLU activations, residual connections, and diffusion-timestep conditioning via additive bias terms.

Bottleneck and decoder. Self-attention at the bottleneck captures long-range spatial dependencies. The decoder progressively upsamples features using transposed convolutions and combines them with corresponding encoder features via skip connections. Time conditioning and cross-attention are applied at each stage to maintain consistency with the conditioning inputs.

Training. The model is trained with the standard noise-prediction objective. Although the physics-driven evolution modifies the marginal distribution of the noisy states, the conditional corruption remains Gaussian with a deterministic mean, yielding the same conditional score parameterization as standard diffusion. Thus, the noise-prediction objective remains valid for the physics-modified forward process. A formal derivation is provided in Appendix A.

3.4 PHYSICS-GUIDED SAMPLING VIA DIFFUSION PRIORS

Standard diffusion sampling relies on learned score functions to iteratively denoise a Gaussian prior toward the data distribution. However, for physical spatio-temporal reconstruction, plausible solutions must also satisfy underlying physical consistency constraints. Inspired by diffusion posterior sampling (Chung et al., 2022), we define a physics-based energy functional over the predicted clean image $\mathbf{x}_0$ to penalize violations of LST spatial smoothness:

$$\mathcal{E}(\mathbf{x}_0) = \left\|\nabla^2 \mathbf{x}_0\right\|_2^2. \tag{6}$$

During sampling, this energy is incorporated through a gradient-based guidance term applied intermittently along the reverse diffusion trajectory. The resulting update combines the standard score-based denoising step with a physics-driven correction, yielding:

$$\mathbf{x} \leftarrow \mathbf{x} - \lambda \nabla_{\mathbf{x}} \mathcal{E}(\mathbf{x}_0), \tag{7}$$

where λ controls the strength of the physics guidance. Fig. 1(b) illustrates the proposed sampling strategy, and Algorithm 2 summarizes the overall procedure. In practice, applying the physics-based constraint at every denoising timestep is computationally expensive. Therefore, we apply this guidance at regular intervals, every 50 timesteps, which yields comparable results while reducing computational cost. The distinction between the physics-guided forward process and sampling-time guidance is intentional. During training, heat-diffusion dynamics are incorporated into the forward process to learn a PDE-structured corruption trajectory. At inference, we use only a lightweight Laplacian-based correction rather than repeatedly enforcing the full PDE, reducing computational cost while retaining a simple physical constraint. This makes the method more suitable for practical and real-time deployment. Inference time is reported in Section 4.4.

Note that both processes in Algorithms 1 and 2 can be extended to other physical PDEs by modifying the governing equation. Here, we use a heat-equation formulation tailored to LST dynamics. This makes PhyDiM applicable to other spatio-temporal physical systems with an appropriate PDE prior.

Algorithm 2 Physics-Guided Diffusion Sampling. The highlighted components denote the physics-based constraint introduced in PhyDiM compared to standard DDPM sampling.

Input: trained model ϵ_θ, conditional inputs $\mathcal{C} = (\mathbf{x}(t_i), t_i)$, target time t^*, noise schedule $\{\alpha_t\}$, guidance scale λ
Output: reconstructed field $\mathbf{x}_0$
Initialize $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T \to 1$ **do**
  Predict noise: $\boldsymbol{\epsilon}_\theta \leftarrow \epsilon_\theta(\mathbf{x}_t, \mathcal{C}, t^*, t)$
  Estimate clean field: $\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\, \boldsymbol{\epsilon}_\theta\right)$
  Physical constraint: $\mathcal{L}_{\text{phys}}(\hat{\mathbf{x}}_0) = \left\|\nabla^2 \hat{\mathbf{x}}_0\right\|_2^2$
  Standard DDPM update: $\mathbf{x}_{t-1}^{\text{ddpm}} = \text{DDPM}(\mathbf{x}_t, \boldsymbol{\epsilon}_\theta, \alpha_t)$
  Apply physics guidance: $\mathbf{x}_{t-1} = \mathbf{x}_{t-1}^{\text{ddpm}} - \lambda \nabla_{\mathbf{x}_t} \mathcal{L}_{\text{phys}}(\hat{\mathbf{x}}_0)$
end for
return $\mathbf{x}_0$

4 Experiments

Datasets. We construct the dataset from LST observations retrieved through Google Earth Engine (Gorelick et al., 2017). Terra MODIS data are obtained from the MOD11A1 (Collection 6.1) daily 1 km product, using *LST_Day_1km* and *LST_Night_1km* for t_1 and t_3, respectively, while Aqua observations from MYD11A1 (Collection 6.1) provide complementary measurements at t_2 and t_4. We consider three large regions of interest (ROIs), shown in Fig. 3: ROI_1 covers eastern France ($46°05'52''$–$49°41'17''$N, $0°14'56''$–$5°24'36''$E), ROI_2 covers central Spain ($38°37'12''$–$42°12'36''$N, $5°39'36''$–$1°07'12''$W), and ROI_3 spans parts of Germany, Belgium, and the Netherlands ($49°08'24''$–$52°43'48''$N, $4°22'48''$–$9°32'24''$E). The dataset spans a 10-year period from 2015 to 2025, covering substantial seasonal and inter-annual variability. Only observations with less than 5% cloud cover are retained. If any of the four daily observations exceeds this threshold, the entire sample is discarded. Missing pixels are filled using adaptive spatial interpolation during training, while cloud-affected pixels are masked and excluded from test evaluation. Images are divided into 80×80 patches (80 km$\times$80 km) with a stride of 20 for training and validation, while non-overlapping patches are used for testing to prevent spatial leakage. Each ROI is temporally split into 80% training, 10% validation, and 10% testing, with no observations from the same time period shared across subsets. The training sets contain 16,184, 12,138, and 12,427 samples for ROI_1, ROI_2, and ROI_3, respectively, while the test sets contain 175, 150, and 150 samples. Because MODIS acquisition times vary slightly, timestamps are rounded to the nearest hour (e.g., 10:20→10:00 and 10:40→11:00). Although ROI_1 and ROI_3 partially overlap spatially, their temporal splitting ensures that overlapping regions are never observed at the same time across different subsets.

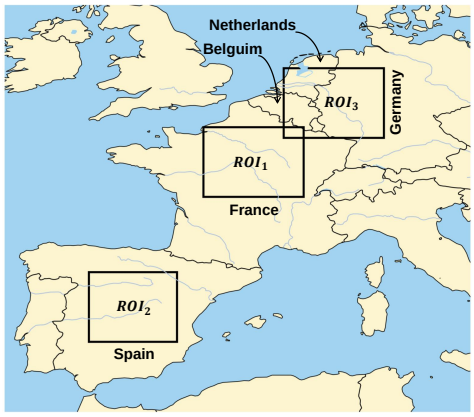


Figure 3: Geographical locations of the three ROIs used for dataset construction.

Implementation Details. Experiments are conducted in PyTorch on an NVIDIA RTX A6000. We use $T = 1000$ diffusion steps with a linear noise schedule ($\beta_{\min} = 8 \times 10^{-4}$, $\beta_{\max} = 2 \times 10^{-2}$). The physics-informed forward process uses $\kappa = 0.1$ and $\Delta t = 0.002$. We train with Adam (learning rate 2×10^{-4}, batch size 64) for 200 epochs, using an exponential moving average with decay 0.999 for evaluation. During inference, physics guidance is applied every 50 diffusion steps with $\lambda = 100$ to compensate for the small magnitude of the physics-based energy gradients. All hyperparameters are selected via grid search (see Section 4.4). Code is provided in the supplementary materials.

Baselines. We evaluate PhyDiM against eight baselines spanning classical interpolation, deep learning, neural operators, and diffusion models. The classical methods include (i) temporal inverse-

distance weighted interpolation (Shepard, 1968), (ii) pixel-wise quadratic temporal regression (Zhu & Woodcock, 2014), and (iii) a first-order harmonic Fourier model (Khormizi et al., 2025). We further consider (iv) a Transformer-based VFI model (VFIT) (Lu et al., 2022), (v) Fourier Neural Operator (FNO) (Li et al., 2020), and (vi) Physics-Informed Neural Operator (PINO) (Li et al., 2024). Since FNO and PINO are originally designed for PDE-based prediction, we adapt them to our sparse conditional spatio-temporal LST reconstruction task. Finally, we compare against (vii) a standard DDPM (Ho et al., 2020) and (viii) a physics-informed DDPM (Bastek et al., 2025), which incorporates physical knowledge through an additional physics-based loss. All methods are evaluated using RMSE, SSIM (Wang et al., 2004), PSNR, and SAM.

4.1 Quantitative Results

Table 1 presents quantitative comparisons across the three ROIs for target times t_1, t_2, t_3, and t_4. Each target includes multiple satellite overpass hours (e.g., $t_1 \in \{2h, 3h, 4h\}$), yielding varying temporal gaps and reconstruction difficulty. This setup evaluates generalization across irregular temporal intervals and diverse diurnal conditions.

In ROI_1, PhyDiM achieves strong performance across most metrics and target times, with notable improvements at the diurnal extremes (t_1 and t_4), reducing RMSE by 5.1% relative to DDPM and 55.41% relative to the best classical baseline. Although its RMSE at t_3 is 0.07 higher than DDPM, PhyDiM consistently improves SSIM, with gains of 7.32%, 14.55%, 11.67%, and 13.33% at t_1-t_4, respectively. Similar SSIM gains are observed in ROI_2 (8.51%, 5.80%, 7.25%, and 12.77%) and ROI_3 (15.15%, 12.50%, 11.67%, and 12.90%). These results indicate consistent improvements in structural reconstruction quality across different temporal configurations and ROIs. The physics-informed DDPM generally underperforms standard DDPM, suggesting that incorporating physical constraints through an additional training loss may complicate optimization. Compared with classical interpolation, Fourier-based models, and VFIT, PhyDiM provides more accurate LST reconstructions, highlighting the benefit of incorporating physical evolution into the diffusion trajectory. FNO and PINO achieve competitive performance through operator learning and PDE-based constraints, but require adaptation from forward PDE prediction to sparse conditional reconstruction. In contrast, PhyDiM directly integrates conditional generative modeling with physics-guided diffusion dynamics, providing greater flexibility for reconstruction across irregular temporal gaps.

Table 1: Quantitative comparison of PhyDiM with temporal inverse-distance weighted interpolation (Shepard, 1968), quadratic temporal regression (Zhu & Woodcock, 2014), harmonic Fourier modeling (Khormizi et al., 2025), VFIT (Lu et al., 2022), FNO (Li et al., 2020), PINO (Li et al., 2024), standard DDPM (Ho et al., 2020), and physics-informed DDPM (Bastek et al., 2025) across ROI_1-ROI_3. Results are reported for target times t_1-t_4. Best results are highlighted in bold.

Region	**ROI_1**															
Time	$t_1 \in \{2h, 3h, 4h\}$				$t_2 \in \{10h, 11h\}$				$t_3 \in \{14h, 15h\}$				$t_4 \in \{20h, 21h\}$			
Metric	RMSE↓	SSIM↑	PSNR↑	SAM↓	RMSE↓	SSIM↑	PSNR↑	SAM↓	RMSE↓	SSIM↑	PSNR↑	SAM↓	RMSE↓	SSIM↑	PSNR↑	SAM↓
Shepard (1968)	15.55	0.01	-3.09	15.46	2.93	0.62	14.76	2.80	4.44	0.64	10.94	2.64	14.27	0.10	-0.17	7.58
Zhu & Woodcock (2014)	13.15	-0.01	12.31	103.40	2.36	**0.63**	15.77	2.67	2.84	0.66	14.96	2.65	12.18	0.20	**12.60**	17.41
Li et al. (2020)	2.65	0.20	10.84	14.41	2.22	0.34	14.11	3.03	3.69	0.39	12.98	3.38	1.99	0.24	12.83	6.36
Li et al. (2024)	2.66	0.20	10.82	14.41	2.22	0.33	14.12	3.04	3.68	0.38	12.98	3.39	1.99	0.24	12.81	6.36
Khormizi et al. (2025)	3.79	0.24	**14.50**	26.92	2.52	0.56	16.41	3.37	2.44	0.62	16.76	3.01	3.75	0.33	15.16	12.45
Shi et al. (2022)	2.18	0.20	12.35	13.83	1.92	0.27	15.21	3.37	2.21	0.33	16.40	3.77	1.76	0.21	13.72	6.79
Ho et al. (2020)	1.97	0.38	14.02	12.29	1.54	0.55	18.24	2.70	1.85	0.60	18.45	2.74	**2.18**	0.45	14.05	7.74
Bastek et al. (2025)	3.31	0.28	12.33	27.41	2.44	0.49	16.25	3.82	2.65	0.58	16.78	3.56	2.82	0.37	13.72	10.15
PhyDiM	**1.87**	**0.41**	13.98	**11.78**	**1.47**	**0.63**	**18.29**	**2.40**	**1.71**	**0.67**	**19.11**	**2.51**	2.25	**0.51**	12.95	**7.28**
Region	**ROI_2**															
Time	$t_1 \in \{2h, 3h\}$				$t_2 \in \{10h, 11h\}$				$t_3 \in \{13h, 14h\}$				$t_4 \in \{21h, 22h\}$			
Shepard (1968)	23.03	-0.01	-2.75	8.06	5.30	**0.74**	12.16	1.95	6.63	**0.74**	10.31	1.92	23.00	0.10	-0.99	6.04
Zhu & Woodcock (2014)	26.96	-0.13	8.72	132.33	5.21	0.73	12.51	2.00	5.18	**0.74**	12.64	1.93	26.28	0.18	8.98	9.65
Li et al. (2020)	2.32	0.20	14.03	6.34	3.22	0.36	14.73	3.33	3.67	0.37	15.46	3.36	2.65	0.22	12.33	4.62
Li et al. (2024)	2.31	0.20	14.04	6.33	3.24	0.34	14.66	3.37	3.68	0.36	15.41	3.40	2.65	0.22	12.35	4.62
Khormizi et al. (2025)	4.80	0.29	13.85	13.21	3.32	0.68	16.92	2.43	3.28	0.71	17.25	2.21	4.75	0.36	14.58	5.91
Shi et al. (2022)	2.38	0.20	13.53	6.49	3.46	0.20	13.99	4.19	3.98	0.23	14.66	4.27	2.45	0.19	12.83	4.94
Ho et al. (2020)	2.14	0.47	15.27	5.54	2.28	0.69	18.60	2.01	2.97	0.69	17.89	2.01	**2.40**	0.47	**14.03**	3.98
Bastek et al. (2025)	2.73	0.38	14.33	7.26	3.14	0.58	16.67	2.69	3.41	0.62	17.33	2.69	2.76	0.37	13.98	5.28
PhyDiM	**2.09**	**0.51**	**15.17**	**5.30**	**2.21**	0.73	**18.68**	**1.91**	**2.89**	**0.74**	**18.03**	**1.89**	2.47	**0.53**	13.31	**3.71**
Region	**ROI_3**															
Time	$t_1 \in \{2h, 3h, 4h\}$				$t_2 \in \{10h\}$				$t_3 \in \{13h, 14h\}$				$t_4 \in \{21h\}$			
Shepard (1968)	14.52	0.04	-1.76	88.39	3.36	**0.67**	12.23	3.88	4.63	0.66	9.04	3.87	14.70	0.05	-1.19	38.47
Zhu & Woodcock (2014)	14.12	0.02	10.73	71.42	3.03	0.66	13.49	3.96	3.14	**0.67**	13.65	3.76	17.72	0.11	10.97	44.46
Li et al. (2020)	2.77	0.11	11.31	43.99	2.86	0.30	12.47	5.57	3.46	0.29	12.56	4.94	2.33	0.16	12.70	35.95
Li et al. (2024)	2.77	0.11	11.32	43.95	2.86	0.29	12.48	5.60	3.45	0.28	12.58	4.95	2.34	0.15	12.68	36.01
Khormizi et al. (2025)	4.06	0.17	13.14	46.41	2.51	0.62	14.93	4.55	2.63	0.64	15.35	4.04	4.70	0.18	13.60	45.92
Shi et al. (2022)	1.97	0.15	14.19	32.58	2.17	0.23	14.39	6.22	2.58	0.24	14.04	5.35	2.10	0.11	13.24	45.63
Ho et al. (2020)	1.98	0.33	14.73	33.40	1.81	0.56	16.71	4.51	2.15	0.60	16.73	4.01	2.18	0.31	14.34	37.25
Bastek et al. (2025)	2.92	0.23	13.25	47.40	2.39	0.51	14.58	5.54	2.40	0.54	17.05	5.23	2.35	0.29	**14.81**	**33.91**
PhyDiM	**1.91**	**0.38**	**14.87**	**30.91**	**1.74**	0.63	**16.86**	**4.05**	**2.08**	**0.67**	**16.70**	**3.57**	**2.13**	**0.35**	14.02	36.33

4.2 Qualitative Results

Fig. 4 presents qualitative comparisons between PhyDiM, DDPM (Ho et al., 2020), and PhyDDPM (Bastek et al., 2025). The results span all three ROIs and diverse acquisition times, including both daytime and nighttime conditions. Additional qualitative results are provided in Appendix B.

In the first example ($t_1 = 3h$, ROI_2), corresponding to nighttime reconstruction, PhyDiM better preserves elongated thermal structures and smooth spatial transitions, while DDPM introduces local noise and PhyDDPM produces overly smooth reconstructions with reduced detail. For the daytime examples ($t_2 = 10h$, ROI_1 and ROI_3), PhyDiM more accurately recovers localized warm regions, thermal gradients, and sharp boundaries despite partially missing observations, whereas DDPM produces noisier patterns and PhyDDPM suppresses local contrasts. At $t_3 = 14h$ in ROI_2, PhyDiM closely matches the morphology and spatial extent of the dominant thermal structures, demonstrating robust reconstruction under dynamic daytime conditions. Finally, the $t_4 = 21h$ example in ROI_1, associated with nighttime reconstruction near the end of the diurnal cycle, shows that PhyDiM preserves the shape, intensity, and location of the dominant warm region while maintaining coherent surrounding temperatures. In contrast, DDPM introduces substantial noise in homogeneous areas, whereas PhyDDM excessively smooths the thermal structure and reduces local contrast.

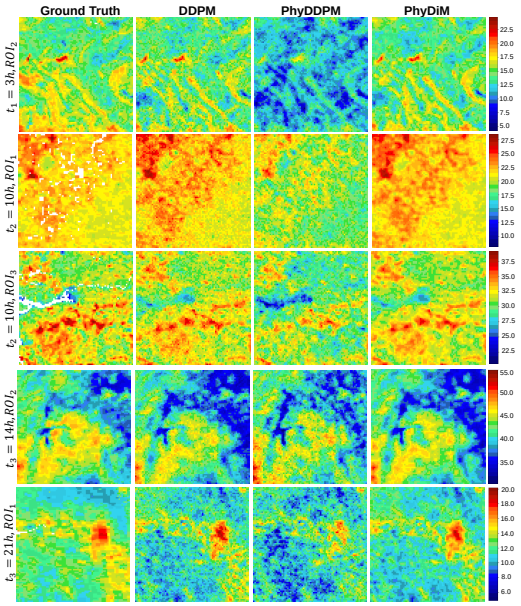


Figure 4: Qualitative comparison between PhyDiM, standard DDPM Ho et al. (2020), and PhyDDPM Bastek et al. (2025).

4.3 Temporal Generalization Analysis

To evaluate temporal consistency, we reconstruct LST at hourly intervals over a complete diurnal cycle. Since ground-truth observations are available only at four acquisition times (t_1, t_2, t_3, t_4), each target hour is conditioned on three observations, excluding the temporally closest one. For example, t_1 is excluded when reconstructing around 2:00. As shown in Fig. 5, PhyDiM produces smooth, physically coherent diurnal trajectories, capturing nighttime cooling, daytime warming, and subsequent afternoon cooling. DDPM captures the overall trend but underestimates temperature variations, whereas PhyDDPM produces overly smooth trajectories with reduced diurnal amplitude. The stable reconstructions despite excluding the nearest observation demonstrate robust continuous-time prediction under sparse temporal conditioning. Additional temporal generalization validation against independent near-ground-based sensor observations is provided in Appendix C.

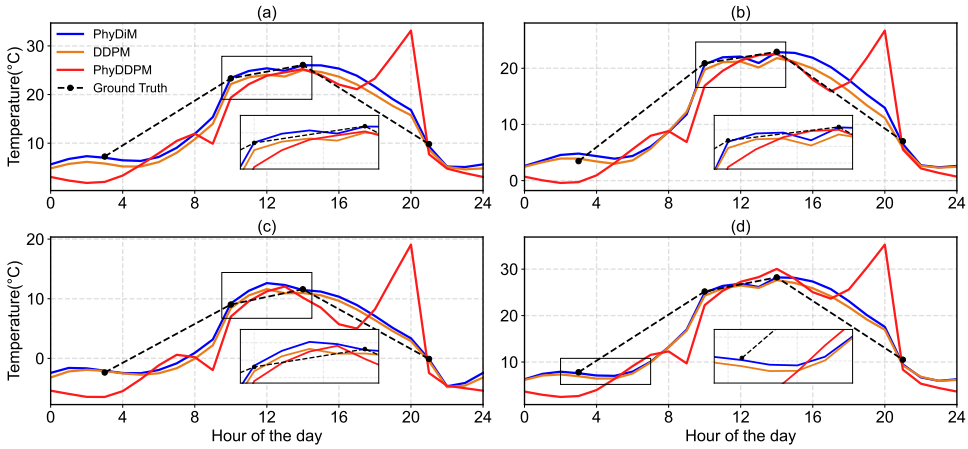


Figure 5: Diurnal LST reconstruction for four representative samples (a–d) using PhyDiM, DDPM (Ho et al., 2020), and PhyDDPM (Bastek et al., 2025).

4.4 Ablation Studies

Physics Guidance Components. Table 2 evaluates the contribution of the physics-guided components on ROI_1 across three configurations: (i) physics only in the forward process, (ii) physics guidance only during reverse sampling, and (iii) the complete framework combining both. Similar trends hold for ROI_{2-3} and are omitted for brevity. Results are averaged over all timesteps. Physics in the forward process improves reconstruction consistency by producing physically coherent noisy states, but without sampling-time guidance, the reverse trajectory remains data-driven and may yield less physically consistent solutions. Sampling-time guidance provides greater improvements by enforcing spatial coherence and suppressing artifacts, although a training-inference discrepancy remains. The complete framework achieves the best performance across all metrics, highlighting the benefit of jointly incorporating physics during both corruption and reconstruction.

Table 2: Ablation of physics guidance in the forward and sampling processes on ROI_1.

Configurations	Forward Diffusion	Sampling Phase	RMSE↓	SSIM↑	PSNR↑	SAM↓
Config1	✓	✗	1.89	0.49	16.15	6.47
Config2	✗	✓	1.82	0.56	16.11	5.87
PhyDiM	✓	✓	**1.81**	**0.60**	**16.78**	**4.06**

Sampling Guidance Frequency. Table 3 evaluates the effect of physics-guidance frequency during inference by applying the Laplacian correction every 50, 20, 10, or 1 diffusion steps, with DDPM as a reference. More frequent guidance increases computational cost, while the 50-step interval provides a favorable accuracy-efficiency trade-off with only a $+0.98$ s runtime overhead.

Table 3: Inference time for different physics-guidance frequencies.

Method	**DDPM**	**PhyDiM(50)**	**PhyDiM(20)**	**PhyDiM(10)**	**PhyDiM(1)**
Time(s)	2.74±0.09	3.72±0.49	4.56±1.07	5.14±1.10	17.42±1.37

Hyperparameter Sensitivity. κ controls the rate of spatial heat diffusion, while Δt determines the physical evolution time in the explicit Euler discretization. Since these parameters are difficult to estimate reliably for MODIS LST pixels, we use $\kappa\Delta t$ as an effective diffusion strength. Small values provide insufficient physical evolution, whereas large values cause excessive smoothing. We therefore select $\kappa\Delta t$ via validation-based grid search across all datasets. We set the guidance scale to $\lambda = 100$ due to the small gradient magnitudes induced by input normalization. As shown in Table 4, small λ values provide limited guidance, whereas large values introduce excessive smoothing.

Table 4: λ sensitivity on ROI_1.

λ	**1**	**10**	**100**	**200**	**400**
RMSE	1.89	1.88	1.82	1.83	2.03
SSIM	0.49	0.50	0.56	0.54	0.37

5 Conclusion

In this work, we proposed PhyDiM, a physics-guided diffusion model for continuous-time LST reconstruction from sparse satellite observations. Unlike existing physics-informed diffusion approaches that impose physical constraints through training losses, PhyDiM incorporates physical priors directly into the diffusion process through physics-informed forward corruption and guided sampling. Experiments across climatically diverse regions spanning France, Spain, Germany, Belgium, and the Netherlands, demonstrate consistent improvements over classical and diffusion-based baselines, reducing RMSE by an agerage 41.7% and improving SSIM by 24.4% over the best classical baseline, and reducing RMSE by an average 2.1% and improving SSIM by 12% over DDPM. Qualitative and temporal evaluations further show improved physical consistency and preservation of localized thermal structures. PhyDiM can also incorporate different physical PDEs, which enables adaptation to a broad range of scientific applications beyond LST reconstruction.

Despite its strong performance, PhyDiM has some limitations. First, its iterative reverse diffusion process remains computationally demanding at inference. Future work will explore accelerated samplers, such as Denoising Diffusion Implicit Models, and assess their robustness under the proposed physics-guided formulation. Second, PhyDiM is currently evaluated only for LST reconstruction from MODIS observations. Moreover, future work will extend the framework to other physics-governed applications, such as air pollution dispersion, ocean temperature, soil moisture, and atmospheric variable reconstruction, where PDE-based priors can guide the generative process.

6 LLM USAGE

We used an LLM selectively for grammar correction and language polishing of parts of the manuscript. All generated suggestions were reviewed and revised by the authors. LLMs were not used to develop the research methodology, conduct experiments, analyze data, or formulate scientific claims, results, or conclusions.

REFERENCES

Saeid Asadzadeh and Carlos Roberto Souza Filho. Numerical modeling of land surface temperature over complex geologic terrains: A remote sensing approach. *Remote Sensing*, 15(19):4877, 2023.

Dayoung Baik and Jaejun Yoo. Dynamic-aware spatio-temporal representation learning for dynamic mri reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 174–184. Springer, 2025.

Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3703–3712, 2019.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis M Kochmann. Physics-informed diffusion models. *arXiv preprint arXiv:2403.14404*, 2024.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 3360–3385, 2025.

Theodore L Bergman. *Fundamentals of heat and mass transfer*. John Wiley & Sons, 2011.

Sofiane Bouaziz, Adel Hafiane, Raphaël Canals, and Rachid Nedjai. Deep learning for spatio-temporal fusion in land surface temperature estimation: A comprehensive survey, experimental analysis, and future trends. *Remote Sensing*, 18(2):289, 2026a.

Sofiane Bouaziz, Adel Hafiane, Raphael Canals, and Rachid Nedjai. Spycer: Semi-supervised physics-guided contextual attention for near-surface air temperature estimation from satellite imagery. *arXiv preprint arXiv:2603.05219*, 2026b.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

TJ Chung. *Computational fluid dynamics*. Cambridge university press, 2010.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 10850–10869, 2023.

Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1472–1480, 2024.

Si-Bo Duan, Zhao-Liang Li, Hua Li, Frank-M Göttsche, Hua Wu, Wei Zhao, Pei Leng, Xia Zhang, and César Coll. Validation of collection 6 modis land surface temperature product using in situ measurements. *Remote sensing of environment*, 225:16–29, 2019.

Lawrence C Evans. *Partial differential equations*, volume 19. American mathematical society, 2022.

Joel H Ferziger, Milovan Perić, and Robert L Street. *Computational methods for fluid dynamics*, volume 3. Springer, 2002.

Paul Gibson. Chapter 1 - a systematic view of remote sensing (second edition). In Shunlin Liang and Jindi Wang (eds.), *Advanced Remote Sensing*, pp. 1–57. Academic Press, 2020. ISBN 978-0-12-815826-5. doi: https://doi.org/10.1016/B978-0-12-815826-5.00001-5. URL https://www.sciencedirect.com/science/article/pii/B9780128158265000015.

Noel Gorelick et al. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.*, 202:18–27, 2017.

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

Xinrong Hu, Yu-Jen Chen, Tsung-Yi Ho, and Yiyu Shi. Conditional diffusion models for weakly supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 756–765. Springer, 2023.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7341–7351, 2024.

Hadi Zare Khormizi, Mohammad Jafari, Hamidreza Ghafarian Malamiri, Ali Tavili, and Hamidreza Keshtkar. Generating modis hourly land surface temperature under clear sky conditions using fourier series analysis. *International Journal of Applied Earth Observation and Geoinformation*, 136:104341, 2025.

Dahyeon Kye, Changhyun Roh, Sukhun Ko, Chanho Eom, and Jihyong Oh. Acevfi: A comprehensive survey of advances in video frame interpolation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Thomas Martin Lehmann, Claudia Gonner, and Klaus Spitzer. Addendum: B-spline interpolation in medical image processing. *IEEE transactions on medical imaging*, 20(7):660–665, 2001.

Randall J LeVeque. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge university press, 2002.

Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

Lixin Li and Peter Revesz. Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems*, 28(3):201–227, 2004.

Tianyi Li, Alessandra S Lanotte, Michele Buzzicotti, Fabio Bonaccorso, and Luca Biferale. Multi-scale reconstruction of turbulent rotating flows with generative diffusion models. *Atmosphere*, 15 (1):60, 2023.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024.

Zongyi Li et al. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Marten Lienen, David Lüdke, Jan Hansen-Palmus, and Stephan Günnemann. From zero to turbulence: Generative modeling for 3d flow simulation. *arXiv preprint arXiv:2306.01776*, 2023.

Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3532–3542, 2022.

Mengting Luo, Nan Zhou, Tao Wang, Linchao He, Wang Wang, Hu Chen, Peixi Liao, and Yi Zhang. Bi-constraints diffusion: A conditional diffusion model with degradation guidance for metal artifact reduction. *IEEE Transactions on Medical Imaging*, 2024.

Ivan Marisca, Andrea Cini, and Cesare Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. *Advances in neural information processing systems*, 35: 32069–32082, 2022.

Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, and Alexander Sorkine-Hornung. Phase-based frame interpolation for video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1410–1418, 2015.

Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus Gross, and Christopher Schroers. Phasenet for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 498–507, 2018.

Chen Qin, Jo Schlemper, Jose Caballero, Anthony N Price, Joseph V Hajnal, and Daniel Rueckert. Convolutional recurrent neural networks for dynamic mr image reconstruction. *IEEE transactions on medical imaging*, 38(1):280–290, 2018.

Kara Marie Schatz, Erik Quintanilla, Shruti Vyas, and Yogesh S Rawat. A recurrent transformer network for novel view action synthesis. In *European Conference on Computer Vision*, pp. 410–426. Springer, 2020.

Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2017.

Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Video frame interpolation and enhancement via pyramid recurrent framework. *IEEE Transactions on Image Processing*, 30:277–292, 2020.

Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pp. 517–524, 1968.

Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17482–17491, 2022.

Qidi Shu, Xiaolin Zhu, Shuai Xu, Yan Wang, and Denghong Liu. Restore-dit: Reliable satellite image time series reconstruction by multimodal sequential diffusion transformer. *Remote Sensing of Environment*, 328:114872, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Juhi Soni, Markus Lange-Hegermann, and Stefan Windmann. Physics-informed diffusion models for unsupervised anomaly detection in multivariate time series. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 202–217. Springer, 2025.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.

Geoffrey K Vallis. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press, 2017.

Thomas J Vandal and Ramakrishna R Nemani. Temporal interpolation of geostationary satellite imagery with optical flow. *IEEE Transactions on Neural Networks and Learning Systems*, 34(7): 3245–3254, 2021.

Nikolaos N Vlassis and WaiChing Sun. Denoising diffusion algorithm for inverse design of microstructures with fine-tuned nonlinear material properties. *Computer Methods in Applied Mechanics and Engineering*, 413:116126, 2023.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

David Vernon Widder. *The heat equation*, volume 67. Academic Press, 1976.

Haoning Wu, Xiaoyun Zhang, Weidi Xie, Ya Zhang, and Yanfeng Wang. Boost video frame interpolation via motion adaptation. *arXiv preprint arXiv:2306.13933*, 2023.

Weifeng Xu, Yaofei Cheng, Mengyuan Luo, Xuzhi Mai, Wenhuan Wang, Wei Zhang, and Yinghui Wang. Progress and limitations in forest carbon stock estimation using remote sensing technologies: a comprehensive review. *Forests*, 16(3):449, 2025.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16010–16021, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Zihao Zhang, Haoran Chen, Haoyu Zhao, Guansong Lu, Yanwei Fu, Hang Xu, and Zuxuan Wu. Eden: Enhanced diffusion for high-quality large-motion video frame interpolation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2105–2115, 2025.

Lili Zhao, Xuncheng Fan, and Tao Hong. Urban heat island effect: remote sensing monitoring and assessment—methods, applications, and future directions. *Atmosphere*, 16(7):791, 2025.

Jie Zhou, Li Jia, Massimo Menenti, and Ben Gorte. On the performance of remote sensing time series reconstruction methods–a spatial comparison. *Remote Sensing of Environment*, 187:367–384, 2016.

Zhe Zhu and Curtis E Woodcock. Continuous change detection and classification of land cover using all available landsat data. *Remote sensing of Environment*, 144:152–171, 2014.

Appendix

A Validity of the Noise-Prediction Objective under Physics-Guided Corruption

We show that the standard noise-prediction objective remains a valid denoising score matching objective when the diffusion corruption is preceded by a deterministic physics-guided transformation.

Background: Standard DDPM.
In a standard DDPM, a clean sample $\mathbf{x}_0$ is corrupted as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{8}$$

and the network is trained to predict $\boldsymbol{\epsilon}$ using:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2\right]. \tag{9}$$

The question is whether this same noise-prediction objective remains valid when $\mathbf{x}_0$ is first transformed by a deterministic physics-guided evolution $F_t(\mathbf{x}_0)$ before the Gaussian perturbation.

We first consider a general corruption process of the form:

$$\mathbf{x}_t = m_t(\mathbf{x}_0) + \sigma_t\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

where $m_t(\cdot)$ is a deterministic differentiable mapping and $\sigma_t > 0$. Conditioned on $\mathbf{x}_0$, the corrupted sample follows (using the standard properties of Gaussian random variables):

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(m_t(\mathbf{x}_0), \sigma_t^2\mathbf{I}\right). \tag{11}$$

Therefore, the conditional score with respect to the corrupted sample is expressed as:

$$\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{\mathbf{x}_t - m_t(\mathbf{x}_0)}{\sigma_t^2}. \tag{12}$$

From Eq. 10, we have $\mathbf{x}_t - m_t(\mathbf{x}_0) = \sigma_t\boldsymbol{\epsilon}$, and hence:

$$\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{\boldsymbol{\epsilon}}{\sigma_t}. \tag{13}$$

Thus, although $m_t(\cdot)$ determines the mean and therefore changes the marginal distribution of $\mathbf{x}_t$, the conditional score has the same noise parameterization as in standard Gaussian diffusion.

For PhyDiM, the deterministic mapping is given by the deterministic physics-guided evolution:

$$m_t(\mathbf{x}_0) = \sqrt{\bar{\alpha}_t}F_t(\mathbf{x}_0), \tag{14}$$

where F_t denotes the heat PDE evolution applied before stochastic perturbation. The forward process is therefore:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}F_t(\mathbf{x}_0) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}. \tag{15}$$

Consequently, the conditional distribution becomes:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}F_t(\mathbf{x}_0), (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{16}$$

and, the conditional score becomes:

$$\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{\boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}}. \tag{17}$$

Result. The conditional score retains the same noise-based parameterization as in standard DDPMs. Therefore, $\boldsymbol{\epsilon}$ remains a valid denoising score matching target, and the standard noise-prediction objective can be used under the physics-guided corruption.

Accordingly, the same noise-prediction objective can be used:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2 \right] \tag{18}$$

where $\mathbf{c}$ denotes the conditioning information. The physics-guided evolution modifies the geometry of the corrupted samples by introducing physically meaningful structures before stochastic perturbation, while preserving the Gaussian perturbation required by the diffusion objective. Larger physical evolution parameters ($\kappa\Delta t$) may affect the complexity of the denoising task by changing the distribution of corrupted samples, but they do not invalidate the score-matching formulation.

B Additional Qualitative Results

In addition to the globally normalized qualitative comparison presented in the main paper, Fig. 6 shows reconstructed LST maps using a per-sample thermal colormap, where each image is independently normalized according to its own temperature range. This per-sample normalization improves the visualization of local thermal variability and spatial organization, making hotspot structures, cold regions, and smooth temperature transitions more distinguishable.

Across all examples, PhyDiM preserves fine-grained thermal structures, including localized hot spots and cold regions, while maintaining globally coherent spatial organization. In particular, in the first example ($t_1 = 3h$, ROI_2), the proposed method accurately reconstructs sharp thermal gradients and elongated warm structures that are partially degraded in DDPM and excessively smoothed in PhyDDPM. Similarly, in the second and third examples ($t_2 = 10h$, ROI_1 and ROI_3), PhyDiM successfully preserves compact high-temperature regions and surrounding cold transition areas, while maintaining sharper and more consistent boundary delineations between hot and cold zones compared to both baselines. Importantly, at later time steps such as ROI_2 at $t_3 = 14h$ and ROI_1 at $t_4 = 21h$, the proposed method continues to produce stable reconstructions, preserving hotspot localization and cold-region structure even under more challenging nighttime conditions and at the extremities of the diurnal cycle.

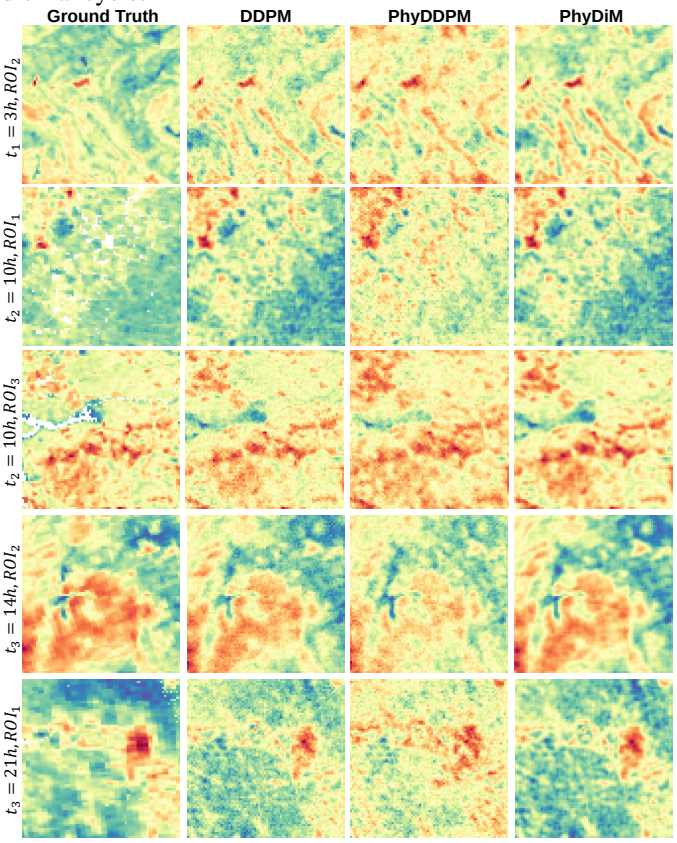


Figure 6: Per-sample normalized qualitative comparison between PhyDiM, DDPM (Ho et al., 2020), and PhyDDPM (Bastek et al., 2025).

C IN-SITU VALIDATION

The temporal generalization evaluation presented in the main article, is based on Terra and Aqua MODIS observations available at the four acquisition times (t_1, t_2, t_3, t_4). Although these results demonstrate that PhyDiM can accurately reconstruct LST at satellite-observed timestamps, they do not by themselves establish the reliability of the reconstructed temperature trajectories at unobserved hourly timestamps. To provide an independent assessment of the hourly reconstructions, we therefore perform an in-situ validation using ground-based meteorological sensors that provide hourly near-surface air temperature (NSAT) measurements.

Ground-based meteorological sensors measure NSAT, typically at a height of approximately 2 m above the ground, whereas PhyDiM reconstructs LST. Since LST and NSAT are physically different quantities, they cannot be compared directly. We therefore use SPyCer (Bouaziz et al., 2026b), a model designed to infer NSAT from LST together with complementary temporal and environmental information. SPyCer reported an average RMSE of $2.27°\mathrm{C}$ in its original evaluation. In our setting, SPyCer provides an intermediate mapping from the hourly LST reconstructed by PhyDiM to estimated hourly NSAT, which can then be directly compared with the independent sensor measurements. The in-situ validation is conducted within Orléans Métropole, France, over a region extending from 47°50'41.77"N to 47°54'1.74"N in latitude and from 1°50'6.98"E to 1°59'36.36"E in longitude, covering approximately 114 km^2 and containing 33 near-ground sensors.

The original SPyCer model was trained using observations acquired primarily around 10:00-11:00. To apply it to the full hourly LST trajectories generated by PhyDiM, we extend its temporal representation by introducing two additional inputs corresponding to the sine and cosine embeddings of the hour of day (The remainder of the SPyCer procedure is kept unchanged). We then retrain SPyCer on an extended dataset covering April 1, 2025, to June 16, 2026, comprising 27,218 hourly samples collected from the 27 training near-ground-based sensors. The remaining six near-ground-based sensors are held out exclusively for testing, allowing us to evaluate generalization to previously unseen sensor locations. For each of the six held-out sensors, we generate hourly LST using PhyDiM and use the retrained SPyCer model to infer the corresponding hourly NSAT. The inferred NSAT is then compared with the actual sensor measurements over the complete hourly time series. The six test sensors are never used during SPyCer training to ensure that their locations remain completely unseen. This strategy provides an independent evaluation of whether the hourly temporal dynamics reconstructed by PhyDiM are consistent with near*ground-based observations at unseen locations.

Table 5 reports the performance of the proposed in-situ validation across the six held-out sensors. PhyDiM achieves an average RMSE of $2.251°\mathrm{C}$ and an average MAE of $1.759°\mathrm{C}$ across the six unseen locations. The performance varies across sensors, with RMSE ranging from $1.503°\mathrm{C}$ to $2.758°\mathrm{C}$. Despite this variability, the model maintains reasonable agreement with the independent hourly sensor measurements across all six locations.

Table 5: In-situ validation results on six held-out near-ground sensors.

Sensor	Number of Hours	RMSE↓	MAE↓
1	1390	1.503	1.168
2	1342	1.966	1.425
3	2440	2.166	1.714
4	1384	2.433	1.863
5	1614	2.682	2.209
6	1818	2.758	2.179
Average	/	**2.251**	**1.759**

Fig. 7 shows the predictions for test sensor 3, which has the largest number of available observations, with 2,440 hourly samples. Panel (a) presents the predictions over the complete test period, spanning more than one year, while the remaining panels show representative daily profiles from different periods of the year: 30 April 2025, 15 August 2025, 12 October 2025, 20 March 2026, and 14 June 2026. Across these periods, the predicted NSAT closely follows the ground truth by capturing the characteristic diurnal cycle from morning warming to the daytime maximum and subsequent cooling through the afternoon and nighttime. This agreement is observed not only in the overall NSAT magnitude but also in the timing and shape of the daily variations. Moreover, the predictions maintain a continuous and physically coherent diurnal evolution throughout the 24-hour cycle.

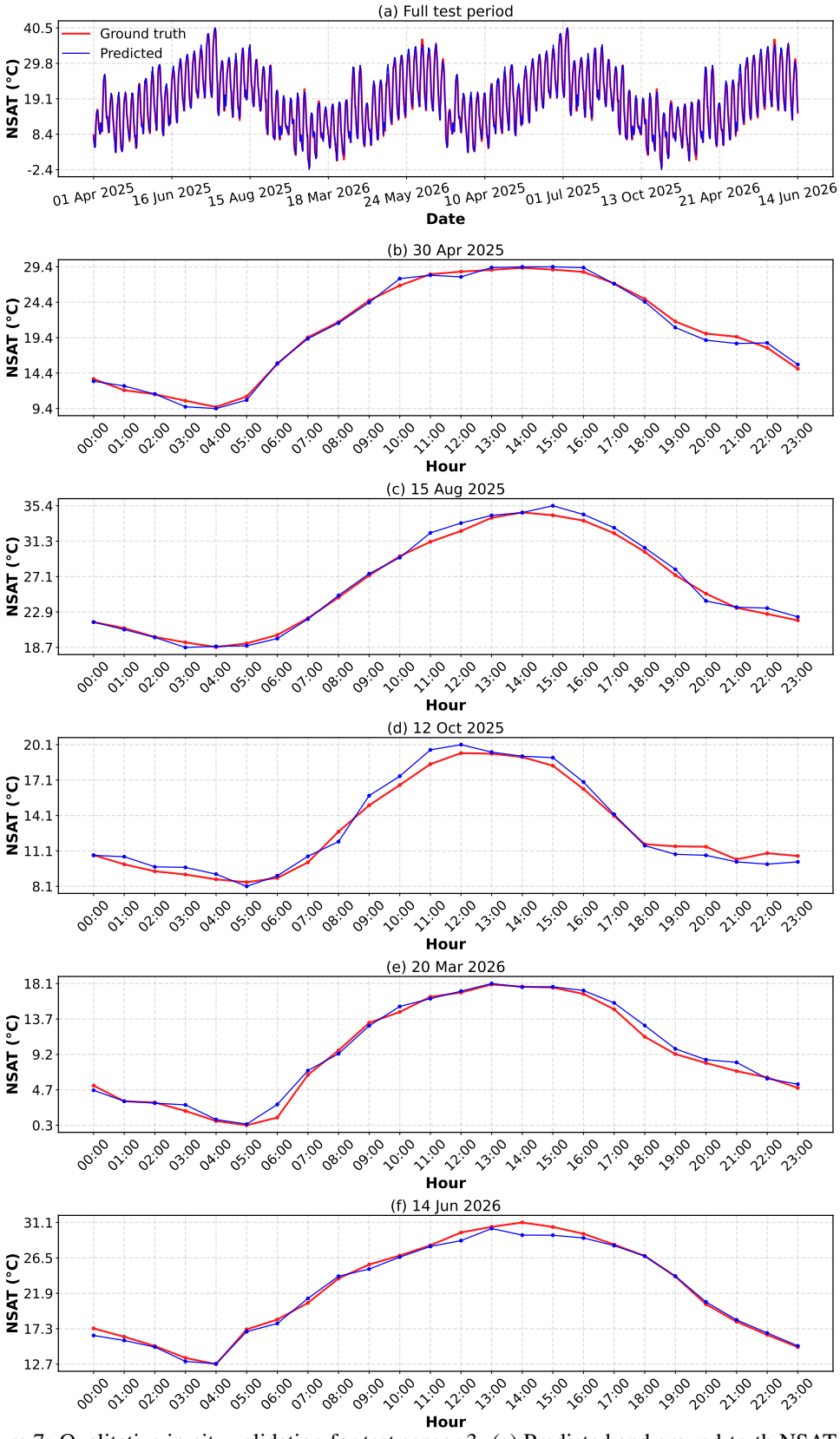


Figure 7: Qualitative in-situ validation for test sensor 3. (a) Predicted and ground-truth NSAT over the complete test period. (b)-(f) Representative daily profiles from different periods of the year.