ALETHEIA: A MULTI-FREQUENCY EDDY CURRENT PULSED THERMOGRAPHY DATASET FOR NEURAL OPERATOR LEARNING IN NONDESTRUCTIVE TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning neural solvers for spatiotemporal partial differential equations (PDEs) under real-world constraints remains a key challenge in scientific machine learning, especially for inverse tasks with sparse and noisy boundary observations. We present the **Aletheia** dataset, the first 3D benchmark for learning data-driven solvers in the context of **nondestructive testing (NDT)**. The dataset simulates eddy-current-induced heating in conductive solids and models the resulting transient heat propagation governed by the heat equation. Aletheia contains over 4,700 high-resolution samples across 10 excitation frequencies (1-100 kHz), each providing volumetric heat source and temperature fields over time. It supports both forward prediction of temperature evolution and inverse reconstruction of internal heat sources or defects from surface infrared measurements. Real infrared thermography data from cracked rail specimens are included for calibration and generalization studies. We define three canonical tasks on both regular and irregular grids and benchmark them using various neural operators. Aletheia establishes a unified platform for evaluating neural PDE solvers under realistic NDT conditions, enabling progress in reliable, data-driven inverse modeling.

1 INTRODUCTION

Neural operator methods, such as the Fourier Neural Operator (FNO) (Li et al., 2021, 2023b; Tran et al., 2023; Xiao et al., 2024) and Transformer-based solvers (Li et al., 2023a; Wu et al., 2024; Lee & Oh, 2024), have emerged as a transformative approach for learning solution operators of partial differential equations (PDEs) directly from data. Unlike traditional methods, these architectures bypass mesh-dependent discretizations, enabling robust generalization across parameterized PDE families. However, they are typically evaluated on academic datasets (e.g., Darcy flow, Navier–Stokes) with fully observed fields and simplified geometries, which fail to capture the complexities of real-world inverse problems. In applications like nondestructive testing (NDT) (Gupta et al., 2022; Xiong et al., 2023; Yuan et al., 2021; Gong et al., 2022; Tuschl et al., 2021), challenges such as sparse or noisy boundary observations, unknown source terms, and heterogeneous domains demand more robust benchmarks (Molinaro et al., 2023; Azizzadenesheli et al., 2024).

In NDT, reconstructing hidden defects (Lin et al., 2023; Zhao et al., 2022; Tao et al., 2022; Wu et al., 2021) from surface temperature measurements, as in inverse heat conduction problems (Silva et al., 2023), is inherently ill-posed: distinct subsurface defects or excitation conditions can produce nearly identical surface temperature patterns (Woodbury et al., 2023), as illustrated in Figure 1. To address this, we employ multi-frequency pulsed induction heating, where different excitation frequencies probe the material at varying depths—lower frequencies penetrate deeper to capture internal defect responses, while higher frequencies reveal surface-level thermal behavior (Liang et al., 2024). As shown in Figure 1, while some frequencies (e.g., 25 kHz) may yield similar surface temperatures for different defects, others (e.g., 9 kHz) reveal distinct patterns. This frequency-dependent response diversity breaks single-frequency ambiguity, enhancing defect discriminability.

Existing PDE-learning benchmarks lack realistic thermal-boundary coupling and 3D scenarios tailored to heat-source inversion or volumetric temperature prediction in NDT. To bridge this gap, we introduce the **Aletheia** dataset (Figure 2), a comprehensive 3D benchmark that integrates high-fidelity

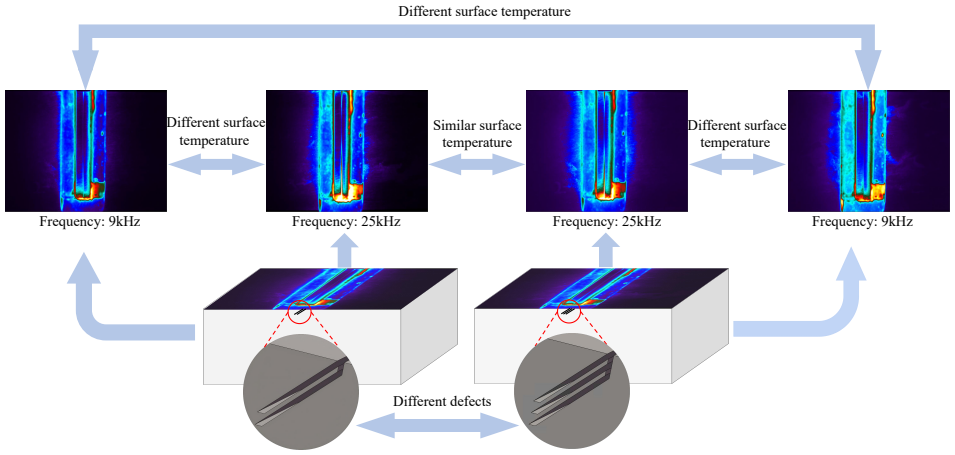


Figure 1: **Multi-frequency stimulation in the Aletheia dataset.** Different defects may produce similar surface temperatures under the same frequency (25 kHz), but show clear differences at another frequency (9 kHz). Using multiple frequencies helps distinguish hidden defects by capturing varied thermal responses at different depths.

simulations with real infrared measurements from rail specimens with internal fatigue cracks. The dataset includes over 4,700 defect cases across 10 excitation frequencies (1–100 kHz), providing volumetric heat-source maps (eddy-current-induced Joule heating) and time-resolved temperature fields on both regular and irregular grids. Calibrated thermography captures transient surface temperature sequences, while multi-frequency conditions supply depth-sensitive signals to mitigate the ill-posedness of surface-only observations.

Using Aletheia, we address three key tasks in eddy current thermography and PDE benchmarking: (1) **forward thermal prediction** of full 3D temperature evolution from known sources; (2) **inverse source reconstruction** of latent heat distributions or defect geometries from sparse surface data; and (3) **out-of-distribution (OOD) generalization** to unseen frequencies, defect shapes, and material variants. Overall, our contributions are summarized as follows:

- We present the first publicly available simulation dataset **Aletheia** in the context of electromagnetic-thermal coupling, enabling the datatization of eddy current thermography.
- Aletheia provides a multi-frequency dimension: data covering a range of excitation frequencies from low to high such as 1—100 kHz, capturing the effect of frequency on the depth and effectiveness of the heat.
- Aletheia contains three-dimensional, temporally-evolving data, such as the evolution of the entire temperature field after pulse heating, not just steady-state or two-dimensional observations
- Aletheia combines high-fidelity simulations and experimental measurements. Simulation data provide comprehensive information on field distributions and "true value" defects, while experimental data introduce real noise and variability and verify the reliability of the simulation.
- Built around the real engineering application of rail crack detection, Aletheia covers a wide range of defect types and sizes with direct engineering relevance.

2 Background

Neural operators have revolutionized data-driven PDE modeling by learning mapping between function spaces (Lu et al., 2021; Li et al., 2021, 2020), but existing benchmarks remain narrowly focused on fully observed, forward-only problems and lack realistic inverse or measurement-sparse scenarios. Eddy-current thermography(ECT) offers a rich real-world setting in NDT, yet no public

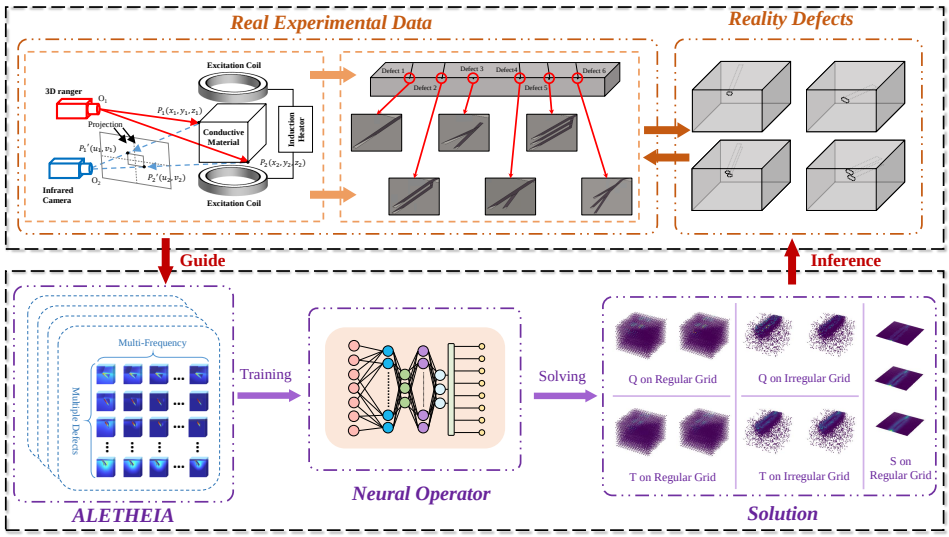


Figure 2: **Overview of the Aletheia dataset.** *Top*: real infrared thermography sequences from rail specimens with fatigue cracks. *Bottom*: synthetic data pipeline generating multi-frequency (1—100 kHz) volumetric heat sources and corresponding transient temperature fields for 4,700+ defect cases, on both regular and irregular grids. This combined sim-to-real benchmark supports forward prediction, inverse reconstruction, and cross-frequency evaluation in NDT.

3D dataset exists for learning heat-field inversion under varying excitation. We therefore position our dataset at the intersection of these gaps, enabling the study of both forward prediction and inverse source reconstruction for transient heat conduction from surface measurements.

2.1 Challenges and Applications of Neural PDE Solvers

Neural operator methods (e.g., DeepONet (Lu et al., 2021), FNO (Li et al., 2021)) learn PDE solution operators directly from data, achieving mesh- and resolution-invariance and outperforming classical surrogates on benchmarks like Darcy flow and NavierStokes (Takamoto et al., 2022). Despite their promise, these models struggle to capture high-frequency components due to spectral truncation in Fourier layers, exhibit poor extrapolation to OOD parameters, and lack robustness under noisy or incomplete boundary observations common in inverse problems. Moreover, most evaluations assume full-field availability, whereas many applications demand reconstructing latent sources or fields from sparse measurements.

2.2 Thermal Holography for NDT

ECT (Gao et al., 2024; Zou et al., 2022; Zu et al., 2023) combines electromagnetic induction heating with infrared imaging to detect subsurface defects by capturing surface temperature anomalies. A common implementation, pulsed ECT (ECPT) (Zhang et al., 2024; Chen et al., 2021), uses short bursts of alternating current to induce volumetric Joule heating in conductive materials, enabling high-sensitivity, non-contact inspection of internal structures (Yin et al., 2021; Ma et al., 2024; Zhang et al., 2021). This process gives rise to the concept of thermal holography (Utadiya et al., 2023), where internal heat sources perturbed by defects such as cracks are inferred from transient surface temperature fields. Mathematically, this is governed by the heat equation with a volumetric heat source:

$$\frac{\partial u}{\partial t} = \alpha \nabla^2 u + q, \tag{1}$$

where $u(\mathbf{x}, t)$ is the temperature field at position $\mathbf{x} = (x, y, z)$ and time t, α is the thermal diffusivity, and $q(\mathbf{x}, t)$ represents the internal heat sources induced by eddy currents. The inverse problem in thermal holography seeks to reconstruct the spatial distribution of the heat source $q(x, y, z)$ from sparse and noisy surface temperature measurements $u(x, y, z_s, t)$, where z_s denotes the surface of the domain $\partial\Omega$, and the observation set is defined over coordinates (x, y, z_s, t). In many practical

settings, the primary goal is to reconstruct the volumetric distribution of these internal heat (Hong et al., 2023; Wang et al., 2021) sources rather than directly visualize the defects themselves. This inverse heat conduction task is inherently ill-posed and only sparse, noisy boundary data are typically available. Despite its critical role in nondestructive testing across domains such as rail, aerospace (Gebrehiwet et al., 2023; Gholizadeh & Gholizadeh, 2022; Jacob & Raddatz, 2022), and pipeline inspection (Cheng et al., 2021; Wang et al., 2024), there exists no standardized benchmark dataset for learning data-driven solvers that can tackle this class of inverse thermal problems under realistic conditions.

2.3 Limitation of existing PDE benchmarks

As summarized in Table 1, most existing PDE benchmark datasets exhibit limited diversity in terms of task settings and data characteristics. Specifically, widely used datasets such as *Darcy Flow*, *NavierStokes*, *Burgers' Equation*, and *Shallow Water* primarily focus on forward and inverse problems in regular 2D geometries with full observations and lack support for more complex learning scenarios. Among commonly used benchmarks (Takamoto et al., 2022; Herde et al., 2024; Dulny et al., 2023) we surveyed none of these datasets simultaneously offer partial observability and multi-task evaluation capabilities. Although the *FWI-F/L/FL* (Zhu et al., 2023) suite introduces inverse modeling and partial observations, it remains confined to 2D domains with relatively simple geometries and lacks support for multitask learning. Similarly, *BubbleML* (Hassan et al., 2023) is an excellent multiphase, multiphysics dataset, yet it still lacks testing capabilities for inverse problems and OOD problems. Furthermore, mainstream and widely used benchmarks predominantly focus on low spatial dimensions(1D or 2D), which falls short of the complexity found in real-world scientific and engineering problems that often involve high-dimensional, irregular domains with heterogeneous observability and spatiotemporal dynamics.

Table 1: Comparison of PDE benchmark datasets. Each checkmark (✓) indicates the presence of a specific feature in the dataset. *Spatial Dim.* denotes the predominant dimensionality used in common benchmarks, not a limitation of the PDE itself.

Benchmark Dataset	Spatial Dim.	Inverse	Partial Obs.	Irregular Geo.	Multi-task	OOD
Advection	1D	✓	✗	✗	✗	✓
Darcy Flow	2D	✓	✗	✗	✗	✓
NavierStokes	1D/2D/3D	✓	✗	✗	✗	✓
Burgers' Equation	1D/2D	✓	✗	✗	✗	✓
Airfoil Flow	2D	✓	✗	✓	✗	✗
Diffusion Reaction	1D/2D	✓	✗	✗	✗	✓
Shallow Water	2D	✗	✗	✗	✗	✗
Plasticity / Elasticity	2D/3D	✗	✗	✓	✗	✗
BubbleML	2D/3D	✗	✗	✓	✓	✗
FWI-F / L / FL	2D	✓	✓	✗	✗	✓
Aletheia (Ours)	3D	✓	✓	✓	✓	✓

These limitations hinder comprehensive evaluation of model generalization, robustness, and versatility. In contrast, our proposed dataset, **Aletheia**, is designed to fill this gap by incorporating inverse tasks, partial and sparse observations, irregular 3D geometries with temporal dynamics, multi-task learning, and OOD generalization, providing a more realistic and challenging benchmark for PDE learning.

3 Dataset construction and details

3.1 Overview of the dataset

Our dataset is derived from two complementary sources. The first source consists of high-fidelity multi-physics simulations, which generate synthetic data by numerically solving a fully coupled electric–magnetic–thermal transient process. The second source comprises experimental measurements, where actual data are collected via frequency-swept pulsed eddy current thermography experiments performed on steel rail samples containing artificial defects. Due to the scarcity and complexity of real defect specimens, experimental data primarily serve to calibrate and validate the simulation

model. Conversely, simulated data are extensively utilized for model training as they offer complete three-dimensional temperature field evolutions, internal heat source distributions, and precise defect geometries—details which real ECPT experiments cannot provide, being limited to surface temperature time series measurements. Recognizing that the real experimental environment closely mimics authentic engineering inspection scenarios, we meticulously calibrated our simulation models. Specifically, we fine-tuned material properties and excitation parameters so that simulated surface temperature curves align closely with experimental measurements, achieving an accuracy within ±1°C. This rigorous calibration ensures both the reliability and realistic nature of our simulated dataset.

3.2 Simulation data acquisition

We used COMSOL Multiphysics[1] to develop a three-dimensional finite element simulation pipeline modeling the coupled process of electromagnetic induction heating and thermal diffusion. For each simulated sample, we first constructed a 3D rail model containing a defect and set material properties such as electrical conductivity σ, thermal conductivity k, and relative permeability μ_r. We then specified the defect geometry and input parameters, including the excitation frequency f and coil current. After the model was built, an adaptive mesh was generated with finer elements in the defect region due to the defect's small size to ensure computational accuracy. The simulation employed a 600 ms pulsed heating process. For more detailed simulation parameter settings, please refer to Section A. When an alternating current pulse at the specified frequency was applied to an excitation coil fixed just above the rail, it induced eddy currents in the metal specimens according to Faraday's law. Meanwhile, according to Ampere's Law, the magnetic field generated by induced eddy currents in turn affects the original field. The Maxwell equations can describe this electromagnetic process:

$$\nabla \cdot \mathbf{D} = \rho, \quad \nabla \cdot \mathbf{B} = 0, \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad \nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}. \tag{2}$$

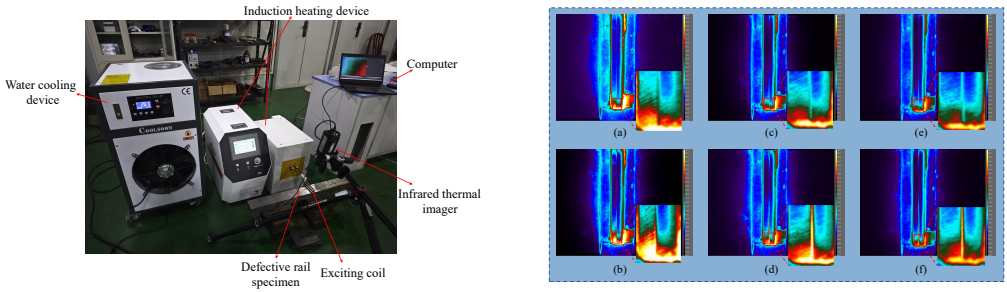
where $\mathbf{D}$ denotes the electric displacement vector, ρ is the free charge density; $\mathbf{B} = \mu\mathbf{H}$ represents the magnetic induction intensity; $\mathbf{H}$ is the magnetic field intensity; $\mathbf{E}$ is the electric field intensity; $\mathbf{J} = \sigma\mathbf{E}$ indicates the eddy current density. Eddy current inside the conductor generated Joule heat due to resistance, acting as an internal heat source: $q(x, y, z) = \frac{|\mathbf{J}|^2}{\sigma}$. After converting this electromagnetic energy into thermal energy, it diffused through thermal conduction in the material.

The depth of eddy current penetration, known as the skin depth, is frequency-dependent and given by: $\delta = \sqrt{\frac{1}{\pi f \mu \sigma}}$, where $\mu = \mu_r \mu_0$ is the magnetic permeability (μ_0 is the permeability of free space). Higher frequencies reduce δ, concentrating eddy currents and heat near the surface, while lower frequencies allow deeper penetration, enabling differentiation of subsurface defect responses. To overcome the inherent ill-posedness of mapping surface temperature back to defect characteristics, we employ multiple excitation frequencies in the pulsed heating process—while some frequencies may yield indistinguishable temperature profiles for certain defect types, others reveal distinct thermal distributions, enabling us to accurately distinguish different types of defects. The electromagnetic and thermal phenomena within our simulations were fully coupled, mutually influencing one another throughout the transient process. Upon completion of each simulation, we systematically recorded the temporal evolution of the temperature field $u(x, y, z, t)$ across the specimen's surface and internal volume, as well as the spatial distribution of the internal heat sources $q(x, y, z)$ generated by induced eddy currents.

3.3 Experimental data acquisition

We additionally performed pulsed eddy current thermography experiments to acquire authentic measurement data. The experimental apparatus comprised a custom-built XZ-series DSP-controlled inverter power supply paired with a specially designed excitation coil featuring a central slot, as depicted in Figure 3. During each experimental run, a sinusoidal pulse current with a duration of 600 ms was applied at selected frequencies within the range of 1—100 kHz. The excitation coil was maintained at a constant lift-off distance of 5 mm above the steel rail sample. A high-resolution FLIR SC6550A infrared thermal imager was employed to capture the evolution of surface temperatures,

[1] https://www.comsol.com



(a) Experimental setup for pulse eddy current thermal imaging.

(b) Thermal imaging results for different defect types and excitation frequencies.

Figure 3: Overview of the pulse eddy current thermal imaging experimental platform and selected thermal imaging results.

providing imagery at a spatial resolution of 640×480 pixels, a frame rate of 50 frames per second, and an intensity depth of 14 bits. This allowed detailed recording of the temporal temperature distribution throughout both the heating and cooling phases. The collected experimental data served to introduce realistic ambient noise into our dataset and was used to rigorously validate the accuracy and reliability of our simulation model.

3.4 Dataset structure

Leveraging our calibrated simulation framework, we systematically generated a comprehensive set of defect samples using automated scripts and parallelized computations. After extensive simulations and rigorous calibration processes, the final dataset comprises a total of 4,782 unique defect instances, categorized into 2,407 open-crack and 2,375 closed-crack samples. Detailed statistics for each defect type are summarized in Table 2. Each defect instance includes simulation data captured across ten discrete excitation frequencies—specifically, 1 kHz, 4 kHz, 9 kHz, 16 kHz, 25 kHz, 36 kHz, 49 kHz, 64 kHz, 81 kHz, and 100 kHz—to reflect varied depth sensitivities and thermal responses. For versatility across different modeling approaches, the thermal and temperature fields within each defect instance were sampled using two distinct strategies: regular grid sampling and irregular point sampling. Regular grid sampling aligns directly with the structured grid employed by experimental infrared imagery, providing consistent surface temperature time-series data for model input. Conversely, the irregular sampling strategy mimics the actual positioning of defects relative to the infrared camera used in experimental setups. These irregularly sampled points follow a multivariate Gaussian distribution concentrated around the central surface line of the specimen, accurately reflecting defect locations encountered in practical inspections. Points near defect regions are sampled densely to ensure higher resolution and precise reconstruction of defect contours, whereas areas less relevant to defect detection are sampled more sparsely, thus reducing computational overhead. This carefully designed sampling method optimally balances computational efficiency with reconstruction accuracy.

Table 2: Number of different types of defects. There are a total of 6 types of defects, please refer to Section C for more detailed information on each type of defect.

Single layer	Type I double-layer	Type II double-layer	Type III double-layer	Type I multi-layer	Type II multi-layer	Total
88	1050	69	2285	90	1200	4782

4 Benchmark

To systematically assess the modeling capabilities of neural operators in 3D spatiotemporal heat transfer problems, we construct a comprehensive benchmark suite based on the Aletheia dataset. This benchmark spans six task settings, encompassing both *Same-Frequency* and *Out-of-Distribution*

scenarios. Each task is evaluated on two types of spatial grids: regular and irregular to rigorously test the generalization and robustness of neural operator models across diverse spatial discretizations. We benchmark several representative neural operator architectures, including FNO and Transolver, under consistent training and evaluation protocols. This unified framework offers a strong baseline for future research and provides critical insight into the strengths and limitations of current neural PDE solvers in realistic NDT contexts.

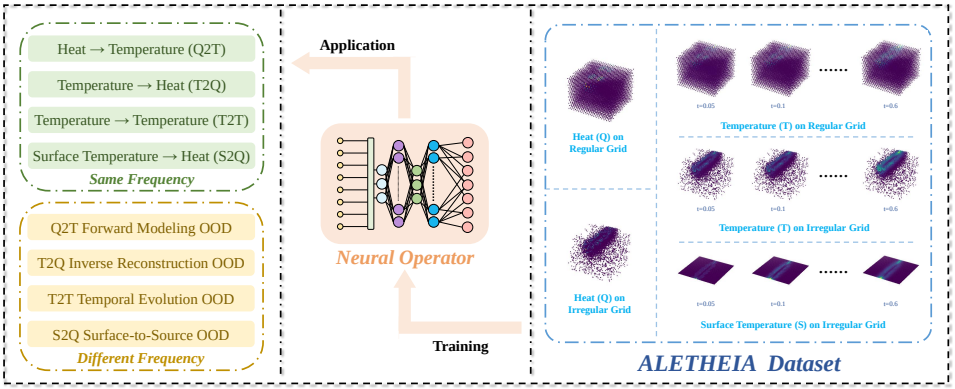


Figure 4: Our benchmark covers tasks such as heat source to temperature, temperature reconstruction, surface temperature reconstruction, and out-of-distribution detection at different frequencies.

4.1 Task Setup

We organize the tasks into two categories according to the alignment between training and testing distributions:

1. Full-frequency tasks (in-distribution / closed-set) These tasks evaluate the accuracy of the model when the training and testing data are sampled from the same excitation frequencies and boundary conditions:

- **Forward Modeling:** Learn the mapping $q(\mathbf{x}, t) \mapsto u(\mathbf{x}, t)$, i.e., predict the transient temperature field given the internal heat source.
- **Inverse Source Reconstruction:** Learn the inverse mapping $u(\mathbf{x}, t) \mapsto q(\mathbf{x}, t)$, recovering the latent volumetric heat source from the transient temperature field $u(x, y, z, t)$.
- **Temporal Evolution Prediction:** Learn $\{u(\mathbf{x}, t_i)\}_{i=0}^{10} \mapsto \{u(\mathbf{x}, t_j)\}_{j=11}^{12}$, forecasting future temperature fields based on historical states.
- **Surface-to-Source Reconstruction:** Learn $s(\mathbf{x}_\partial, t) \mapsto q(\mathbf{x}, t)$, Reconstructing the implicit volumetric heat source from the top surface temperature trace $s(x, y, z, t)$, when the values of z are all 0.

2. Out-of-distribution (OOD) generalization tasks These tasks introduce distributional shifts to test generalization to unseen physical conditions, such as novel excitation frequencies or altered initial/boundary settings:

- **Forward OOD Generalization:** Predict $u(\mathbf{x}, t)$ from $q(\mathbf{x}, t)$ when the excitation frequency used at test time differs from those seen during training.
- **Inverse OOD Generalization:** Recover $q(\mathbf{x}, t)$ from $u(\mathbf{x}, t)$ when testing using excitation frequencies outside the distribution of the training set.
- **Temporal OOD Generalization:** Predict future thermal states from earlier temperature observations when historical or boundary conditions are shifted.
- **Surface-to-Source OOD Generalization:** Recover $q(\mathbf{x}, t)$ from $s(\mathbf{x}_\partial, t)$ when testing using excitation frequencies outside the distribution of the training set.

To further enhance diversity and realism, each task is evaluated on both **regular** and **irregular** spatial grids. This enables us to systematically investigate the robustness of neural operators under varying spatial discretization schemes, mimicking real-world sensing constraints.

4.2 Experimental Setup

We conducted benchmarks on both regular and irregular 3D geometric datasets generated from two-layer defect simulations. Each sample was downsampled to 8000 unstructured points, and each experiment involved 600 samples. Prior to training, all input data were normalized using global statistical measures.

To handle the dataset, we developed a custom VTU parser that supports normalization and enables two data partitioning strategies: Full-Frequency (FF) and Out-of-Distribution (OOD). In the FF setting, the entire dataset is loaded and randomly shuffled at the sample level to maintain statistical uniformity across training and testing sets. In the OOD setting, each sample group contains 10 simulations, each corresponding to a different frequency. Within each group, 80% of the samples (i.e., the first 8 frequencies) are used for training, while the remaining 20% (i.e., the last 2 frequencies) are reserved for testing. Although the internal order of samples may vary due to file loading mechanisms, the frequency-based grouping is consistent across all computational environments. This structured split ensures the model is evaluated on frequencies not encountered during training, enabling a more rigorous assessment of generalization under distribution shifts.

We compare a range of baseline and state-of-the-art neural operator models. These include Fourier-based models such as FNO (Li et al., 2021), FFNO (Tran et al., 2023), FCNO (Li et al., 2024), and GeoFNO (Li et al., 2023b), as well as attention-based models like LNO (Wang & Wang, 2024) and Transolver (Wu et al., 2024). Additionally, we include DeepONet and a simple MLP as baseline models to provide a fair and comprehensive comparison.

All models were trained under the same configurations, please refer to Section F for more detailed configurations.

4.3 Experimental Results

We show the performance of representative operator learning models from multiple perspectives: in-distribution vs OOD generalization, inverse task complexity (T2Q), and metric diversity that spans global error (MSE, RMSE), structural fidelity (SSIM), and extreme case sensitivity (Max, bRMSE) in Figure 5. Detailed quantitative results , more experimental results and visualizations can be found in Section H. The experiments were trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

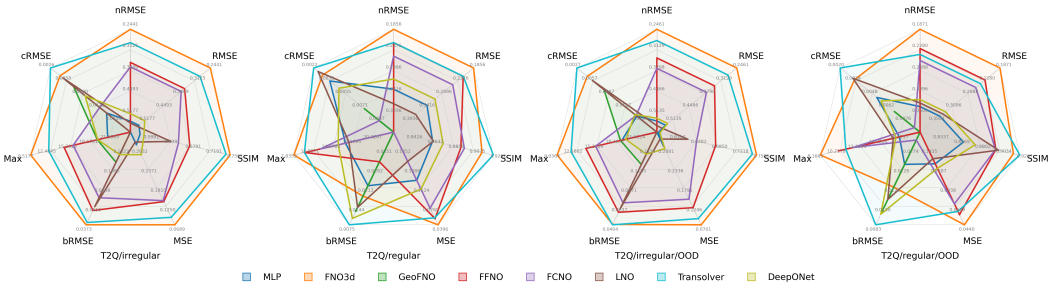


Figure 5: Comparison of the performance of eight models under the temperature to heat (T2Q) inversion task based on seven evaluation metrics.

Among the four T2Q tasks, FNO demonstrated the most reliable performance in recovering volumetric sources from whole-field temperature data. This performance exhibits pronounced spectral-induced bias characteristics: the globally fourier layer encoded smoothing of frequency-structured thermodynamics enables effective extrapolation during excitation displacement, thereby maintaining stable performance in the inverse mapping of observed complete three-dimensional temperature fields. Even when task difficulty increases due to irregular grids or frequencies beyond the design range, FNO maintains an outer envelope on core error axes. This indicates spectral operators capture large-scale

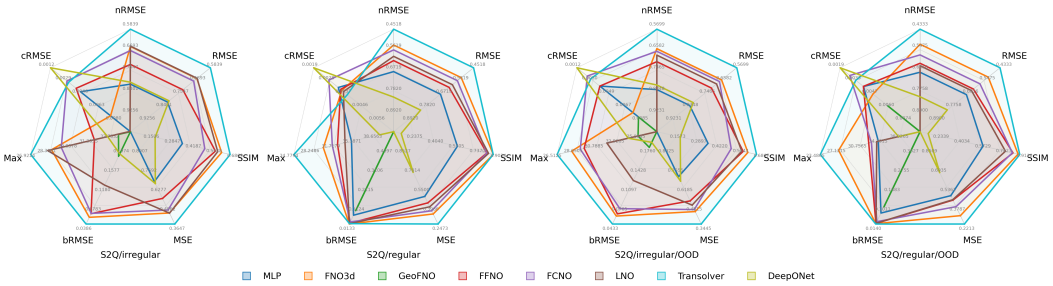


Figure 6: Comparison of the performance of eight models under the surface temperature to heat (S2Q) inversion task based on seven evaluation metrics.

thermal conduction with minimal aliasing effects while preserving global accuracy more effectively than alternative methods.

Transolver demonstrates superior performance in 'surface-to-source' scenarios (S2Q) where only the surface temperature is accessible and internal states remain unobservable. Across regular and irregular layouts, as well as under out-of-distribution conditions, it consistently achieves optimal performance across multiple metrics, reflecting its ability to more faithfully reconstruct internal heat sources from boundary signals. Its attention-based data-dependent weights generate adaptive receptive fields that simultaneously fuse long-range cross-surface coupling effects and highlight critical boundary features, effectively suppressing pseudo-correlations induced by sparse or heterogeneous sampling.

It is noteworthy that, compared to T2Q with full-field observations, S2Q only accesses top surface temperatures and cannot directly observe the interior, substantially reducing identifiability and exacerbating ill posedness. Empirically, all models degrade on S2Q, with the drop most pronounced on irregular meshes and under OOD excitations. It proves that limited observations and sparse/heterogeneous sampling further weaken the inverse problem's identifiability and promote error accumulation across space and frequency. This mirrors the core challenge of NDT: inferring internal heat sources/defects from boundary-only measurements is inherently information-limited.

In summary, operator choice should match the characteristics of the task. FNO is preferable when full-field observations are available and the goal is high global accuracy or out-of-distribution extrapolation of inverse mappings. In contrast, Transolver excels when measurements are confined to surfaces or irregularly sampled, and the priority lies in preserving structural fidelity and controlling worst-case risks. The two operator types are thus complementary: spectral operators reduce displacement errors, while attention operators safeguard structural integrity and robustness under heterogeneous sensing conditions.

5 Conclusion

We have developed and released the first large-scale, multi-frequency coupled 3D Pulsed Eddy Current Thermography benchmark dataset, aimed at advancing the application of machine learning methods in internal crack detection for NDT. This dataset is derived from high fidelity electromagnetic thermal multiphysics simulation data and calibrated with real infrared thermal imaging experimental data, covering various defect types and frequency response scenarios. It supports a range of forward and inverse modeling tasks. Benchmark results show that neural operator models can effectively learn the dynamic process of heat diffusion and reconstruct the primary internal heat source distribution from surface temperature sequences. However, challenges remain in accurately reconstructing fine-grained three-dimensional morphologies of complex defects.

This dataset is the first to encompass the electromagnetic-thermal coupling response of rail materials under different excitation frequencies, providing a standardized testing platform for research into neural network surrogate models, operator learning methods, and defect inversion. Looking ahead, we plan to extend this framework to additional materials, such as composites and pipeline structures, as well as to a broader range of NDT scenarios, further promoting the widespread application of data-driven modeling in industrial inspections. We hope that the Aletheia benchmark will serve as a foundational resource for multi-physics modeling and model generalization research.

Ethics Statement

This work adheres to the ICLR Code of Ethics. Our study concerns methodological advances for learning neural operators on thermo-electromagnetic simulations and lab measurements of inanimate rail specimens. It involves no human subjects, personally identifiable information, or sensitive user data. Real measurements are acquired from controlled ECPT experiments on metal rails; no living beings are involved, and the procedures present no biological, environmental, or privacy risks. The proposed benchmark aims to improve the reliability of inverse modeling for nondestructive testing. While the dataset and models may inform industrial inspection workflows, they do not directly enable harmful applications. Any future deployment in safety-critical domains must consider regulatory, ethical, and societal constraints beyond the scope of this work (e.g., responsible use, failure modes under distribution shift). We report all methods and results transparently and disclose no conflicts of interest or external sponsorship. All experiments were designed and conducted in accordance with standards of research integrity.

Reproducibility Statement

We provide the information necessary to reproduce our results. Dataset construction details (simulation pipeline, experimental setup, sampling strategies, frequencies, and annotations) are described in Section 3 and Sections A to C. Task definitions, data splits, data distribution types, and evaluation metrics are specified in Section 4 and Section E. For each model, we list architectures, hyperparameters, training schedules, and preprocessing/normalization in Section 4.2, with additional tables in Section F. We report the exact point counts and time steps, and use standardized evaluation scripts. The dataset and anonymized code will be made publicly available together with scripts to reproduce all tables and figures.

References

Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024.

Xiaotian Chen, Guiyun Tian, Song Ding, Junaid Ahmed, and Wai Lok Woo. Tomographic reconstruction of rolling contact fatigues in rails using 3d eddy current pulsed thermography. *IEEE Sensors Journal*, 21(17):18488–18496, 2021.

Jingui Cheng, Lei Xu, and Li Chao. A review of two types of non-destructive testing technique for pressure pipelines. *Insight-Non-Destructive Testing and Condition Monitoring*, 63(6):326–333, 2021.

Andrzej Dulny, Andreas Hotho, and Anna Krause. Dynabench: A benchmark dataset for learning dynamical systems from low-resolution data. Jun 2023.

Yuan Gao, Zheng Liang, Liang Zhang, Ting Zheng, Jiawei Zhou, and Jiyu Zheng. Quantification of depth and morphology of internal corrosion defects by stepped eddy current thermography skewness under weak excitation conditions. *Measurement*, 229:114454, 2024.

Lijalem Gebrehiwet, Abdi Chimido, Wondmagegn Melaku, and Eshet Tesfaye. A review of common aerospace composite defects detection methodologies. *Int. J. Res. Publ. Rev*, 4:1829–1846, 2023.

Samira Gholizadeh and S Gholizadeh. Impact behaviours and non-destructive testing (ndt) methods in carbon fiber composites in aerospace industry: a review. *Authorea Preprints*, 10, 2022.

Wendong Gong, Muhammad Firdaus Akbar, Ghassan Nihad Jawad, Mohamed Fauzi Packeer Mohamed, and Mohd Nadhir Ab Wahab. Nondestructive testing technologies for rail inspection: A review. *Coatings*, 12(11):1790, 2022.

Mridul Gupta, Muhsin Ahmad Khan, Ravi Butola, and Ranganath M Singari. Advances in applications of non-destructive testing (ndt): A review. *Advances in Materials and Processing Technologies*, 8 (2):2286–2307, 2022.

Sheikh Md Shakeel Hassan, Arthur Feeney, Akash Dhruv, Jihoon Kim, Youngjoon Suh, Jaiyoung Ryu, Yoonjin Won, and Aparna Chandramowlishwaran. Bubbleml: A multi-physics dataset and benchmarks for machine learning. *arXiv preprint arXiv:2307.14623*, 2023.

Maximilian Herde, Bogdan Raonic, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for pdes. *Advances in Neural Information Processing Systems*, 37:72525–72624, 2024.

Yang Hong, Yicheng Ma, Shuang Wen, and Zhiqiang Sun. A reconstructed approach for online prediction of transient heat flux and interior temperature distribution in thermal protect system. *International Communications in Heat and Mass Transfer*, 148:107055, 2023.

Geo Jacob and Florian Raddatz. Data fusion for the efficient ndt of challenging aerospace structures: a review. *NDE 4.0, Predictive Maintenance, and Communication and Energy Systems in a Globally Networked World*, 12049:126–135, 2022.

Seungjun Lee and Taeil Oh. Inducing point operator transformer: A flexible and scalable architecture for solving pdes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 153–161, 2024.

Ye Li, Ting Du, Yiwen Pang, and Zhongyi Huang. Component fourier neural operator for singularly perturbed differential equations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13691–13699, 2024.

Zijie Li, Dule Shu, and Amir Barati Farimani. Scalable transformer for pde surrogate modeling. *Advances in Neural Information Processing Systems*, 36:28010–28039, 2023a.

Zongyi Li, NikolaB. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, AndrewM. Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *International Conference on Learning Representations,International Conference on Learning Representations*, Feb 2020.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for pdes on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023b.

Yiping Liang, Libing Bai, Lulu Tian, Xu Zhang, and Yong Gao. Thermal parameter reconstruction imaging for interlayer defect detection in ecpt. *IEEE Transactions on Industrial Informatics*, 2024.

Qiang Lin, Shenghui Jiang, Haidong Tian, Haohao Ding, Wenjian Wang, Jun Guo, and Qiyue Liu. Study on non-destructive testing of rail rolling contact fatigue crack based on magnetic barkhausen noise. *Wear*, 528:204965, 2023.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *Nature Machine Intelligence*, pp. 218–229, Mar 2021. doi: 10.1038/s42256-021-00302-5. URL http://dx.doi.org/10.1038/s42256-021-00302-5.

Xiaojie Ma, Song Ding, Yiqing Wang, Cheng Song, Qing Zhang, and Jie Shen. Characterisation and evaluation of paint-coated marine corrosion in carbon steel based on pulsed eddy current thermography and bemd noise-reducing method. *Nondestructive Testing and Evaluation*, pp. 1–17, 2024.

Roberto Molinaro, Yunan Yang, Björn Engquist, and Siddhartha Mishra. Neural inverse operators for solving pde inverse problems. In *International Conference on Machine Learning*, pp. 25105–25139. PMLR, 2023.

Maria Inês Silva, Evgenii Malitckii, Telmo G Santos, and Pedro Vilaça. Review of conventional and advanced non-destructive testing techniques for detection and characterization of small-scale defects. *Progress in Materials Science*, 138:101155, 2023.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, Oct 2022.

Xian Tao, Dapeng Zhang, Wenzhi Ma, Zhanxin Hou, ZhenFeng Lu, and Chandranath Adak. Unsupervised anomaly detection for surface defects with dual-siamese network. *IEEE Transactions on Industrial Informatics*, 18(11):7707–7717, 2022.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tmIiMPl4IPa.

Christoph Tuschl, Beate Oswald-Tranta, and Sven Eck. Inductive thermography as non-destructive testing for railway rails. *Applied Sciences*, 11(3):1003, 2021.

Subhash Utadiya, Vismay Trivedi, Gyanendra Sheoran, Atul Srivastava, Daniel Claus, Humberto Cabrera, and Arun Anand. Digital holographic imaging of thermal signatures and its use in inhomogeneity identification. *Optics and Lasers in Engineering*, 160:107227, 2023.

Gang Wang, Xiangjie Qin, Dongyang Han, and Zhiyuan Liu. Study on seepage and deformation characteristics of coal microstructure by 3d reconstruction of ct images at high temperatures. *International Journal of Mining Science and Technology*, 31(2):175–185, 2021.

Tian Wang and Chuang Wang. Latent neural operator for solving forward and inverse pde problems. *arXiv preprint arXiv:2406.03923*, 2024.

Yuqin Wang, Fei Song, Qingshan Feng, Weibiao Qiao, Shaohua Dong, Yangyang Jiang, and Qianli Ma. Basic theory and applications of oil and gas pipeline non-destructive testing methods. *Energies*, 17(24):6366, 2024.

Keith A Woodbury, Hamidreza Najafi, Filippo De Monte, and James V Beck. Inverse heat conduction: Ill-posed problems. 2023.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for pdes on general geometries, 2024. URL http://arxiv.org/abs/2402.02366.

Zhenning Wu, Yiming Deng, Jinhai Liu, and Lixing Wang. A reinforcement learning-based reconstruction method for complex defect profiles in mfl inspection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–10, 2021.

Zipeng Xiao, Siqi Kou, Hao Zhongkai, Bokai Lin, and Zhijie Deng. Amortized fourier neural operators. *Advances in Neural Information Processing Systems*, 37:115001–115020, 2024.

Longhui Xiong, Guoqing Jing, Jingru Wang, Xiubo Liu, and Yuhua Zhang. Detection of rail defects using ndt methods. *Sensors*, 23(10):4627, 2023.

Hui Yin, Jianping Peng, Xiang Zhang, Kang Tian, Yu Zhang, and Jianqiang Guo. Quantification of closed cracks in railways using eddy current pulsed thermography. *Applied Optics*, 60(17): 5195–5202, 2021.

Fei Yuan, Yating Yu, Linfeng Li, and Guiyun Tian. Investigation of dc electromagnetic-based motion induced eddy current on ndt for crack detection. *IEEE sensors journal*, 21(6):7449–7457, 2021.

Xiang Zhang, Jianping Peng, Luquan Du, Jie Bai, Lingfan Feng, Jianqiang Guo, and Xiaorong Gao. Detection of fatigue microcrack using eddy current pulsed thermography. *Journal of Sensors*, 2021 (1):6647939, 2021.

Xu Zhang, Libing Bai, Yiping Liang, Jinsheng Yang, Jiangshan Ai, and Chao Ren. Surface defect 3-d profile reconstruction using eddy current pulsed thermography. *IEEE Transactions on Industrial Informatics*, 2024.

Jianming Zhao, Wei Li, Xin'an Yuan, Xiaokang Yin, Xiao Li, Qinyu Chen, and Jianxi Ding. An end-to-end physics-informed neural network for defect identification and 3-d reconstruction using rotating alternating current field measurement. *IEEE Transactions on Industrial Informatics*, 19 (7):8340–8350, 2022.

Min Zhu, Shihang Feng, Youzuo Lin, and Lu Lu. Fourier-deeponet: Fourier-enhanced deep operator networks for full waveform inversion with improved accuracy, generalizability, and robustness. *Computer Methods in Applied Mechanics and Engineering*, 416:116300, 2023.

Xingxing Zou, Libin Wang, Jiaqing Wang, Jie Liu, Hao Ma, and Yi Bao. Nondestructive evaluation of carbon fiber reinforced polymer (cfrp)-steel interfacial debonding using eddy current thermography. *Composite Structures*, 284:115133, 2022.

Ruili Zu, Yang Yang, Xianfu Huang, Dacheng Jiao, Jiaye Zhao, and Zhanwei Liu. A stress detection method for metal components based on eddy current thermography. *NDT & E International*, 133: 102762, 2023.

A Simulation Experiment Details

In COMSOL Multiphysics, the simulation domain consisted of a steel rail geometry with dimensions set to 100 $\times 70 \times 45$ which is embedded artificial defects of configurable size, orientation, and depth, surrounded by air domain; The rail material used in the physical experiments was U71Mn steel, and the material parameters used in the simulation were modified from U71Mn properties based on calibration against experimental surface temperature measurements. The constant material properties for steel, air, and excitation coil were assigned as listed in Table 3. All external surfaces of the rail were thermally insulated except for the top surface, which had a convective cooling boundary condition to ambient air, with a heat transfer coefficient set to 5 W/(m^2·K); the excitation coil was driven via an applied current boundary condition to simulate inductive heating. The initial temperature of the entire model was uniformly set to 20 °C. A free tetrahedral mesh was employed, featuring fine elements around defect regions and within the electromagnetic skin depth, while coarser elements were used elsewhere; an adaptive refinement scheme further adjusted the mesh based on local conductivity gradients. The simulation was run using a time-dependent solver with a backward differentiation formula (BDF) scheme, covering a time span from 0 to 0.6 s with a fixed time step of 0.05 s.

Table 3: Material property parameters. In the ECPT experiment, due to the extremely short heating process and relatively small temperature rise changes, the impact on the various properties of the material can be ignored. Therefore, the material's property parameters are set to a fixed constant. $\sim$ represents the physical quantity of the material that is not involved in actual calculations.

Material	Conductivity (S/m)	Relative permeability	Relative dielectric constant	Thermal conductivity (W/m·K)	Density (kg/m^3)	Constant pressure heat capacity (J/kg·K)
Steel	1.3×10^7	200	2	48	8000	450
Air	0	1	1	0.0257	1.205	1005
Copper	5.998×10^7	1	1	$\sim$	$\sim$	$\sim$

B Data Generation and Details

The diversity of defect samples in our dataset is achieved through systematic variations in defect orientation, depth, and length. For single-layer defects, the angle between the defect extension direction and the upper surface differs from sample to sample. In double-layer and multi-layer defects, not only do the individual cracks vary in their inclination to the surface, but the angles between multiple cracks also change. These controlled variations result in a wide range of geometric configurations. Detailed statistical parameters of defect angles are presented in Table 4.

Table 4: Angle parameters for different types of defects. Among them, angle 1 represents the angle formed between the defect direction and the upper surface, while angles 2 and 3 represent the angles between multiple cracks. All units are based on angle system.

Defect type	**Range of different angles**		
	angle I	angle II	angle III
Single layer defect	3∼90	-	-
Type I double-layer defect	24.47∼75.47	5.72∼81.27	-
Type II double-layer defect	3∼71	-	-
Type III double-layer defect	10.06∼80.06	0.83∼24.03	-
Type I multi-layer defect	0∼89	-	-
Type II multi-layer defect	30∼90	12.09∼31.1	12.12∼31.12

The dataset contains a total of 4,782 samples, with an overall size of approximately 1.89 terabytes. These samples are organized into six main folders, each corresponding to a specific defect type and named accordingly. Each sample folder is labeled by its defect type followed by one or more

angle values that indicate the orientation of the defect. For example, names may include a single angle for single-layer defects or two to three angles for more complex configurations. Within each sample folder, there are three subfolders storing different types of data. The first subfolder contains surface temperature data sampled on a structured grid. The second provides full three-dimensional temperature and heat field data, also on a structured grid with a uniform spacing of one unit. The third subfolder includes similar 3D data, but sampled on an unstructured grid. The unstructured sampling strategy, which distributes 50,000 points per sample, is detailed in the main text. Table 5 summarizes the number of sampling points for each data type.

Table 5: Sampling point statistics for different types of data. Here, $\mathrm{T}_{\mathrm{surf}}$ denotes surface temperature measurements, $\{q(\mathbf{x},t), u(\mathbf{x},t)\}_{\text{structured}}$ represents paired heat source and temperature field data sampled on regular grids, and $\{q(\mathbf{x},t), u(\mathbf{x},t)\}_{\text{unstructured}}$ denotes the same sampled irregularly.

Data type	x	y	z	Sampling points
$\mathrm{T}_{\mathrm{surf}}$	150	105	-	10750
$\{q,u\}_{\text{structured}}$	50	30	30	45000
$\{q,u\}_{\text{unstructured}}$	-	-	-	50000

Each sample is simulated under ten excitation frequencies ranging from 1 to 100 kHz, with each frequency corresponding to one VTU format file. Both surface and volumetric temperature data are time-resolved from 0 to 0.6 seconds, with a time step of 0.05 seconds, resulting in 13 time-series data stored in each VTU file. This structure ensures that the dataset supports diverse tasks involving spatial, temporal, and multi-frequency analysis.

C Defect Details

Our dataset encompasses six distinct internal crack defect types, spanning single-layer, double-layer, and multi-layer configurations and including both open and closed crack cases. Six types of defects are shown in Figure 7 and the detailed parameters of each defect are shown in Table 6.

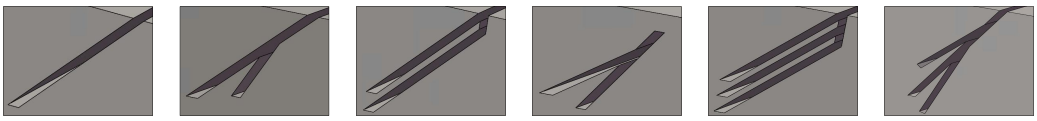


Figure 7: 6 types of defects: (a)Single layer defect (b)Type I double-layer defect (c)Type II double-layer defect (d)Type III double-layer defect (e)Type I multi-layer defect (f)Type II multi-layer defect

Table 6: Detailed parameters of 6 types of defects. The units of depth, width, and length are all in millimeters. In double-layer and multi-layer defects, the length of the defect refers to the length of the longest crack.

Defect type	Sample size	Depth	Width	Length	Open/Closed
Single layer defect	88	0.2∼3.82	0.1∼0.2	3.82	Open
Type I double-layer defect	1050	1.17∼4.1	0.2	4.53	Open
Type II double-layer defect	69	0.24∼4.02	0.1∼0.2	4.22	Open
Type III double-layer defect	2285	1.40∼3.75	0.2	3.29	Closed
Type I multi-layer defect	90	0.81∼2.41	0.2	2.4	Closed
Type II multi-layer defect	1200	1.89∼3.12	0.15	3.12	Open

D DEFECT RECONSTRUCTION FROM HEAT FIELD

We conducted a simple test using a basic 3D CNN to reconstruct the heat source field $Q(x, y, z)$. The results demonstrate that Q can effectively capture defect morphology: high-Q-value dense zones align with defect surfaces, Q-value decay indicates defect propagation direction, and low-Q-value regions correspond to internal cavities. Leveraging a 3D CNN with a large receptive field, we extract global spatial features from the reconstructed Q-field and use a regression network to predict key defect parameters such as angle, length, and depth. Experimental results (Table 7) show reliable predictions for these parameters, with length exhibiting the highest accuracy due to its distinct extension feature in the Q-field. The width parameter shows low variance in the dataset, resulting in negligible predictive signal rather than a model limitation. Therefore, as long as our neural operator can reconstruct Q accurately, it can serve as a reliable basis for predicting multidimensional defect characteristics.

Table 7: Regression evaluation indicators for crack parameters

Attributes	Model	MSE	RMSE	MAE
Crack angle	MLP	0.23	0.48	0.38
	FNO	0.17	0.41	0.35
	Transolver	0.20	0.44	0.30
Crack length	MLP	0.0020	0.044	0.034
	FNO	0.0012	0.035	0.028
	Transolver	0.0011	0.033	0.020
Crack depth	MLP	1.24	1.11	0.91
	FNO	0.85	0.92	0.80
	Transolver	1.02	1.01	0.70
Crack width	MLP	1.00×10^{-14}	1.00×10^{-7}	7.86×10^{-8}
	FNO	1.26×10^{-14}	1.12×10^{-7}	1.10×10^{-7}
	Transolver	1.43×10^{-14}	1.19×10^{-7}	1.16×10^{-7}

E METRICS

Standard methods for calculating the root mean square error (RMSE) of test data fail to capture important optimization criteria in scientific machine learning. It is not enough to fit (usually sparse) data well if the physical laws of the underlying problem are seriously violated. Therefore, they must be evaluated using appropriate metrics. Furthermore, a single evaluation metric is not sufficient to compare differences in the ability of different methods to infer unseen time steps and parameters, which are important but not yet fully explored evaluation criteria for machine learning alternative models. We used the PDEBench evaluation metrics and the SSIM evaluation metrics, as shown in Table 8

The normalized RMSE is a variant of the RMSE to provide scale-independent information defined as:

$$\text{nRMSE} \equiv \frac{\|u_{\text{pred}} - u_{\text{true}}\|_2}{\|u_{\text{true}}\|_2}, \tag{3}$$

where $\|u\|_2$ is the L_2-norm of a (vector-valued) variable u, and $u_{\text{true}}, u_{\text{pred}}$ are true and predicted values, respectively. The maximum error measures the model's worst prediction, which quantifies both local performance and models' stability of their prediction.

cRMSE is defined as

$$\text{cRMSE} \equiv \frac{\|\sum u_{\text{pred}} - \sum u_{\text{true}}\|_2}{N}, \tag{4}$$

which measures the deviation of the prediction from some physically conserved value.

bRMSE measures the error at the boundary, indicating if the model understands the boundary condition properly.

Table 8: Evaluation indicators provided by PDEBench

Scope	Acronym	Metric
Data view	RMSE	root-mean-squared-error
	nRMSE	normalized RMSE (ensuring scale independence)
	max error	maximum error (local worst case; also proxy for stability of time-stepping)
Physics view	cRMSE	RMSE of conserved value (deviation from conserved physical quantity)
	bRMSE	RMSE on boundary (whether boundary condition can be learned)
	fRMSE low	RMSE in Fourier space, low frequency regime (wavelength dependence)
	fRMSE mid	RMSE in Fourier space, medium frequency regime
	fRMSE high	RMSE in Fourier space, high frequency regime

Finally, fRMSE measures the error in low/middle/high-frequency ranges defined as:

$$\sqrt{\frac{\sum_{k_{\min}}^{k_{\max}} |\mathcal{F}(u_{\text{pred}}) - \mathcal{F}(u_{\text{true}})|^2}{k_{\max} - k_{\min} + 1}}, \tag{5}$$

where $\mathcal{F}$ is a discrete Fourier transformation, and $k_{\min}, k_{\max}$ are the minimum and maximum indices in Fourier coordinates.

In PDEBench paper, the low/middle/high-frequency regions are defined as:

- Low: $k_{\min} = 0$, $k_{\max} = 4$
- Middle: $k_{\min} = 5$, $k_{\max} = 12$
- High: $k_{\min} = 13$, $k_{\max} = \infty$

This allows a quantitative discussion of the model performance's dependence on the wavelength. In the multidimensional cases, the $|\mathcal{F}(u_{\text{pred}} - u_{\text{true}})(k)|^2$ in the angular coordinate direction is first integrated and summed along the k coordinate.

F Experimental Setup Details

To ensure a fair comparison, this section details all hyperparameters and training configurations employed across the models. For the FNO family of models, given their similar parameter sets, these hyperparameters are presented in Table 9.

Table 9: Shared architectural settings for the FNO family baselines. All models use the same spectral bandwidth and channel widths to ensure fairness. (m_x, m_y, m_z) denotes spectral modes, $Width$ denotes the linear transformation applied on the spatial domain, $Grid/scale$ denotes the resolution of the gird.

Model	(m_x, m_y, m_z)	$Width$	$Grid/scale$	Model Variant
FNO	$(12, 12, 8)$	32	$(20, 20, 20)$	Original Fourier Neural Operator
GeoFNO	$(12, 12, 8)$	32	$(20, 20, 20)$	Geometry-Aware FNO
FFNO	$(12, 12, 8)$	32	$(20, 20, 20)$	Factorized FNO
FCNO	$(12, 12, 8)$	32	$(20, 20, 20)$	Factorized Cosine Neural Operator

Transolver. We instantiate Transolver with hidden size $n_{hidden} = 256$, depth $n_{layers} = 8$, and spatial dimension $space_{dim} = 3$. Multi-head attention uses $n_{head} = 8$ with an MLP expansion ratio of 2. We further use $slice_{num} = 32$ and disable positional unification. The same configuration is used for all tasks.

LNO. The LNO is consisting of $n_{block} = 4$ operator blocks with spectral resolution $n_{mode} = 256$ and hidden width $n_{\text{dim}} = 128$. Attention uses $n_{\text{head}} = 8$ and $n_{layer} = 2$ transformer layers with `GELU` activations and the vanilla attention kernel; temporal modeling is disabled.

MLP. We employ a point-wise multilayer perceptron with hidden width 32, and 4 layers in total, serving as a lightweight regression baseline.

DeepONet. The DeepONet baseline adopts a branch–trunk decomposition with width 32 in both subnetworks. The branch network has depth 2; the trunk network has depth 3, receiving the functional input of dimension 13 and producing a scalar output of dimension 1. All other settings follow the original implementation.

All models were trained under the same configuration: batch size of 1, learning rate of 0.0001, and 200 training epochs. Training employed the Adam optimizer alongside the OneCycleLR learning rate scheduler. All experiments were conducted on a single NVIDIA RTX 4090 GPU with 24GB of memory. The models were trained using mean squared error (MSE) loss, and evaluated using SSIM along with several PDEBench metrics. Regarding the part of evaluation indicators, lease refer to Section E for detailed definitions.

Detailed configurations reference Table 10. Apart from DeepONet, each model permits customisation of input and output channels. Due to DeepONet's constraint that its output channel must be limited to 1, DeepONet participated solely in the T2Q and S2Q tasks.

Table 10: Training configuration across tasks. Points denote the number of samples. *In → Out ch.* denotes input and output channels, *Sample Points* denotes the number of downsampling points for each data, *Epochs* denotes total training epochs, *LR* denotes initial learning rate, *Sample Data* denotes the number of simulated data used for training, *Optimizer* denotes the optimiser employed for training, *Batch* denotes the batch size used for training, *Training Times* denotes the number of repeated training sessions for each task

Task	In → Out ch.	Sample Points	Epochs	LR	Sample Data	Optimizer	Batch	Training Times
T2Q	$13 \to 1$	8 000 (vol.)	200	1×10^{-4}	600	Adam	1	5
S2Q	$13 \to 1$	8 000 (vol.) + 8 000 (surf.)						
T2T	$11 \to 2$	8 000 (vol.)						
Q2T	$1 \to 13$	8 000 (vol.)						

G LLM USE DISCLOSURE

We used large language models (LLMs) only for paper grammar and wording edits, minor LaTeX formatting for tables and figures, lightweight source code checks and plotting assistance.LLMs were *not* used to generate scientific claims, design or run experiments, analyze results, create data or alter data, nor to draft substantive technical content. All scientific content, analyses, and conclusions were authored and verified by the authors, no confidential submission materials were provided to third-party LLM services. We take full responsibility for the submission.

H MORE EXPERIMENTAL RESULTS

Here we show the performance comparison of various neural operator models in related tasks, including Forward Modeling, Inverse Source Reconstruction, Temporal Evolution Prediction and Surface-to-Source Reconstruction tasks, involving regular/irregular grids and full-frequency/OOD environments. Multiple tables quantify the seven metrics of each model in different scenarios (The smaller the value of the indicator with a downward arrow, the better; the larger the value of the indicator with an upward arrow, the better), which cover a variety of models such as MLP, Transolver, FNO, and so on.

Table 11: Quantitative results for the temperature-to-heat (T2Q) inversion task on irregular grids under full-frequency setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.311 ± 0.004	0.578 ± 0.009	0.550 ± 0.004	0.550 ± 0.004	0.014 ± 0.002	20.658 ± 0.153	0.156 ± 0.009
DeepONet	0.281 ± 0.004	0.586 ± 0.009	0.525 ± 0.004	0.525 ± 0.004	0.010 ± 0.001	19.215 ± 0.212	0.127 ± 0.005
FNO	0.069 ± 0.001	0.759 ± 0.005	0.244 ± 0.002	0.244 ± 0.002	0.005 ± 0.001	9.518 ± 0.185	0.037 ± 0.002
GeoFNO	0.349 ± 0.012	0.559 ± 0.021	0.586 ± 0.010	0.586 ± 0.010	0.007 ± 0.001	19.161 ± 0.055	0.117 ± 0.022
FFNO	0.139 ± 0.002	0.678 ± 0.003	0.354 ± 0.002	0.354 ± 0.002	0.019 ± 0.001	14.350 ± 0.210	0.056 ± 0.003
FCNO	0.142 ± 0.001	0.655 ± 0.005	0.369 ± 0.001	0.369 ± 0.001	0.012 ± 0.001	15.668 ± 0.076	0.072 ± 0.015
LNO	0.319 ± 0.030	0.641 ± 0.008	0.556 ± 0.027	0.556 ± 0.027	0.005 ± 0.001	23.952 ± 1.956	0.061 ± 0.021
Transolver	0.091 ± 0.003	0.751 ± 0.003	0.289 ± 0.004	0.289 ± 0.004	0.003 ± 0.000	12.127 ± 0.355	0.040 ± 0.001

Table 12: Quantitative results of the temperature-to-heat (T2Q) inversion task on irregular grids under OOD setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.301 ± 0.003	0.577 ± 0.008	0.544 ± 0.003	0.544 ± 0.003	0.014 ± 0.003	19.268 ± 0.118	0.157 ± 0.012
DeepONet	0.294 ± 0.005	0.580 ± 0.010	0.535 ± 0.004	0.535 ± 0.004	0.014 ± 0.001	19.859 ± 0.217	0.129 ± 0.008
FNO	0.070 ± 0.001	0.759 ± 0.007	0.246 ± 0.002	0.246 ± 0.002	0.005 ± 0.000	9.036 ± 0.200	0.041 ± 0.002
GeoFNO	0.343 ± 0.017	0.575 ± 0.018	0.580 ± 0.015	0.580 ± 0.015	0.008 ± 0.003	19.019 ± 0.098	0.116 ± 0.018
FFNO	0.120 ± 0.003	0.683 ± 0.002	0.339 ± 0.004	0.339 ± 0.004	0.018 ± 0.001	13.563 ± 0.513	0.056 ± 0.014
FCNO	0.145 ± 0.001	0.648 ± 0.005	0.373 ± 0.001	0.373 ± 0.001	0.014 ± 0.001	15.112 ± 0.056	0.068 ± 0.008
LNO	0.322 ± 0.019	0.632 ± 0.006	0.559 ± 0.017	0.559 ± 0.017	0.006 ± 0.001	24.755 ± 0.815	0.060 ± 0.007
Transolver	0.088 ± 0.003	0.751 ± 0.001	0.283 ± 0.002	0.283 ± 0.002	0.003 ± 0.000	11.235 ± 0.288	0.040 ± 0.002

Table 13: Quantitative results of the temperature-to-heat (T2Q) inversion task on regular grids under full-frequency setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.127 ± 0.001	0.863 ± 0.011	0.337 ± 0.002	0.337 ± 0.002	0.004 ± 0.001	15.904 ± 0.127	0.022 ± 0.010
DeepONet	0.110 ± 0.002	0.875 ± 0.013	0.311 ± 0.004	0.311 ± 0.004	0.005 ± 0.001	14.769 ± 0.092	0.010 ± 0.002
FNO	0.040 ± 0.001	0.917 ± 0.009	0.186 ± 0.001	0.186 ± 0.001	0.003 ± 0.000	7.836 ± 0.067	0.018 ± 0.003
GeoFNO	0.222 ± 0.017	0.821 ± 0.013	0.446 ± 0.018	0.446 ± 0.018	0.007 ± 0.003	23.442 ± 1.164	0.012 ± 0.003
FFNO	0.052 ± 0.001	0.894 ± 0.002	0.219 ± 0.001	0.219 ± 0.001	0.010 ± 0.000	9.523 ± 0.113	0.031 ± 0.008
FCNO	0.072 ± 0.004	0.898 ± 0.005	0.253 ± 0.005	0.253 ± 0.005	0.009 ± 0.000	12.022 ± 0.549	0.042 ± 0.014
LNO	0.164 ± 0.105	0.864 ± 0.083	0.374 ± 0.121	0.374 ± 0.121	0.003 ± 0.001	16.451 ± 5.906	0.014 ± 0.006
Transolver	0.053 ± 0.003	0.929 ± 0.002	0.219 ± 0.002	0.219 ± 0.002	0.002 ± 0.000	9.473 ± 0.248	0.008 ± 0.001

Table 14: Quantitative results of the temperature-to-heat (T2Q) inversion task on regular grids under OOD setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.126 ± 0.001	0.861 ± 0.011	0.341 ± 0.002	0.341 ± 0.002	0.005 ± 0.001	16.300 ± 0.068	0.024 ± 0.012
DeepONet	0.120 ± 0.003	0.875 ± 0.013	0.327 ± 0.004	0.327 ± 0.004	0.006 ± 0.000	15.936 ± 0.186	0.011 ± 0.001
FNO	0.044 ± 0.001	0.919 ± 0.008	0.187 ± 0.001	0.187 ± 0.001	0.003 ± 0.000	8.167 ± 0.094	0.018 ± 0.004
GeoFNO	0.168 ± 0.012	0.810 ± 0.037	0.391 ± 0.016	0.391 ± 0.016	0.007 ± 0.005	18.423 ± 0.833	0.012 ± 0.007
FFNO	0.058 ± 0.001	0.899 ± 0.001	0.225 ± 0.003	0.225 ± 0.003	0.009 ± 0.000	10.014 ± 0.227	0.032 ± 0.006
FCNO	0.072 ± 0.002	0.898 ± 0.004	0.249 ± 0.004	0.249 ± 0.004	0.009 ± 0.000	11.540 ± 0.332	0.030 ± 0.006
LNO	0.133 ± 0.003	0.902 ± 0.006	0.333 ± 0.002	0.333 ± 0.002	0.003 ± 0.000	15.662 ± 0.277	0.015 ± 0.006
Transolver	0.062 ± 0.002	0.927 ± 0.002	0.237 ± 0.002	0.237 ± 0.002	0.002 ± 0.000	10.781 ± 0.363	0.008 ± 0.002

Table 15: Quantitative results for the surface temperature to heat (S2Q) inversion task on irregular grids under full-frequency setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.667 ± 0.017	0.366 ± 0.008	0.813 ± 0.011	0.813 ± 0.011	0.004 ± 0.002	34.211 ± 0.159	0.237 ± 0.070
DeepONet	0.653 ± 0.023	0.183 ± 0.006	0.805 ± 0.014	0.805 ± 0.014	0.001 ± 0.001	32.556 ± 0.519	0.189 ± 0.044
FNO	0.443 ± 0.004	0.630 ± 0.009	0.656 ± 0.003	0.656 ± 0.003	0.008 ± 0.001	28.445 ± 0.201	0.053 ± 0.002
GeoFNO	1.022 ± 0.017	0.017 ± 0.002	1.011 ± 0.008	1.011 ± 0.008	0.010 ± 0.004	33.338 ± 0.012	0.184 ± 0.019
FFNO	0.547 ± 0.011	0.588 ± 0.004	0.731 ± 0.007	0.731 ± 0.007	0.004 ± 0.000	31.579 ± 0.444	0.061 ± 0.004
FCNO	0.464 ± 0.004	0.517 ± 0.015	0.673 ± 0.003	0.673 ± 0.003	0.003 ± 0.000	29.157 ± 0.303	0.062 ± 0.005
LNO	0.442 ± 0.066	0.604 ± 0.004	0.654 ± 0.048	0.654 ± 0.048	0.010 ± 0.003	28.170 ± 3.192	0.122 ± 0.027
Transolver	0.365 ± 0.008	0.687 ± 0.003	0.584 ± 0.006	0.584 ± 0.006	0.003 ± 0.001	26.922 ± 0.241	0.039 ± 0.002

Table 16: Quantitative results for the surface temperature to heat (S2Q) inversion task on irregular grids under OOD setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.695 ± 0.019	0.365 ± 0.005	0.830 ± 0.012	0.830 ± 0.012	0.004 ± 0.002	37.203 ± 0.201	0.209 ± 0.082
DeepONet	0.663 ± 0.021	0.180 ± 0.009	0.810 ± 0.013	0.810 ± 0.013	0.001 ± 0.001	34.130 ± 0.483	0.190 ± 0.051
FNO	0.437 ± 0.005	0.629 ± 0.009	0.654 ± 0.004	0.654 ± 0.004	0.008 ± 0.001	28.072 ± 0.205	0.058 ± 0.005
GeoFNO	1.029 ± 0.018	0.025 ± 0.004	1.011 ± 0.004	1.011 ± 0.004	0.008 ± 0.003	35.187 ± 0.059	0.181 ± 0.016
FFNO	0.516 ± 0.002	0.598 ± 0.003	0.710 ± 0.002	0.710 ± 0.002	0.004 ± 0.000	29.409 ± 0.366	0.062 ± 0.007
FCNO	0.452 ± 0.008	0.518 ± 0.009	0.665 ± 0.006	0.665 ± 0.006	0.002 ± 0.000	28.992 ± 0.635	0.071 ± 0.012
LNO	0.485 ± 0.113	0.588 ± 0.091	0.682 ± 0.074	0.682 ± 0.074	0.010 ± 0.003	31.849 ± 3.624	0.121 ± 0.064
Transolver	0.344 ± 0.006	0.687 ± 0.004	0.570 ± 0.004	0.570 ± 0.004	0.003 ± 0.001	26.512 ± 0.206	0.043 ± 0.001

Table 17: Quantitative results for the surface temperature to heat (S2Q) inversion task on regular grids under full-frequency setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.473 ± 0.066	0.603 ± 0.067	0.680 ± 0.049	0.680 ± 0.049	0.003 ± 0.001	34.790 ± 3.014	0.061 ± 0.027
DeepONet	0.672 ± 0.035	0.134 ± 0.047	0.816 ± 0.023	0.816 ± 0.023	0.002 ± 0.001	39.705 ± 1.240	0.509 ± 0.366
FNO	0.330 ± 0.002	0.903 ± 0.002	0.537 ± 0.002	0.537 ± 0.002	0.004 ± 0.000	29.897 ± 0.145	0.013 ± 0.001
GeoFNO	1.004 ± 0.003	0.071 ± 0.011	1.002 ± 0.001	1.002 ± 0.001	0.006 ± 0.002	42.126 ± 0.006	0.085 ± 0.005
FFNO	0.421 ± 0.009	0.849 ± 0.003	0.620 ± 0.006	0.620 ± 0.006	0.003 ± 0.000	32.368 ± 0.776	0.018 ± 0.003
FCNO	0.356 ± 0.017	0.850 ± 0.003	0.563 ± 0.011	0.563 ± 0.011	0.003 ± 0.000	31.456 ± 1.195	0.025 ± 0.007
LNO	0.388 ± 0.069	0.863 ± 0.025	0.598 ± 0.070	0.598 ± 0.070	0.004 ± 0.001	33.665 ± 5.873	0.016 ± 0.002
Transolver	0.247 ± 0.013	0.902 ± 0.004	0.452 ± 0.007	0.452 ± 0.007	0.004 ± 0.001	24.779 ± 0.559	0.015 ± 0.002

Table 18: Quantitative results for the surface temperature to heat (S2Q) inversion task on regular grids under OOD setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.473 ± 0.066	0.603 ± 0.067	0.680 ± 0.049	0.680 ± 0.049	0.003 ± 0.001	34.790 ± 3.014	0.061 ± 0.027
DeepONet	0.672 ± 0.035	0.134 ± 0.047	0.816 ± 0.023	0.816 ± 0.023	0.002 ± 0.001	39.705 ± 1.240	0.509 ± 0.366
FNO	0.330 ± 0.002	0.903 ± 0.002	0.537 ± 0.002	0.537 ± 0.002	0.004 ± 0.000	29.897 ± 0.145	0.013 ± 0.001
GeoFNO	1.004 ± 0.003	0.071 ± 0.011	1.002 ± 0.001	1.002 ± 0.001	0.006 ± 0.002	42.126 ± 0.006	0.085 ± 0.005
FFNO	0.421 ± 0.009	0.849 ± 0.003	0.620 ± 0.006	0.620 ± 0.006	0.003 ± 0.000	32.368 ± 0.776	0.018 ± 0.003
FCNO	0.356 ± 0.017	0.850 ± 0.003	0.563 ± 0.011	0.563 ± 0.011	0.003 ± 0.000	31.456 ± 1.195	0.025 ± 0.007
LNO	0.388 ± 0.069	0.863 ± 0.025	0.598 ± 0.070	0.598 ± 0.070	0.004 ± 0.001	33.665 ± 5.873	0.016 ± 0.002
Transolver	0.247 ± 0.013	0.902 ± 0.004	0.452 ± 0.007	0.452 ± 0.007	0.004 ± 0.001	24.779 ± 0.559	0.015 ± 0.002

Table 19: Quantitative results of the heat-to-temperature (Q2T) forward task on irregular grids under full-frequency setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.979 ± 0.001	0.011 ± 0.001	0.989 ± 0.001	0.989 ± 0.001	0.001 ± 0.000	16.445 ± 0.011	0.421 ± 0.017
FNO	0.019 ± 0.001	0.939 ± 0.001	0.108 ± 0.009	0.110 ± 0.013	0.005 ± 0.003	2.276 ± 0.468	0.055 ± 0.010
GeoFNO	0.450 ± 0.002	0.573 ± 0.003	0.667 ± 0.002	0.667 ± 0.002	0.008 ± 0.002	13.095 ± 0.336	0.325 ± 0.006
FFNO	0.046 ± 0.001	0.854 ± 0.001	0.193 ± 0.001	0.193 ± 0.001	0.016 ± 0.001	4.268 ± 0.057	0.106 ± 0.006
FCNO	0.065 ± 0.002	0.783 ± 0.004	0.235 ± 0.005	0.235 ± 0.005	0.011 ± 0.000	5.176 ± 0.074	0.139 ± 0.015
LNO	0.037 ± 0.002	0.917 ± 0.002	0.158 ± 0.004	0.158 ± 0.004	0.006 ± 0.001	3.882 ± 0.265	0.060 ± 0.004
Transolver	0.017 ± 0.002	0.952 ± 0.004	0.108 ± 0.004	0.109 ± 0.001	0.003 ± 0.002	2.183 ± 0.318	0.054 ± 0.005

Table 20: Quantitative results of the heat-to-temperature (Q2T) forward task on irregular grids under OOD setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.979 ± 0.001	0.011 ± 0.001	0.990 ± 0.000	0.990 ± 0.000	0.001 ± 0.000	16.789 ± 0.012	0.401 ± 0.019
FNO	0.018 ± 0.001	0.938 ± 0.001	0.103 ± 0.002	0.103 ± 0.002	0.004 ± 0.000	1.913 ± 0.070	0.058 ± 0.002
GeoFNO	0.463 ± 0.003	0.577 ± 0.011	0.667 ± 0.021	0.663 ± 0.030	0.014 ± 0.012	13.857 ± 0.783	0.309 ± 0.004
FFNO	0.048 ± 0.001	0.853 ± 0.001	0.194 ± 0.002	0.194 ± 0.002	0.015 ± 0.001	4.578 ± 0.133	0.100 ± 0.007
FCNO	0.064 ± 0.002	0.785 ± 0.005	0.232 ± 0.004	0.232 ± 0.004	0.010 ± 0.000	5.098 ± 0.197	0.134 ± 0.010
LNO	0.038 ± 0.003	0.918 ± 0.002	0.161 ± 0.008	0.161 ± 0.008	0.006 ± 0.000	4.090 ± 0.376	0.064 ± 0.003
Transolver	0.018 ± 0.001	0.950 ± 0.001	0.110 ± 0.003	0.110 ± 0.003	0.002 ± 0.000	2.049 ± 0.291	0.057 ± 0.003

Table 21: Quantitative results of the heat-to-temperature (Q2T) forward task on regular grids under full-frequency setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.751 ± 0.008	0.250 ± 0.013	0.865 ± 0.005	0.865 ± 0.005	0.001 ± 0.000	23.476 ± 0.794	0.162 ± 0.068
FNO	0.018 ± 0.000	0.964 ± 0.002	0.103 ± 0.001	0.103 ± 0.001	0.004 ± 0.000	3.786 ± 0.100	0.017 ± 0.001
GeoFNO	0.419 ± 0.001	0.831 ± 0.003	0.631 ± 0.001	0.632 ± 0.001	0.003 ± 0.000	25.503 ± 0.007	0.167 ± 0.003
FFNO	0.035 ± 0.002	0.948 ± 0.002	0.157 ± 0.004	0.157 ± 0.004	0.012 ± 0.002	6.295 ± 0.224	0.023 ± 0.003
FCNO	0.036 ± 0.000	0.933 ± 0.001	0.165 ± 0.001	0.165 ± 0.001	0.015 ± 0.001	5.818 ± 0.057	0.033 ± 0.003
LNO	0.038 ± 0.003	0.962 ± 0.005	0.164 ± 0.009	0.164 ± 0.009	0.003 ± 0.001	6.544 ± 0.415	0.013 ± 0.002
Transolver	0.018 ± 0.001	0.977 ± 0.001	0.111 ± 0.002	0.111 ± 0.002	0.002 ± 0.000	3.971 ± 0.157	0.012 ± 0.002

Table 22: Quantitative results of the heat-to-temperature (Q2T) forward task on regular grids under OOD setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ↓	SSIM ↑	RMSE ↓	nRMSE ↓	cRMSE ↓	Max ↓	bRMSE ↓
MLP	0.755 ± 0.011	0.245 ± 0.012	0.867 ± 0.007	0.867 ± 0.007	0.001 ± 0.000	23.717 ± 1.066	0.179 ± 0.054
FNO	0.019 ± 0.000	0.965 ± 0.003	0.107 ± 0.002	0.107 ± 0.002	0.003 ± 0.000	4.023 ± 0.124	0.017 ± 0.002
GeoFNO	0.440 ± 0.000	0.835 ± 0.002	0.648 ± 0.000	0.648 ± 0.000	0.003 ± 0.001	27.766 ± 0.023	0.149 ± 0.003
FFNO	0.030 ± 0.001	0.945 ± 0.002	0.149 ± 0.003	0.149 ± 0.003	0.013 ± 0.002	5.616 ± 0.116	0.026 ± 0.004
FCNO	0.044 ± 0.001	0.928 ± 0.001	0.179 ± 0.001	0.179 ± 0.001	0.015 ± 0.001	6.607 ± 0.097	0.039 ± 0.005
LNO	0.040 ± 0.003	0.957 ± 0.006	0.173 ± 0.007	0.173 ± 0.007	0.004 ± 0.000	6.728 ± 0.296	0.015 ± 0.003
Transolver	0.021 ± 0.001	0.978 ± 0.001	0.117 ± 0.002	0.117 ± 0.002	0.002 ± 0.000	4.580 ± 0.128	0.012 ± 0.001

Table 23: Quantitative results of the temperature-to-temperature (T2T) prediction task on irregular grids under full-frequency setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ($\times 10^{-3}$) ↓	SSIM ↑	RMSE ($\times 10^{-2}$) ↓	nRMSE ($\times 10^{-2}$) ↓	cRMSE ($\times 10^{-3}$) ↓	Max ↓	bRMSE ($\times 10^{-3}$) ↓
MLP	954.897 ± 2.459	0.010356 ± 0.000623	97.718 ± 0.126	97.718 ± 0.126	2.340 ± 0.946	6.464 ± 0.623	41.718 ± 9.210
FNO	0.004 ± 0.001	0.999915 ± 0.000018	0.188 ± 0.022	0.188 ± 0.022	0.070 ± 0.022	0.235 ± 0.064	0.706 ± 0.208
GeoFNO	0.223 ± 0.223	0.996296 ± 0.003234	1.300 ± 0.775	1.300 ± 0.775	0.761 ± 0.705	1.290 ± 0.116	1.688 ± 1.573
FFNO	0.030 ± 0.003	0.999097 ± 0.000090	0.544 ± 0.027	0.544 ± 0.027	0.700 ± 0.094	1.013 ± 0.051	1.595 ± 0.212
FCNO	0.630 ± 0.020	0.984715 ± 0.000547	2.476 ± 0.194	2.476 ± 0.194	5.106 ± 0.162	1.857 ± 0.160	5.345 ± 0.445
LNO	5.735 ± 0.414	0.947556 ± 0.001475	7.532 ± 0.396	7.532 ± 0.396	23.904 ± 3.602	4.022 ± 0.615	8.755 ± 1.087
Transolver	0.002 ± 0.0001	0.999996 ± 0.000004	0.116 ± 0.012	0.116 ± 0.012	0.020 ± 0.007	0.185 ± 0.044	0.296 ± 0.059

Table 24: Quantitative results of the temperature-to-temperature (T2T) prediction task on irregular grids under OOD setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ($\times 10^{-3}$) ↓	SSIM ↑	RMSE ($\times 10^{-2}$) ↓	nRMSE ($\times 10^{-2}$) ↓	cRMSE ($\times 10^{-3}$) ↓	Max ↓	bRMSE ($\times 10^{-3}$) ↓
MLP	951.890 ± 2.649	0.010708 ± 0.000649	97.560 ± 0.142	97.560 ± 0.142	3.009 ± 1.940	6.236 ± 0.821	45.946 ± 5.980
FNO	0.005 ± 0.001	0.999879 ± 0.000021	0.212 ± 0.025	0.212 ± 0.025	0.068 ± 0.022	0.241 ± 0.055	0.737 ± 0.199
GeoFNO	0.227 ± 0.221	0.996368 ± 0.003169	1.276 ± 0.123	1.276 ± 0.123	2.576 ± 2.553	1.266 ± 0.115	1.720 ± 0.982
FFNO	0.031 ± 0.003	0.999113 ± 0.000092	0.542 ± 0.016	0.542 ± 0.016	0.698 ± 0.175	1.008 ± 0.071	1.642 ± 0.206
FCNO	0.638 ± 0.021	0.984639 ± 0.000550	2.494 ± 0.196	2.494 ± 0.196	5.194 ± 0.168	1.879 ± 0.112	5.458 ± 0.545
LNO	5.786 ± 0.419	0.955559 ± 0.001581	7.544 ± 0.358	7.544 ± 0.358	23.763 ± 3.616	4.117 ± 0.576	9.661 ± 0.857
Transolver	0.002 ± 0.000	0.999994 ± 0.000005	0.119 ± 0.013	0.119 ± 0.013	0.019 ± 0.007	0.196 ± 0.051	0.308 ± 0.043

Table 25: Quantitative results of the temperature-to-temperature (T2T) prediction task on regular grids under full-frequency setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ($\times 10^{-3}$) ↓	SSIM ↑	RMSE ($\times 10^{-2}$) ↓	nRMSE ($\times 10^{-2}$) ↓	cRMSE ($\times 10^{-3}$) ↓	Max ↓	bRMSE ($\times 10^{-3}$) ↓
MLP	955.422 ± 3.412	0.010155 ± 0.000517	97.760 ± 0.170	97.760 ± 0.170	2.540 ± 1.643	6.329 ± 0.634	42.009 ± 7.311
FNO	0.005 ± 0.001	0.999886 ± 0.000019	0.199 ± 0.023	0.199 ± 0.023	0.070 ± 0.027	0.243 ± 0.062	0.685 ± 0.248
GeoFNO	0.207 ± 0.190	0.996455 ± 0.003413	1.223 ± 0.149	1.223 ± 0.149	2.306 ± 2.158	1.273 ± 0.115	1.669 ± 0.955
FFNO	0.031 ± 0.003	0.999106 ± 0.000085	0.547 ± 0.019	0.547 ± 0.019	0.706 ± 0.183	1.017 ± 0.061	1.607 ± 0.255
FCNO	0.662 ± 0.023	0.984206 ± 0.000535	2.550 ± 0.200	2.550 ± 0.200	5.116 ± 0.161	1.869 ± 0.146	5.360 ± 0.407
LNO	5.607 ± 0.449	0.955190 ± 0.001672	7.489 ± 0.433	7.489 ± 0.433	23.578 ± 3.794	4.116 ± 0.638	9.949 ± 1.256
Transolver	0.002 ± 0.000	0.999995 ± 0.000004	0.117 ± 0.013	0.117 ± 0.013	0.019 ± 0.007	0.197 ± 0.050	0.287 ± 0.055

Table 26: Quantitative results of the temperature-to-temperature (T2T) prediction task on regular grids under OOD setting, trained and evaluated on the *Type I double-layer* subset of the Aletheia dataset. Each entry reports mean ± standard deviation over repeated runs.

Model	MSE ($\times 10^{-3}$) ↓	SSIM ↑	RMSE ($\times 10^{-2}$) ↓	nRMSE ($\times 10^{-2}$) ↓	cRMSE ($\times 10^{-3}$) ↓	Max ↓	bRMSE ($\times 10^{-3}$) ↓
MLP	955.924 ± 3.168	0.010191 ± 0.000537	97.788 ± 0.163	97.788 ± 0.163	2.573 ± 1.603	6.335 ± 0.650	42.184 ± 6.998
FNO	0.004 ± 0.001	0.999888 ± 0.000018	0.200 ± 0.023	0.200 ± 0.023	0.068 ± 0.018	0.241 ± 0.060	0.702 ± 0.215
GeoFNO	0.204 ± 0.193	0.996451 ± 0.003315	1.215 ± 0.148	1.215 ± 0.148	2.313 ± 2.150	1.273 ± 0.117	1.628 ± 0.848
FFNO	0.030 ± 0.003	0.999097 ± 0.000083	0.546 ± 0.020	0.546 ± 0.020	0.703 ± 0.195	1.017 ± 0.053	1.642 ± 0.212
FCNO	0.661 ± 0.021	0.984198 ± 0.000534	2.547 ± 0.199	2.547 ± 0.199	5.119 ± 0.161	1.871 ± 0.142	5.341 ± 0.374
LNO	5.576 ± 0.416	0.955002 ± 0.001634	7.488 ± 0.421	7.488 ± 0.421	23.548 ± 3.665	4.120 ± 0.646	9.966 ± 1.093
Transolver	0.002 ± 0.000	0.999995 ± 0.000004	0.117 ± 0.013	0.117 ± 0.013	0.019 ± 0.007	0.199 ± 0.048	0.280 ± 0.125

I MORE VISUALIZATIONS

The following presents visualisations of all experimental results across all tasks, including Forward Modeling, Inverse Source Reconstruction, Temporal Evolution Prediction and Surface-to-Source Reconstruction, involving regular/irregular grids and full-frequency/OOD environments. We have selected experimental results from FNO, Transolver, GeoFNO, and LNO models for display, with visualisations provided for inputs and outputs under each configuration.

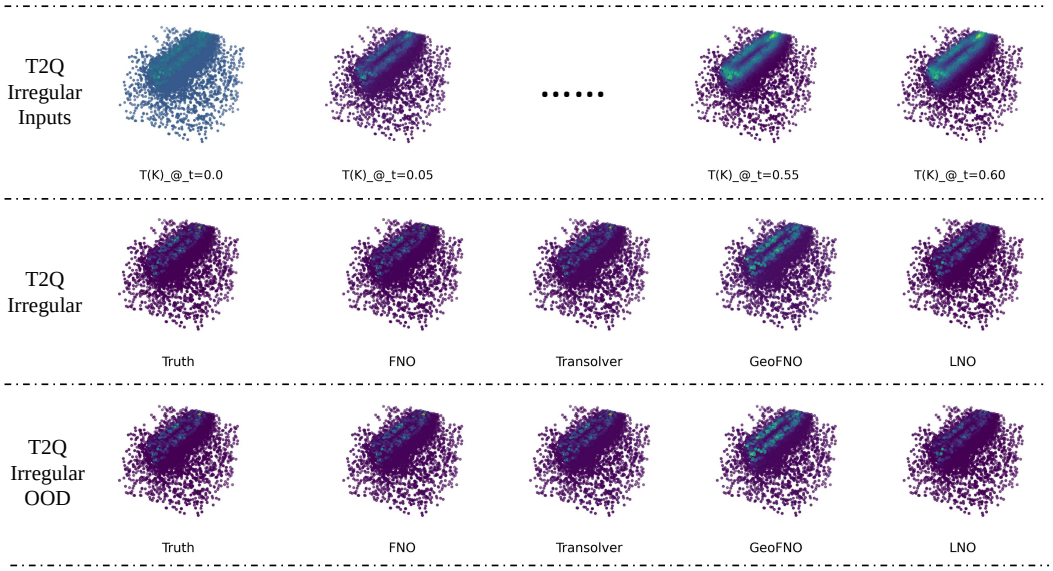


Figure 8: Visualization of experimental results under the temperature to heat (T2Q) inverse task on irregular grids, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

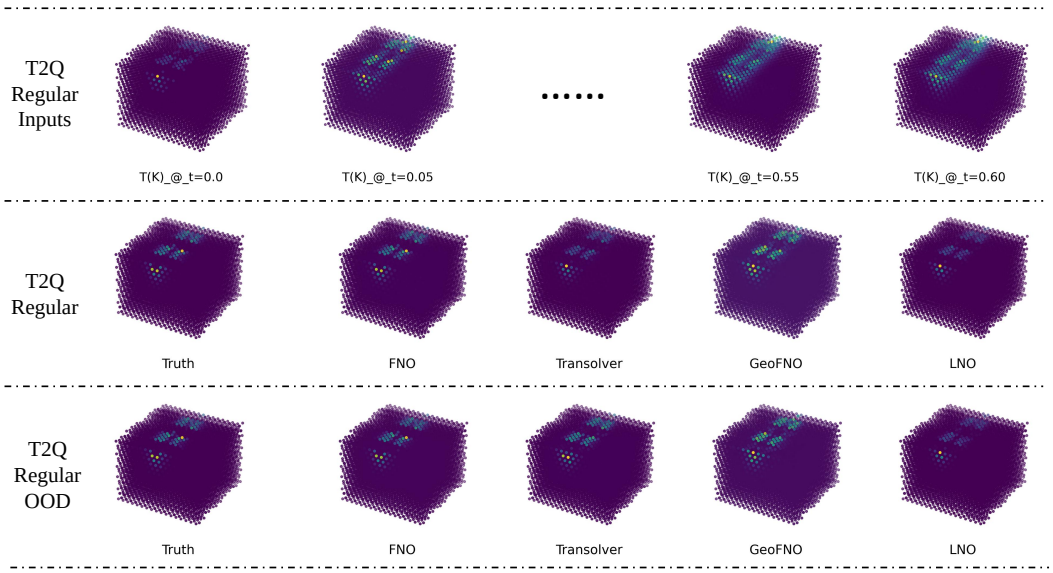


Figure 9: Visualization of experimental results under the temperature to heat (T2Q) inverse task on regular grids, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

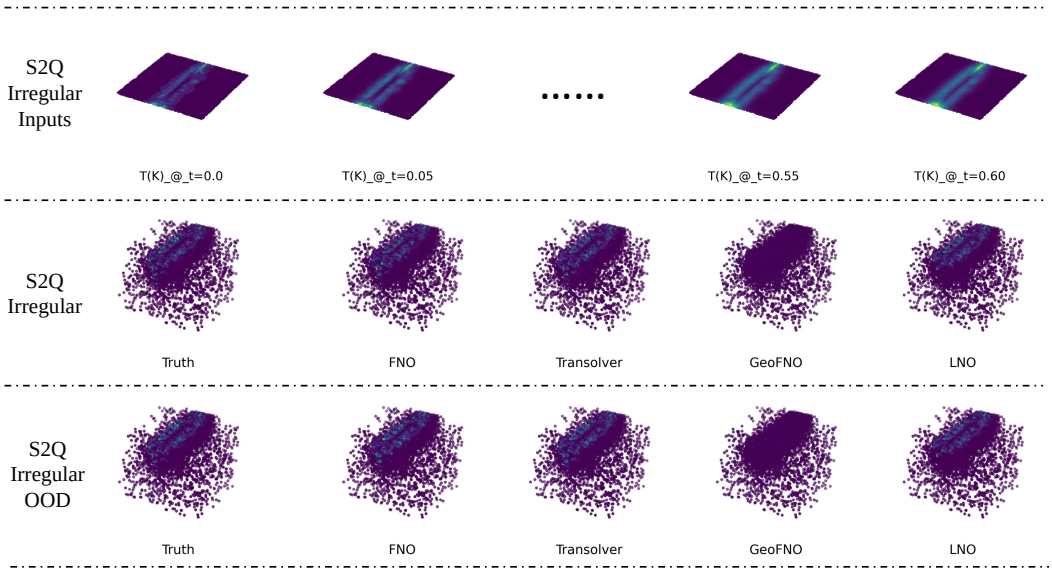


Figure 10: Visualization of experimental results under the surface temperature to heat (S2Q) inverse task on irregular grids, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

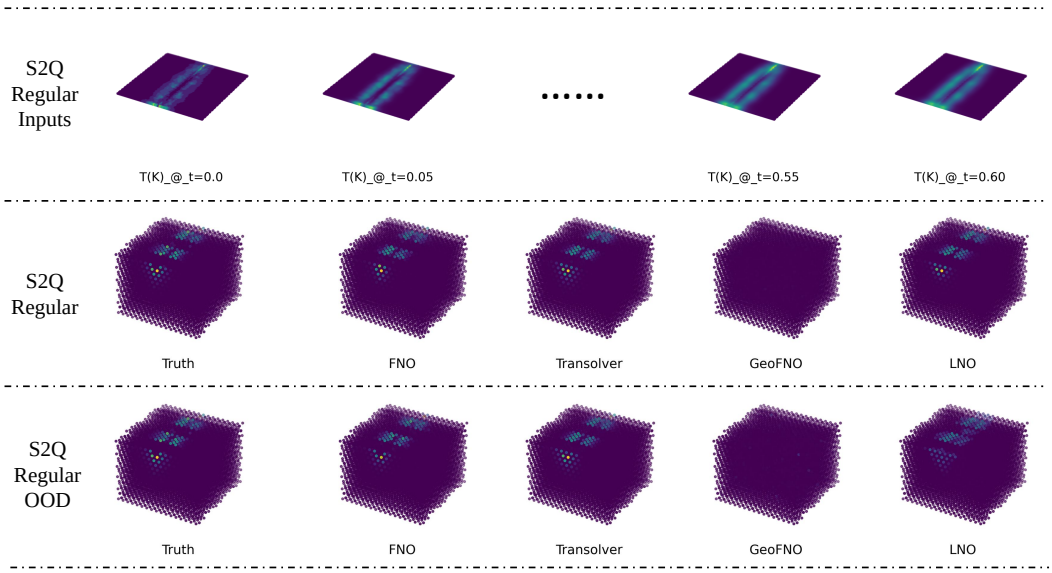


Figure 11: Visualization of experimental results under the surface temperature to heat (S2Q) inverse task on regular grids, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

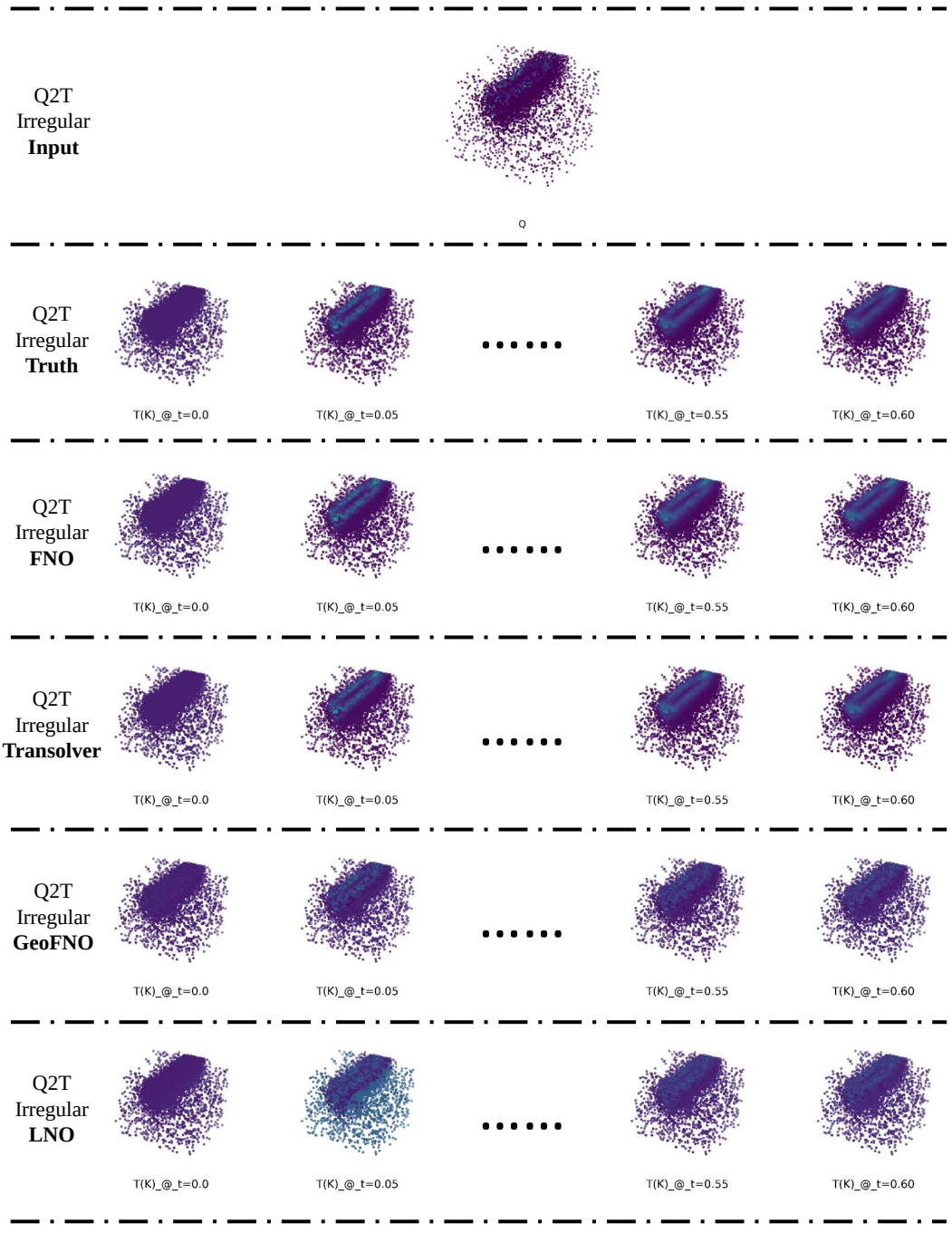


Figure 12: Visualization of experimental results under the heat-to-temperature (Q2T) forward task on irregular grids under full-frequency setting, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

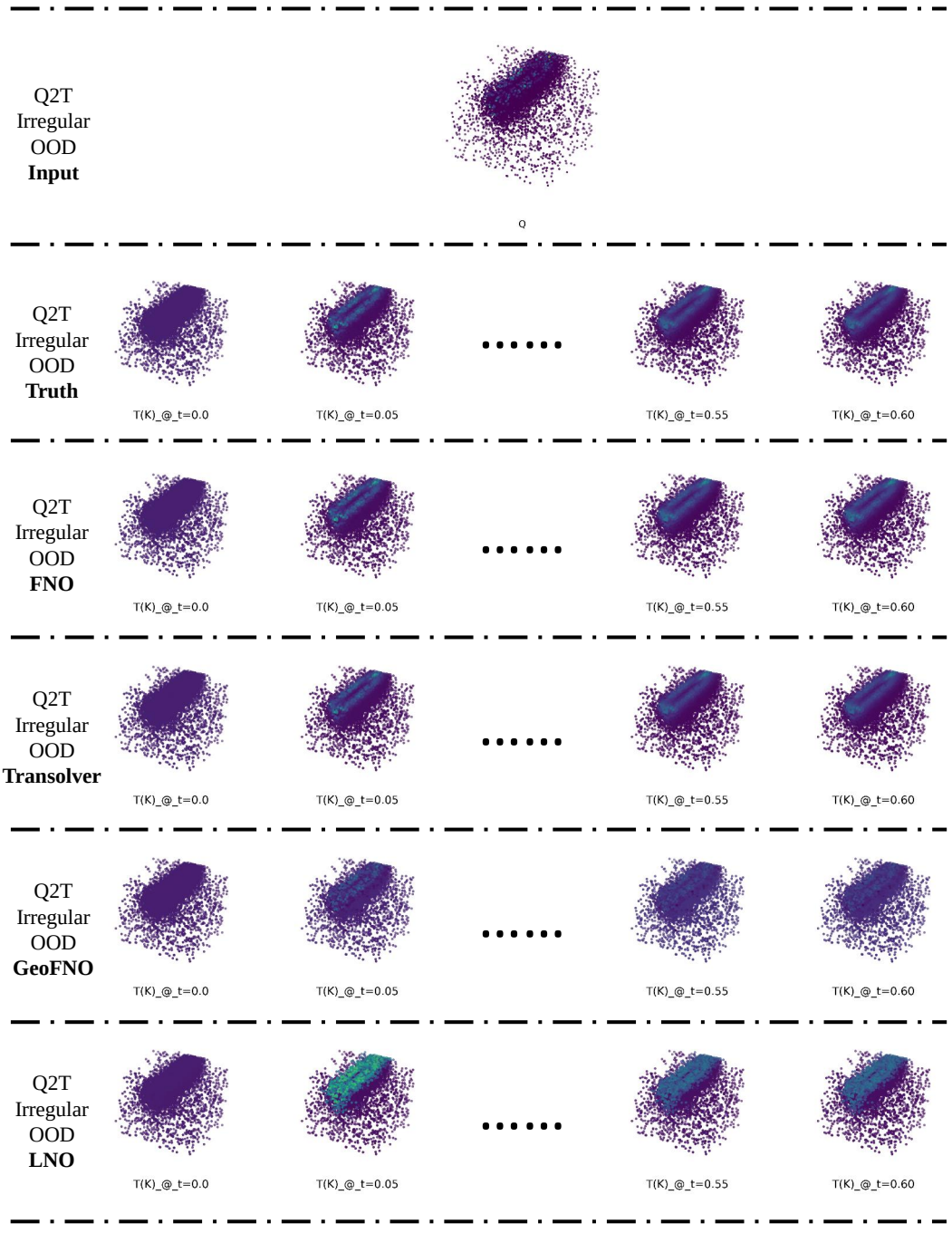


Figure 13: Visualization of experimental results under the heat-to-temperature (Q2T) forward task on irregular grids under OOD setting, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

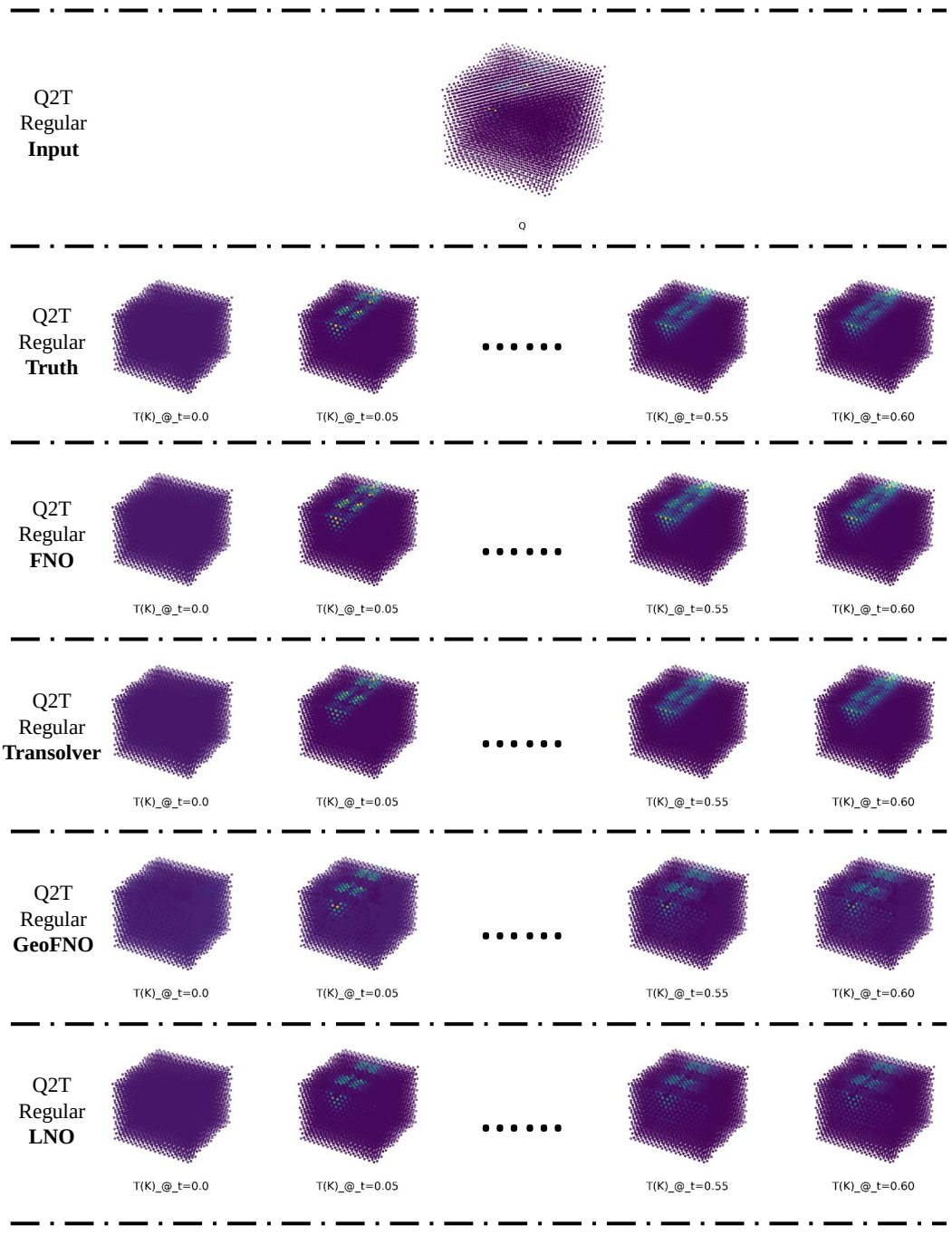


Figure 14: Visualization of experimental results under the heat-to-temperature (Q2T) forward task on regular grids under full-frequency setting, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

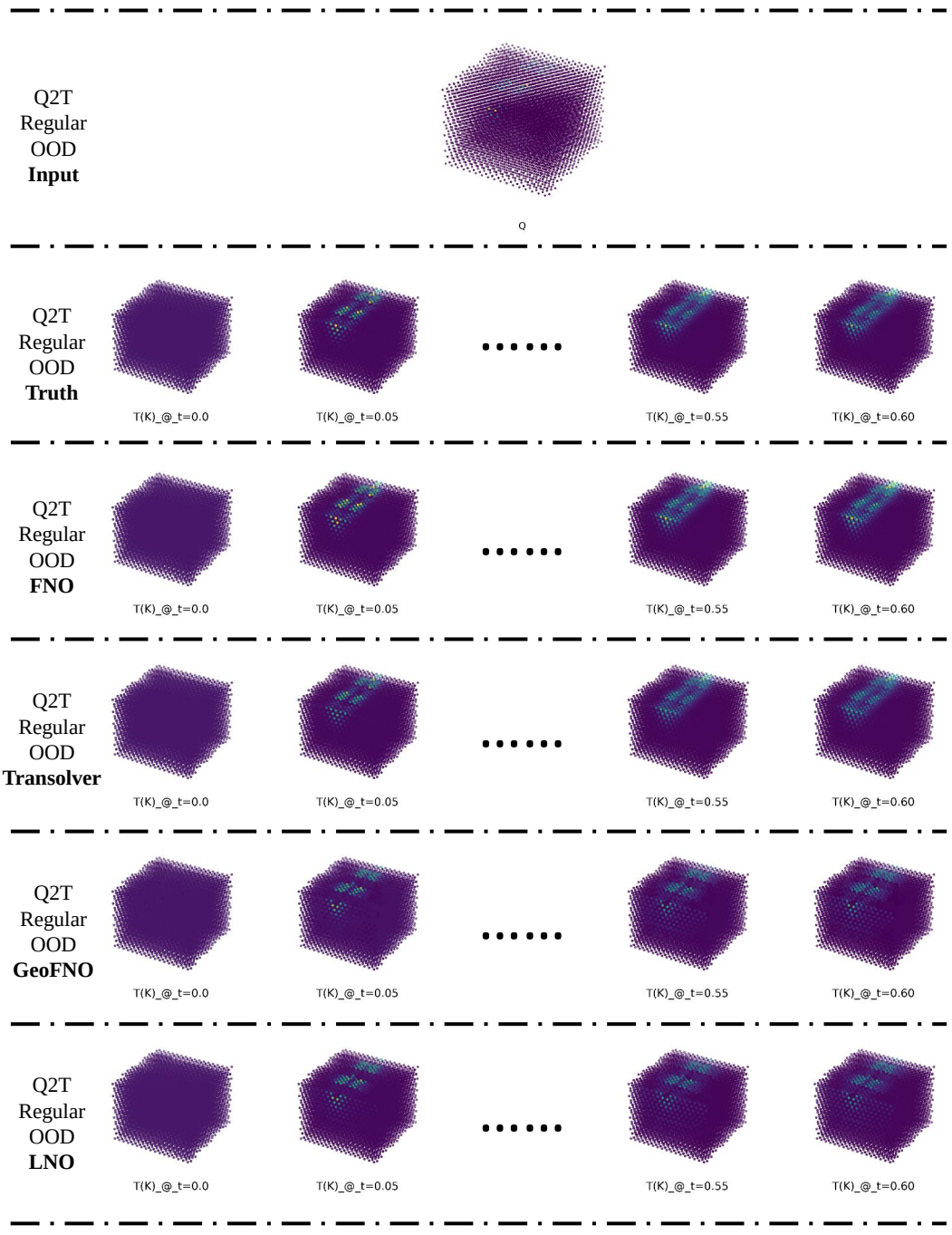


Figure 15: Visualization of experimental results under the heat-to-temperature (Q2T) forward task on regular grids under OOD setting, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

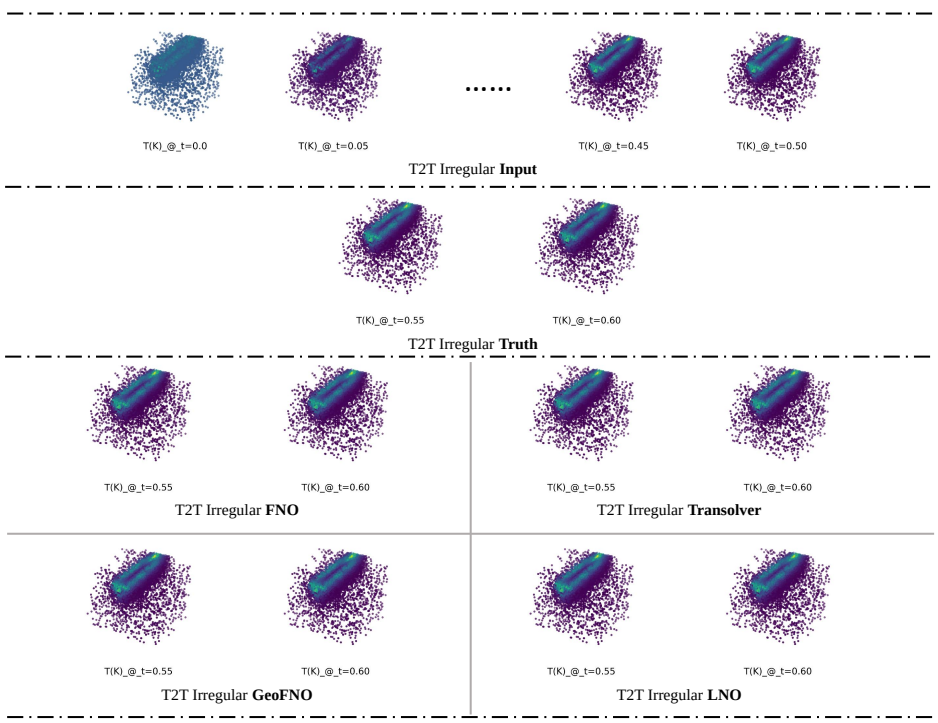


Figure 16: Visualization of experimental results under the temperature-to-temperature (T2T) prediction task on irregular grids under full-frequency setting, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

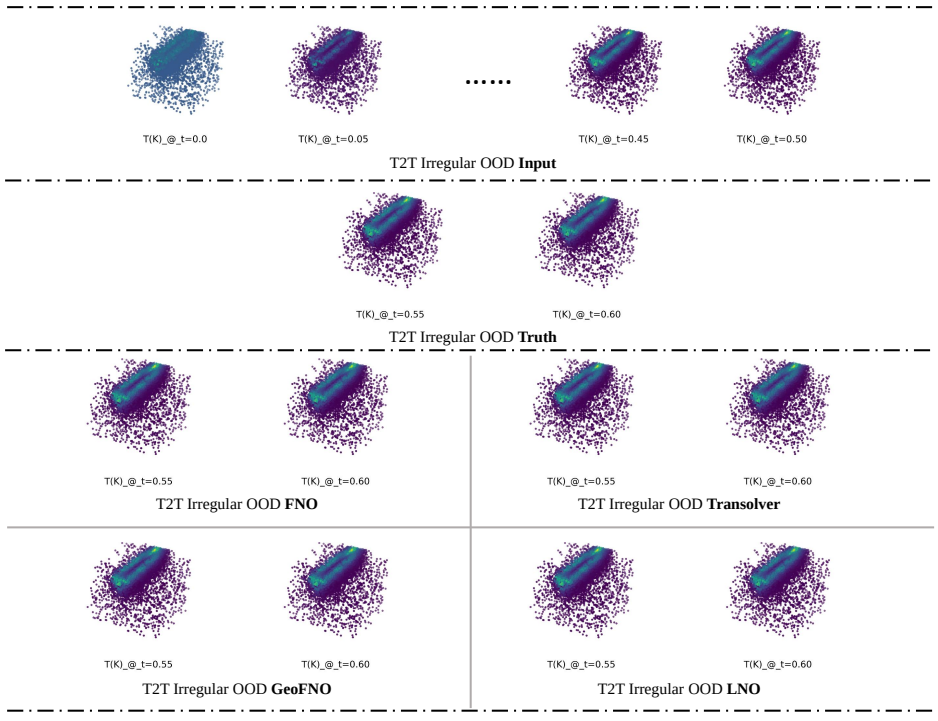


Figure 17: Visualization of experimental results under the temperature-to-temperature (T2T) prediction task on irregular grids under OOD setting, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

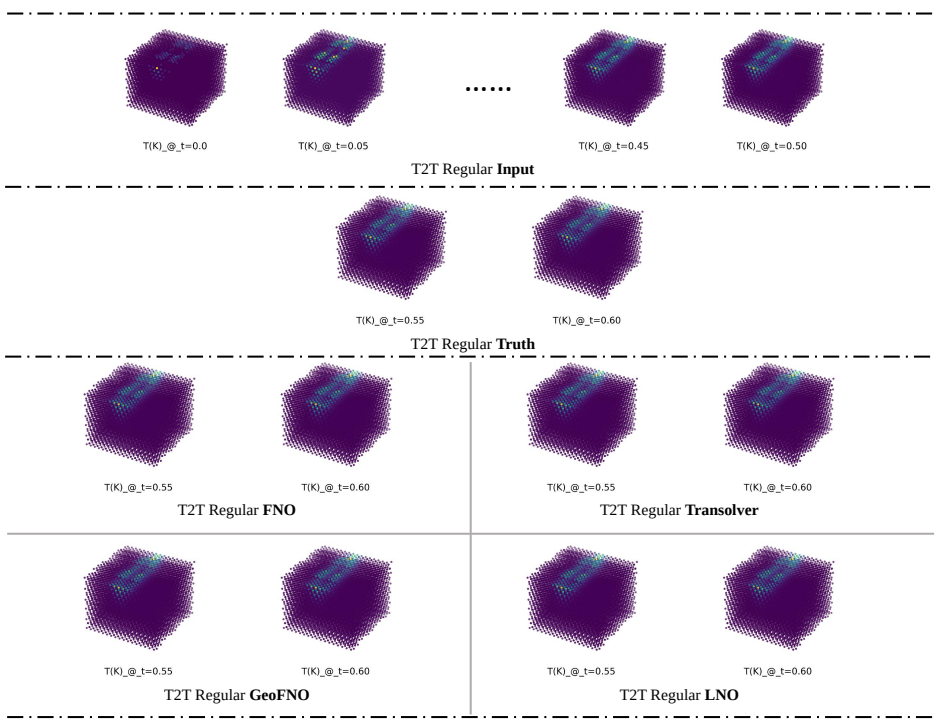


Figure 18: Visualization of experimental results under the temperature-to-temperature (T2T) prediction task on regular grids under full-frequency setting, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.

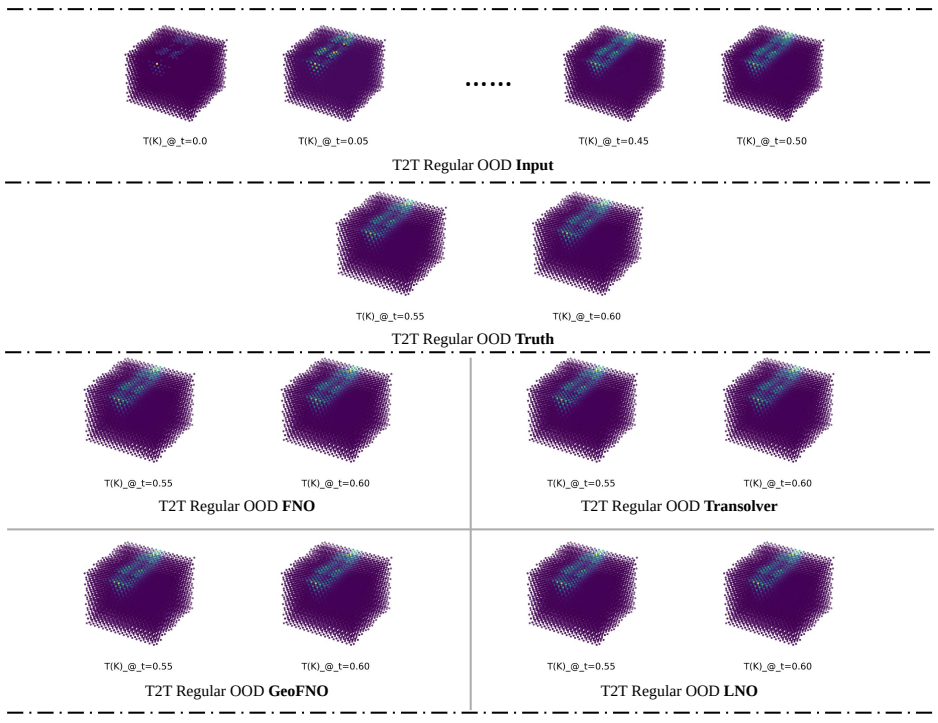


Figure 19: Visualization of experimental results under the temperature-to-temperature (T2T) prediction task on regular grids under OOD setting, the experiment was trained and evaluated on the Type I double-layer subset of the Aletheia dataset.