WaveBench: Benchmark Datasets for Modeling Wave Propagation PDEs Supplementary Material

A Details of the time-harmonic datasets

	Wavespeed c	Frequency $\omega/2\pi$
Acoustic wave	Isotropic GRF	10 Hz
		15 Hz
		20 Hz
		40 Hz
	Anisotropic GRF	10 Hz
		15 Hz
		20 Hz
		40 Hz
Elastic wave	Anisotropic GRF	10 Hz
		15 Hz
		20 Hz
		40 Hz

Table 2: **Summary of 12 time-harmonic datasets.** Each dataset corresponds to a governing time-harmonic wave equation (acoustic or elastic), a type of wavespeed (isotropic GRF or anisotropic GRF) and frequency (10, 15, 20, or 40 Hz). **Each dataset consists of 49,000 training samples, 500 validation samples, and 500 test samples.** The roles of the wavespeed c and frequency ω in time-harmonic wave propagation can be seen in the Helmholtz equation (1).

A summary of our time-harmonic datasets are provided in Table 2.

A.1 Acoustic time-harmonic datasets

Boundary condition. In the experiments, we follow a geophysical configuration following Benitez et al. (2023). The boundary is divided into two distinct parts. We refer to the interface with the acoustic medium as Γ_1, where Dirichlet zero conditions (free surface) are enforced. In the remaining portion of the boundary, Γ_2 we assume *absorbing boundary conditions* to prevent waves from reflecting back to the medium. We define the boundary conditions of the equation (1) as follows.

$$p(\boldsymbol{x},\omega) = 0\,, \qquad \text{on } \Gamma_1 \text{ (free surface)}, \tag{6a}$$

$$\left(\partial_\nu - \frac{i\omega}{c(\boldsymbol{x})}\right) p(\boldsymbol{x},\omega) = 0\,, \qquad \text{on } \Gamma_2 \text{ (absorbing boundary conditions)}. \tag{6b}$$

The absorbing boundary conditions correspond to the fact that the domain is a numerical restriction of the earth (Engquist & Majda, 1977).

Source, frequencies, and the domain. Among time-harmonic experiments, we use the same domain and a fixed point source for all frequencies:

$$\text{Experiment config.} \begin{cases} \text{The domain is 2D having the size } 1.27\times 1.27\text{km}^2 \\ \text{The point source is at the top center of the domain} \\ 50\,000 \text{ GRF wave speeds generated, imposing } \ 1.5\text{km}\,\text{s}^{-1} \le c(x) \le 5\text{km}\,\text{s}^{-1} \\ \text{The data are } p \text{ that solve (1) at frequency } \omega/(2\pi) = 10,\ 15,\ 20 \text{ and } 40 \text{ Hz}. \end{cases} \tag{7}$$

Wavespeed. To stay in the statistical learning setup, we randomly generate wavespeed as the composition of an affine transformation and a Gaussian random field with the Whittle–Matérn covariance as described in Benitez et al. (2023). The smoothness parameter ν of the field is set to be 1 fot all the cases and for the two experiments we change $\boldsymbol{\lambda} = (\lambda_x, \lambda_y)$ as follows: (a) for the isotropic case we choose $\boldsymbol{\lambda} = (0.1, 0.1)$, and (b) for the anisotropic case we have $\boldsymbol{\lambda} = (0.2, 0.5)$.

A.2 Elastic isotropic time-harmonic datasets

Governing equation. In contrast with the acoustic wave problem which models a scalar pressure field p, the elastic case works with the displacement vector field $\boldsymbol{u}$. In 2D cases, the displacement field $\boldsymbol{u}$ contains two components (one per direction), which we denoted by $\boldsymbol{u}_x$ and $\boldsymbol{u}_z$; each component is complex-valued. The elastic isotropic time-harmonic equation has the form

$$-\rho(\boldsymbol{x})\omega^2 \boldsymbol{u}(\boldsymbol{x}) - \nabla\Big(\lambda(\boldsymbol{x})\nabla\cdot\boldsymbol{u}(\boldsymbol{x})\Big) - \nabla\cdot\Big(\mu(\boldsymbol{x})\Big[\nabla\boldsymbol{u}(\boldsymbol{x}) + \big(\nabla\boldsymbol{u}(\boldsymbol{x})\big)^\top\Big]\Big) = 0, \tag{8}$$

where ω is the angular frequency, $\rho(x)$ is the density, $\lambda(x)$ and $\mu(x)$ are the Lamé parameters (in particular, $\mu(x)$ is the shear modulus). In elastic media, two body waves propagate, the P-wave (compressional or primary wave) and S-wave (shear or secondary wave). Each wave is associated with a wavespeed, respectively $c_p(x)$ and $c_s(x)$, which can be used to characterize the medium as an alternative to the Lamé parameters, and are given by,

$$c_p(\boldsymbol{x}) \coloneqq \sqrt{\frac{\lambda(\boldsymbol{x}) + 2\mu(\boldsymbol{x})}{\rho(\boldsymbol{x})}}, \qquad c_s(\boldsymbol{x}) \coloneqq \sqrt{\frac{\mu(\boldsymbol{x})}{\rho(\boldsymbol{x})}}. \tag{9}$$

Throughout our elastic wave experiments, we let P-wavespeeds c_p, the S-wavespeed c_s be realizations of anisotropic GRFs while the density is kept to $\rho = 1$. More precisely, we first generate c_p as a GRF with values between 2.5 and 5.5 km.s^{-1}, we then generate a GRF function $\mathfrak{c}$ with values between 0.35 and 0.50 that we use as a scaling to create $c_s = \mathfrak{c}\, c_p$. Therefore, c_s is a randomly scaled version of c_p. This choice is motivated as c_p and c_s are physical properties of a medium that are expected to contain the same geometry of structures (e.g., μ and ρ appear in both c_p and c_s in (9)). For boundary conditions, we follow the same configuration as for the acoustic case with absorbing boundary conditions on the lateral and bottom boundaries, and a free-surface condition for the upper surface where (representing the interface between the ground and the air), see, e.g., Faucher (2017).

B Details of the time-varying datasets

To simulate wave propagation for both Reverse Time Continuation (RTC) and Inverse Scattering (IS) problems, we use the open-source `j-wave package` (Stanziola et al., 2023). The j-wave package simulates the wave dynamics in (3) by an equivalent system of first-order equations (Treeby et al., 2012; Pierce, 2019):

$$\begin{aligned}
\frac{\partial u}{\partial t} &= -\frac{1}{b_0}\nabla q, && \text{(momentum conservation)}\\
\frac{\partial b}{\partial t} &= -b_0 \nabla\cdot u, && \text{(mass conservation)}\\
q &= c^2 b. && \text{(pressure-density relation)}
\end{aligned}$$

where $u = u(\boldsymbol{x}, t)$ is called the acoustic particle velocity and b_0 is ambient density. Radiating boundary conditions are enforced with a perfectly matched layer (PML), following the default setting of `j-wave` (Stanziola et al., 2023).

In our simulation, the domain is represented as a square grid, with dimensions of 1.024 km $\times$ 1.024 km discretized into a 128×128 array. Recall that for the time-varying experiments, the wavespeed c can be a realization of an isotropic GRF, anisotropic GRF, or a Gaussian lens. In the case of isotropic and anisotropic GRF, the wavespeeds are taken from the time-harmonic datasets; in the case of the Gaussian-lens wavespeed,

Problem	Wavespeed c	Initial pressure $q(\cdot, 0)$
Reverse time continuation (RTC)	Gaussian lens	Thick lines
		MNIST
	Isotropic GRF	Thick lines
		MNIST
	Anisotropic GRF	Thick lines
		MNIST
Inverse source (IS)	Gaussian lens	Thick lines
		MNIST
	Isotropic GRF	Thick lines
		MNIST
	Anisotropic GRF	Thick lines
		MNIST

Table 3: **Summary of the 12 time-varying datasets.** Each dataset corresponds to specific problem types (reverse time continuation or inverse source), wavespeed variations (Gaussian lens, isotropic GRF, or anisotropic GRF), and initial pressure characteristics (thick lines or MNIST). The thick line initial pressure datasets consist of in-distribution samples: they contain **9000 training samples, 500 validation samples, and 500 testing samples**. The MNIST pressure dataset consists of out-of-distribution (OOD) samples used exclusively for testing and comprises **500 samples**.

the wavespeed is a point mass situated at the grid coordinates of $(50, 55)$ blurred by a Gaussian filter with a standard deviation 50 in both spatial directions. Across all types of wavespeeds considered in our simulations, the minimum wavespeed is normalized to a value of 1.4 $\mathrm{km\,s^{-1}}$, while the maximum wavespeed is normalized to 4 $\mathrm{km\,s^{-1}}$. The propagation time for both the RTC and IS simulations is set to be $T = 0.2$ s.

For both RTC and IS datasets, the initial pressure datasets $q(\cdot, 0)$ can either be thick lines or MNIST images, represented by 128×128 arrays; see Figure 3 and Figure 4. The thick lines represent pressures that are used as in-distribution samples for training and evaluating the model. They are box-like patterns of random sizes, orientations, and locations in the domain, following the approach in Kothari et al. (2020). Each sample contains 5 to 10 boxes, uniformly distributed. The box centroids are sampled on the discretized grid of the domain. Dimensions of boxes are sampled from uniform distributions: length from $[50, 100]$, width from $[20, 40]$, and orientation from $[0, \pi]$. The dataset consists of 9000 training samples, 500 validation samples, and 500 test samples. Additionally, there are 500 out-of-distribution (OOD) MNIST pressure samples for testing.

The IS problem is more challenging than RTC. This is because in IS we only get to measure the wave pressure at the top of the domain. That is, the sensor locations $\mathcal{S}$ in the measurements $[q(\boldsymbol{x}, t)]_{\boldsymbol{x}\in\mathcal{S}, t\in\mathcal{T}}$ correspond to the topmost coordinates of the domain (excluding the size of PML). The time steps $\mathcal{T}$ consist of 128 equidistant intervals within the range of $[0, T]$. These settings result in the sensor record $[q(\boldsymbol{x}, t)]_{\boldsymbol{x}\in\mathcal{S}, t\in\mathcal{T}}$ having a square image-like appearance as in Figure 4. To make the IS problem more tractable to solve, we use the following way to prepare the initial pressure $q(\cdot, 0)$. We resize thick line and MNIST images that with an original size 128×128 into the size of 64×64 and put them on the top center of the domain. The remaining entries are filled with zeros. Consequently, all nonzero entries of the initial wave pressure $q(\cdot, T)$ are concentrated in the top center of the domain. This configuration, with the sensors positioned at the top, allows for better reception of propagated waves and helps mitigate the ill-posedness of the problem.

C Full experimental results of the acoustic time-harmonic datasets

Wavespeed c	Freq. $\omega/2\pi$	FNO-depth-4	FNO-depth-8	U-Net-ch-32	U-Net-ch-64	UNO-modes-12	UNO-modes-16
Isotropic GRF	10 Hz	0.063	0.040	0.073	0.063	0.064	0.054
	15 Hz	0.093	0.057	0.116	0.087	0.106	0.081
	20 Hz	0.122	0.070	0.157	0.106	0.147	0.114
	40 Hz	0.283	0.165	0.286	0.191	0.407	0.301
Anisotropic GRF	10 Hz	0.059	0.025	0.144	0.119	0.074	0.051
	15 Hz	0.098	0.039	0.204	0.165	0.123	0.093
	20 Hz	0.135	0.060	0.230	0.176	0.171	0.129
	40 Hz	0.315	0.172	0.321	0.231	0.422	0.343

Table 4: In-distribution performance comparison of models on the test folds of the time-harmonic datasets. The error metric is the relative L2 error $\|p - \widehat{p}\|_{L^2} / \|p\|_{L^2}$ between the ground-truth p and prediction $\widehat{p}$.

Wavespeed c	Freq. $\omega/2\pi$	FNO-depth-4	FNO-depth-8	U-Net-ch-32	U-Net-ch-64	UNO-modes-12	UNO-modes-16
Isotropic GRF	10 Hz	0.485	0.379	0.527	0.506	0.489	0.458
	15 Hz	0.633	0.464	0.638	0.620	0.674	0.618
	20 Hz	0.758	0.533	0.751	0.717	0.770	0.747
	40 Hz	1.152	0.895	0.883	0.891	0.893	0.951
Anisotropic GRF	10 Hz	0.560	0.388	0.541	0.527	0.376	0.382
	15 Hz	0.771	0.518	0.671	0.656	0.498	0.483
	20 Hz	0.812	0.612	0.754	0.725	0.607	0.599
	40 Hz	1.018	0.887	0.905	0.898	0.803	0.950

Table 5: OOD performance comparison of models on the test folds of the time-harmonic datasets. The table layout is similar to Table 4. The wavespeed c reported in the left column shows the OOD wavespeed used in test data; for instance, the three "Isotropic GRF" rows are based on models trained on the corresponding Anisotropic GRF versions, and vice versa.

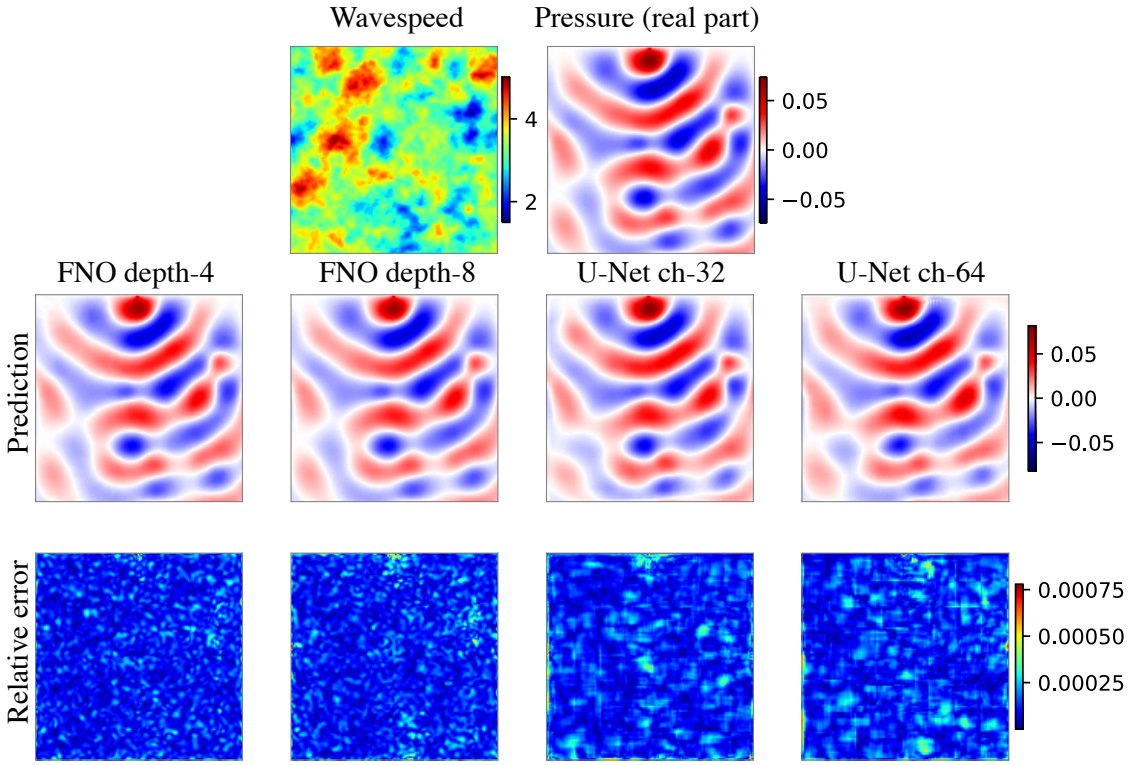

Figure 8: In-distribution test performance of models on an acoustic time-harmonic dataset. The time-harmonic dataset is configured with isotropic GRF wavespeed and frequency $\omega/2\pi = 10$Hz. The figure layout is same as Figure 5 in the main text.

D Full experimental results of the elastic time-harmonic datasets

Freq. $\omega/2\pi$	FNO-depth-4	FNO-depth-8	U-Net-ch-32	U-Net-ch-64	UNO-modes-12	UNO-modes-16
10 Hz	0.080	0.039	0.154	0.141	0.103	0.076
15 Hz	0.130	0.072	0.229	0.217	0.205	0.127
20 Hz	0.225	0.124	0.267	0.225	0.264	0.204
40 Hz	0.504	0.365	0.497	0.470	0.534	0.490

Table 6: In-distribution performance comparison of models on the test folds of the elastic time-harmonic datasets. Wavespeeds are anisotropic GRFs for all frequencies. The error metric is the relative L2 error $\|\boldsymbol{u} - \widehat{\boldsymbol{u}}\|_{L^2} / \|\boldsymbol{u}\|_{L^2}$ between the ground-truth $\boldsymbol{u}$ and prediction $\widehat{\boldsymbol{u}}$.

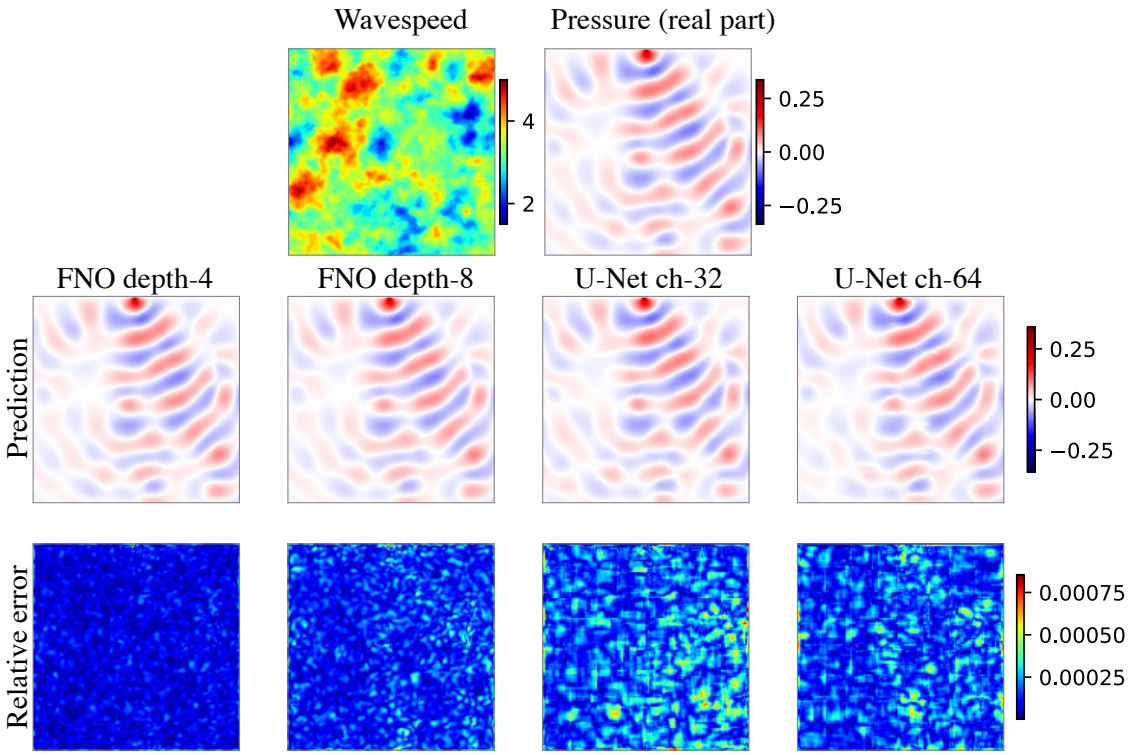

Figure 9: In-distribution test performance of models on a acoustic time-harmonic dataset. The time-harmonic dataset is configured with with isotropic GRF wavespeed and frequency $\omega/2\pi = 15$Hz. The figure layout is same as Figure 5 in the main text.

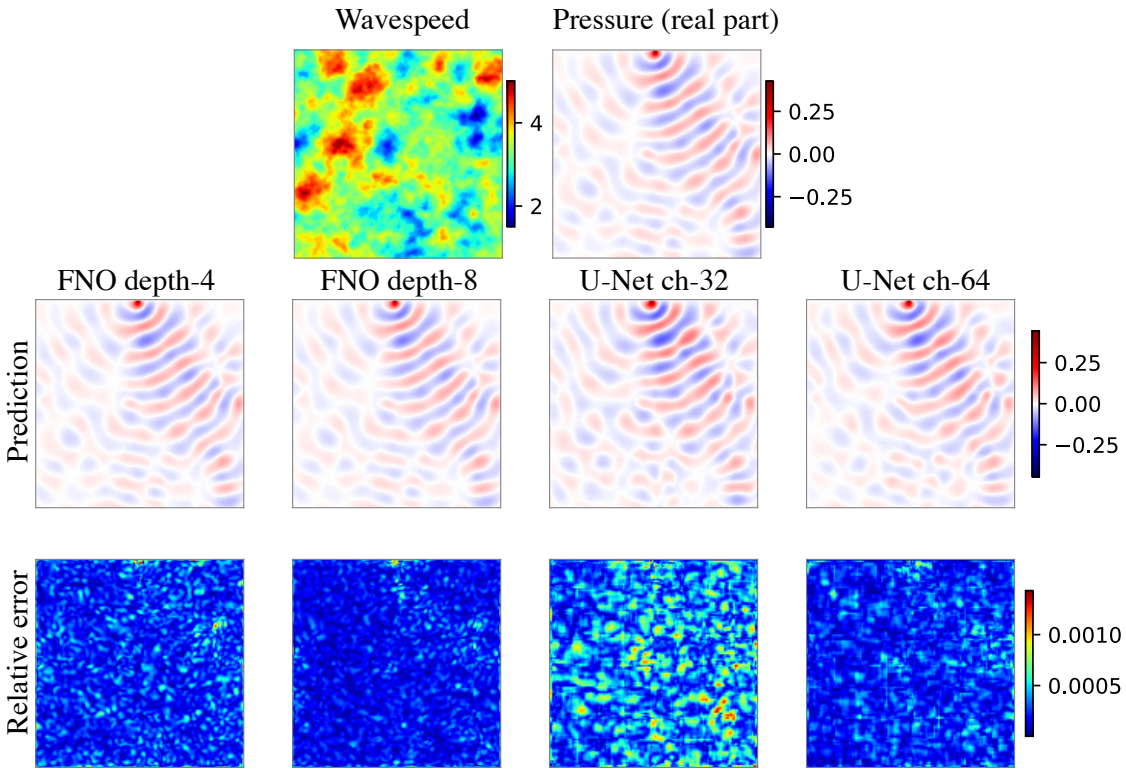

Figure 10: In-distribution test performance of models on a acoustic time-harmonic dataset. The time-harmonic dataset is configured with isotropic GRF wavespeed and frequency $\omega/2\pi = 20$Hz. The figure layout is same as Figure 5 in the main text.

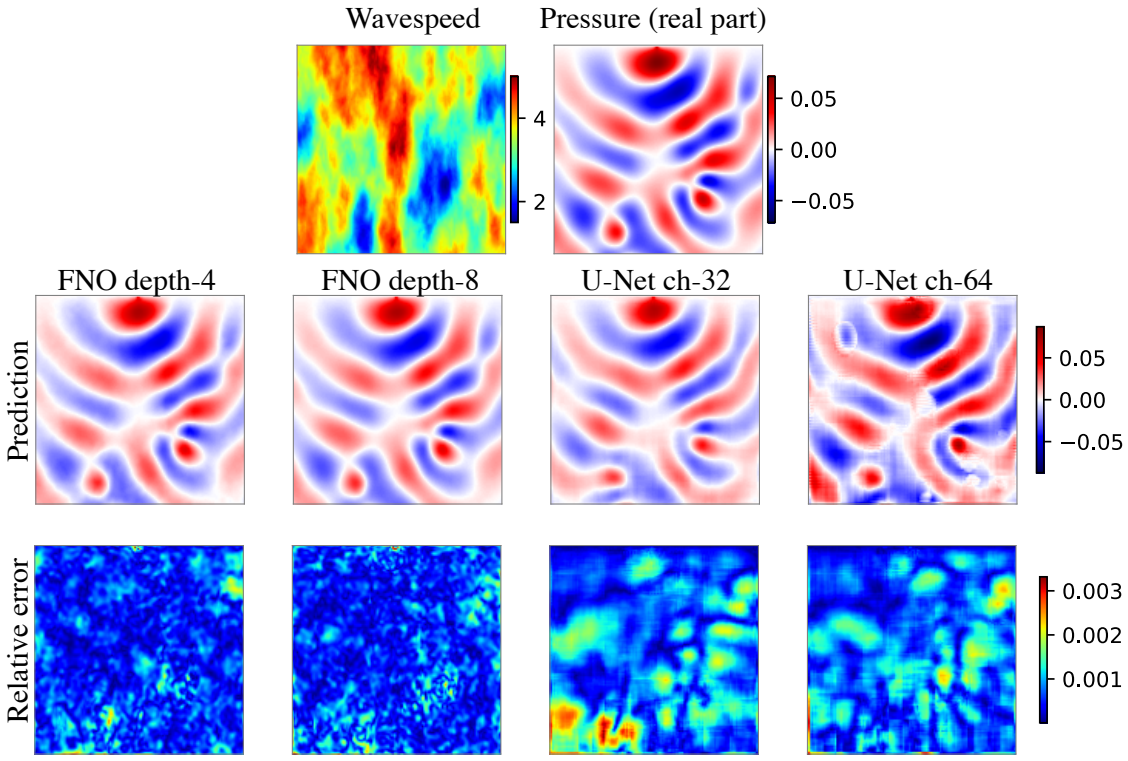

Figure 11: In-distribution test performance of models on a acoustic time-harmonic dataset. The time-harmonic dataset is configured with anisotropic GRF wavespeed and frequency $\omega/2\pi = 10$Hz. The figure layout is same as Figure 5 in the main text.

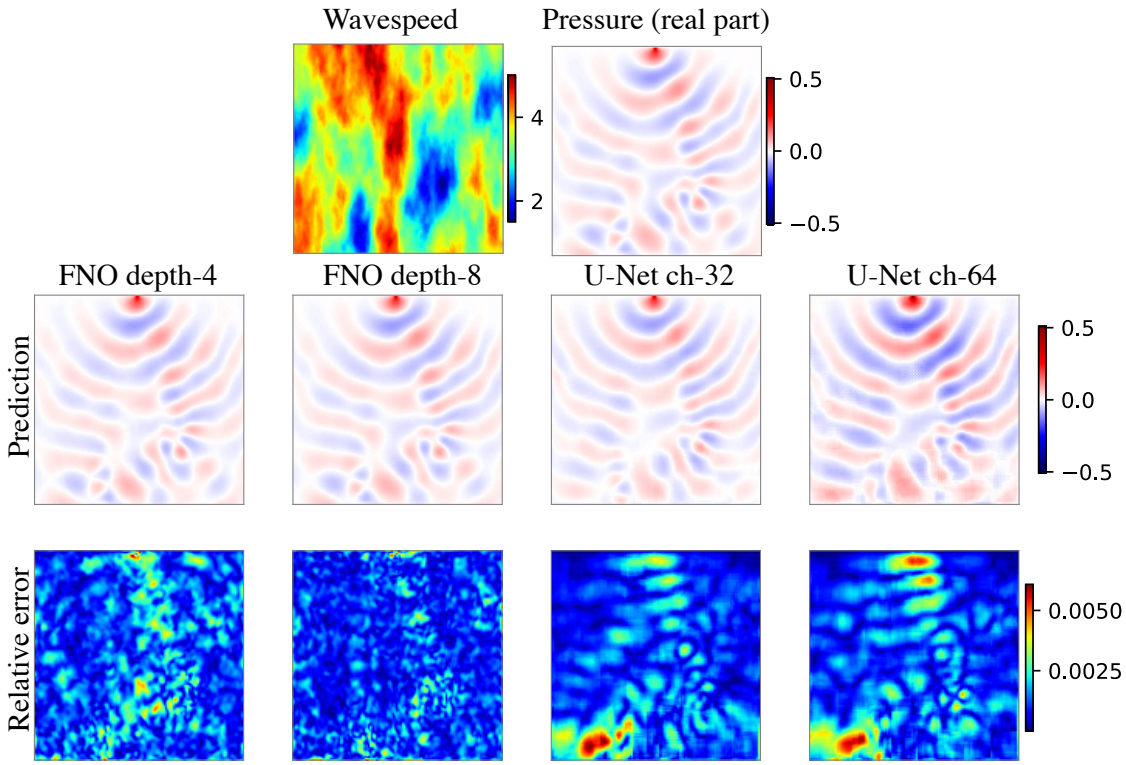

Figure 12: In-distribution test performance of models on a acoustic time-harmonic dataset. The time-harmonic dataset is configured with anisotropic GRF wavespeed and frequency $\omega/2\pi = 15$Hz. The figure layout is same as Figure 5 in the main text.

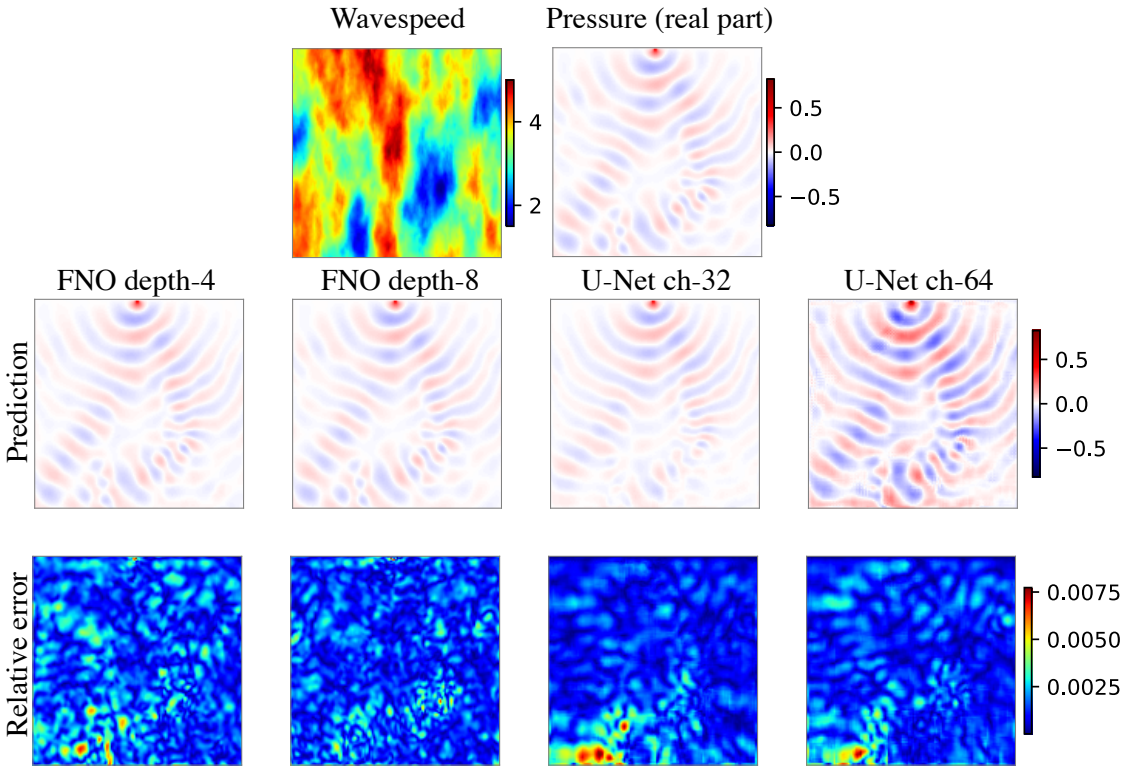

Figure 13: In-distribution test performance of models on a acoustic time-harmonic dataset. The time-harmonic dataset is configured with anisotropic GRF wavespeed and frequency $\omega/2\pi = 20$Hz. The figure layout is same as Figure 5 in the main text.

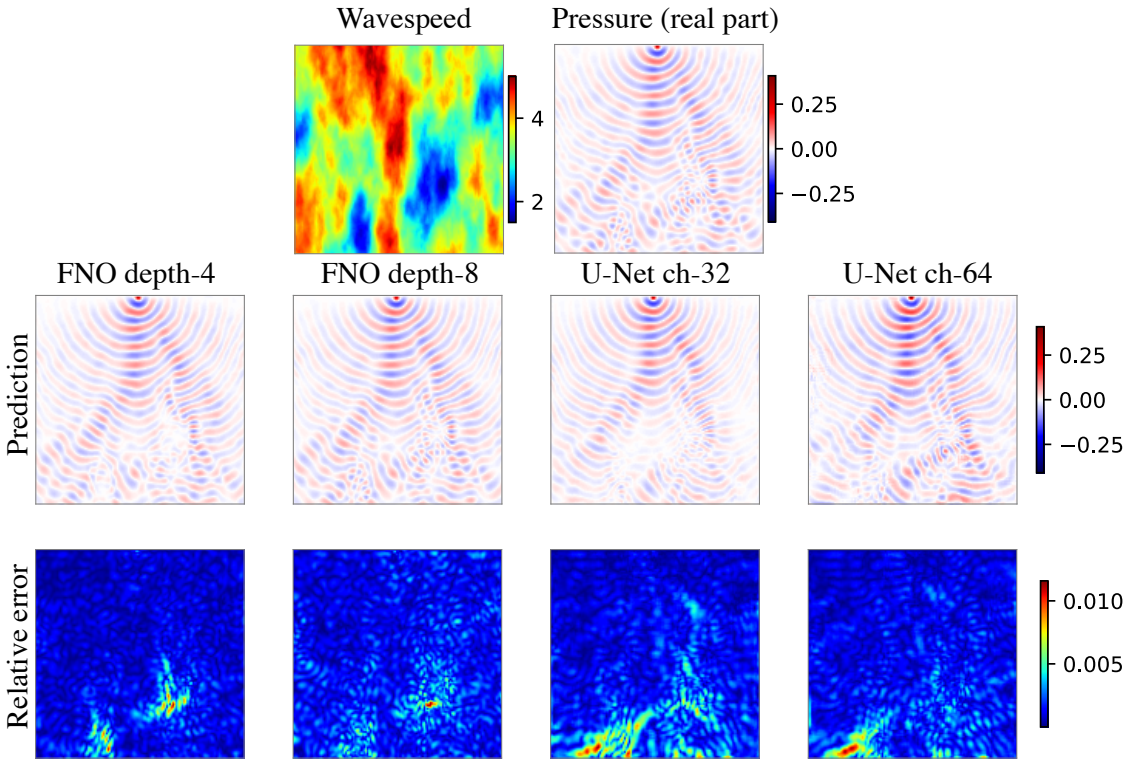

Figure 14: In-distribution test performance of models on a acoustic time-harmonic dataset. The time-harmonic dataset is configured with anisotropic GRF wavespeed and frequency $\omega/2\pi = 40$Hz. The figure layout is same as Figure 5 in the main text.

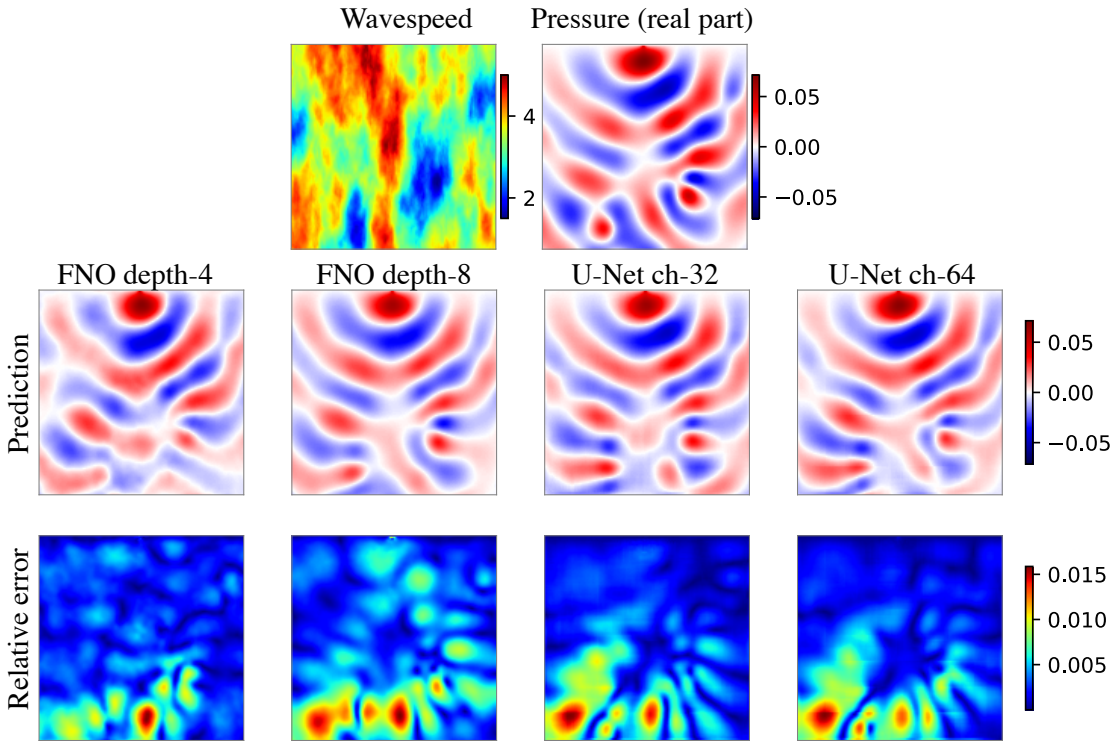

Figure 15: OOD test performance of models on a time-harmonic dataset. The model is trained on the time-harmonic dataset with *isotropic* GRF wavespeed and frequency $\omega/2\pi = 10$Hz, but tested on the *anisotropic* version instead. The figure layout follows from Figure 6 in the main text.

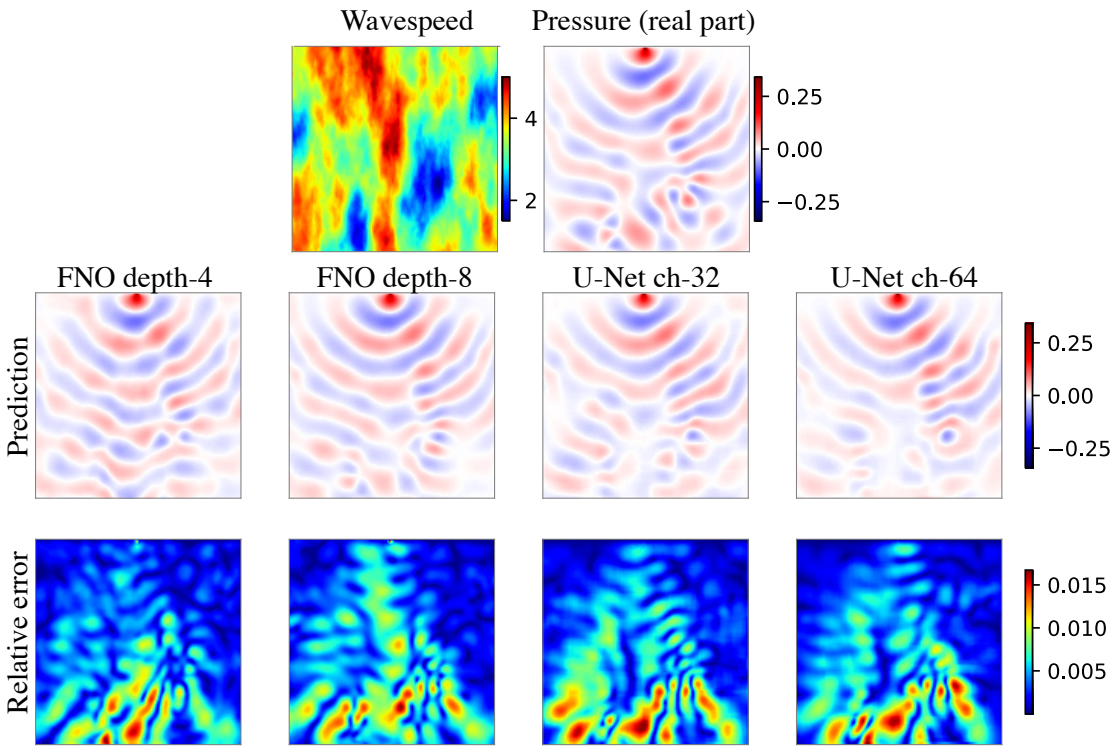

Figure 16: OOD test performance of models on a time-harmonic dataset. The model is trained on the time-harmonic dataset with *isotropic* GRF wavespeed and frequency $\omega/2\pi = 15$Hz, but tested on the *anisotropic* version instead. The figure layout follows from Figure 6 in the main text.

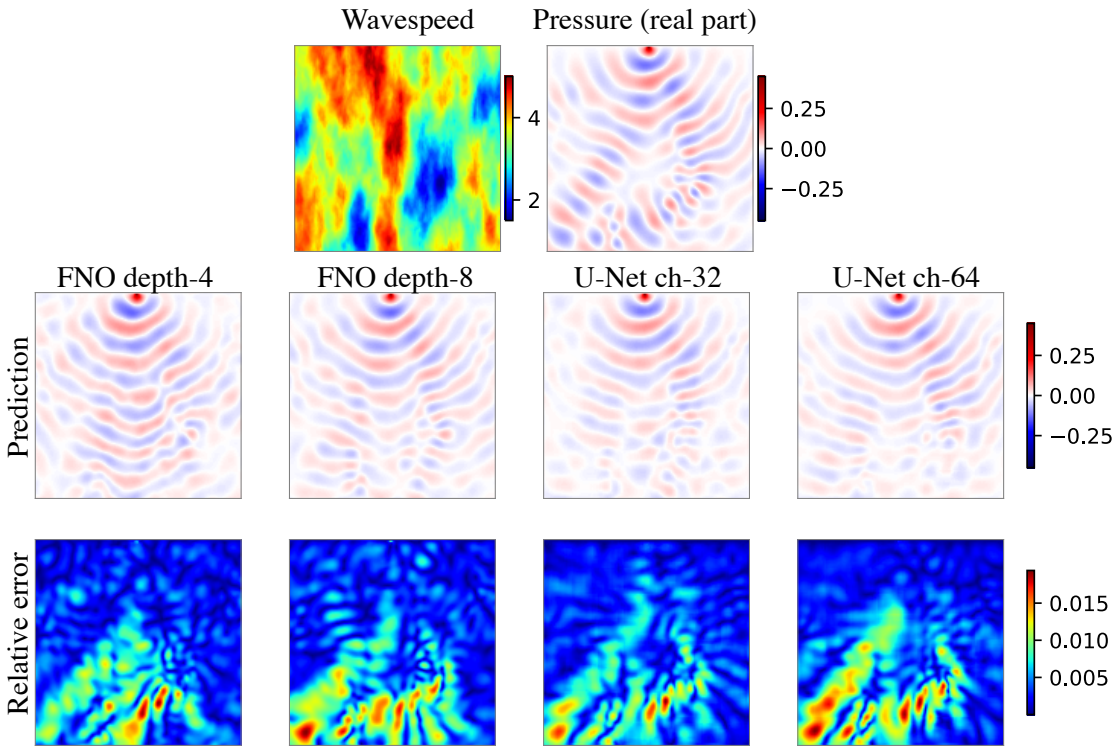

Figure 17: OOD test performance of models on a time-harmonic dataset. The model is trained on the time-harmonic dataset with *isotropic* GRF wavespeed and frequency $\omega/2\pi = 20$Hz, but tested on the *anisotropic* version instead. The figure layout follows from Figure 6 in the main text.

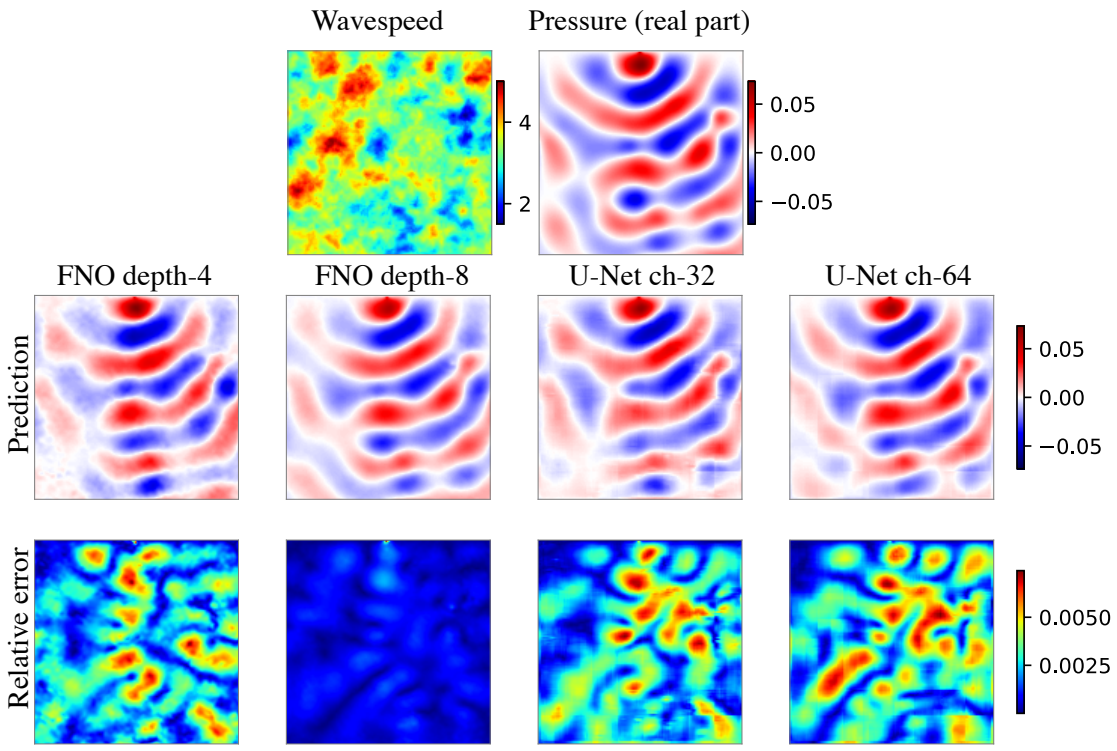

Figure 18: OOD test performance of models on a time-harmonic dataset. The model is trained on the time-harmonic dataset with *isotropic* GRF wavespeed and frequency $\omega/2\pi = 10$Hz, but tested on the *anisotropic* version instead. The figure layout follows from Figure 6 in the main text.

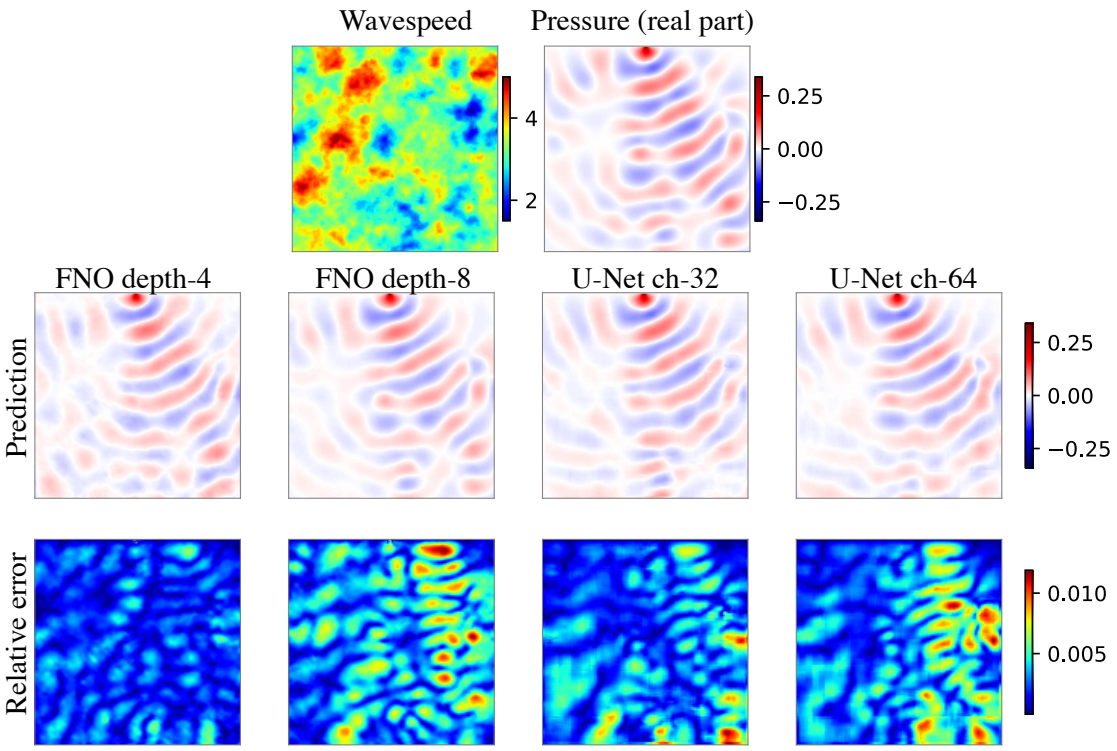

Figure 19: OOD test performance of models on a time-harmonic dataset. The model is trained on the time-harmonic dataset with *anisotropic* GRF wavespeed and frequency $\omega/2\pi = 15$Hz, but tested on the *isotropic* version instead. The figure layout follows from Figure 6 in the main text.

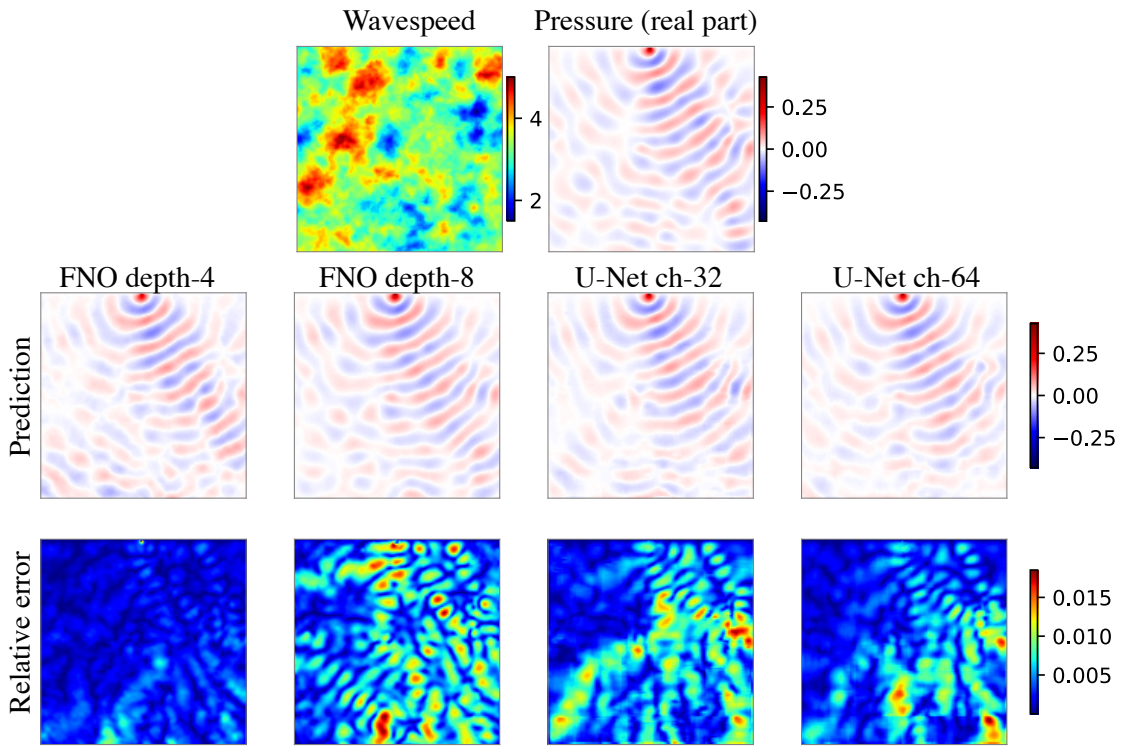

Figure 20: OOD test performance of models on a time-harmonic dataset. The model is trained on the time-harmonic dataset with *anisotropic* GRF wavespeed and frequency $\omega/2\pi = 20$Hz, but tested on the *isotropic* version instead. The figure layout follows from Figure 6 in the main text.

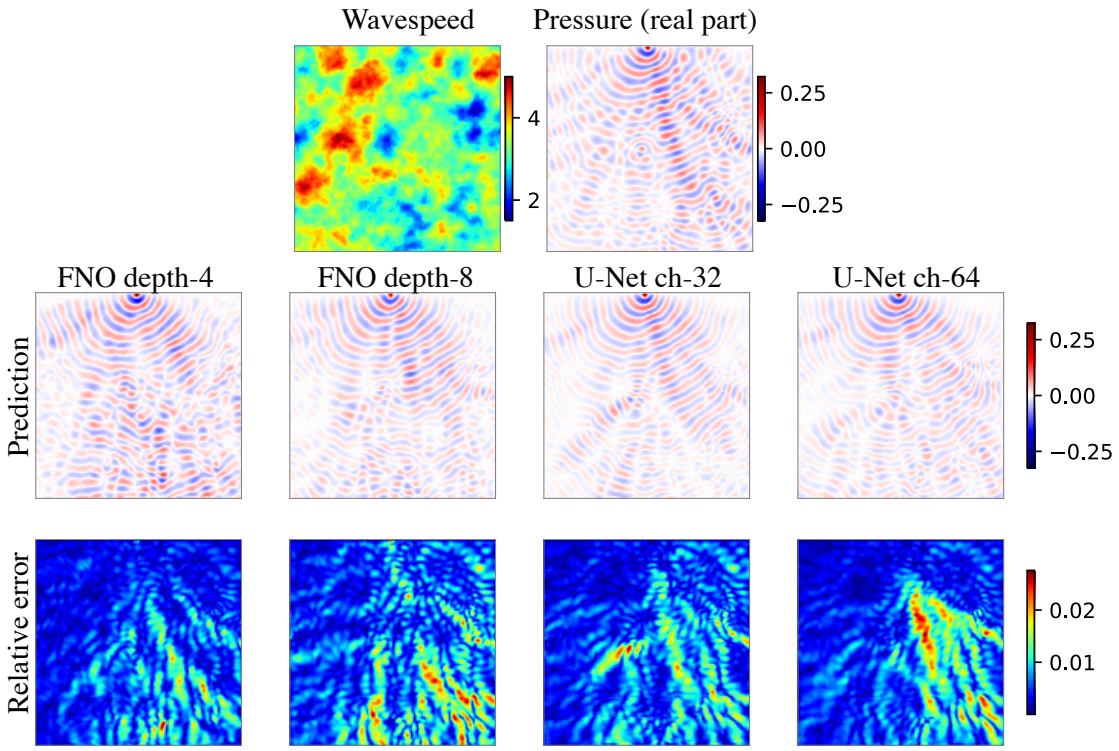

Figure 21: OOD test performance of models on a time-harmonic dataset. The model is trained on the time-harmonic dataset with *anisotropic* GRF wavespeed and frequency $\omega/2\pi = 40$Hz, but tested on the *isotropic* version instead. The figure layout follows from Figure 6 in the main text.

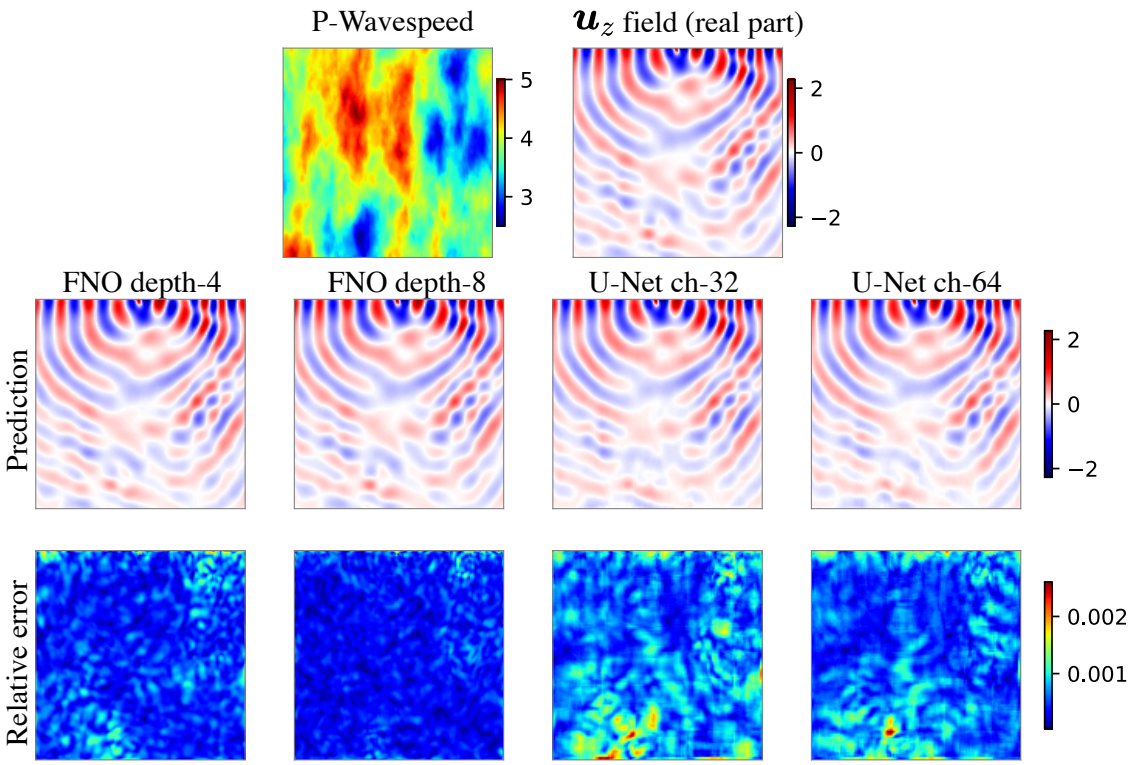

Figure 22: In-distribution test performance of models on an elastic time-harmonic dataset. The time-harmonic dataset is configured with anisotropic GRF wavespeed and frequency $\omega/2\pi = 10$Hz.

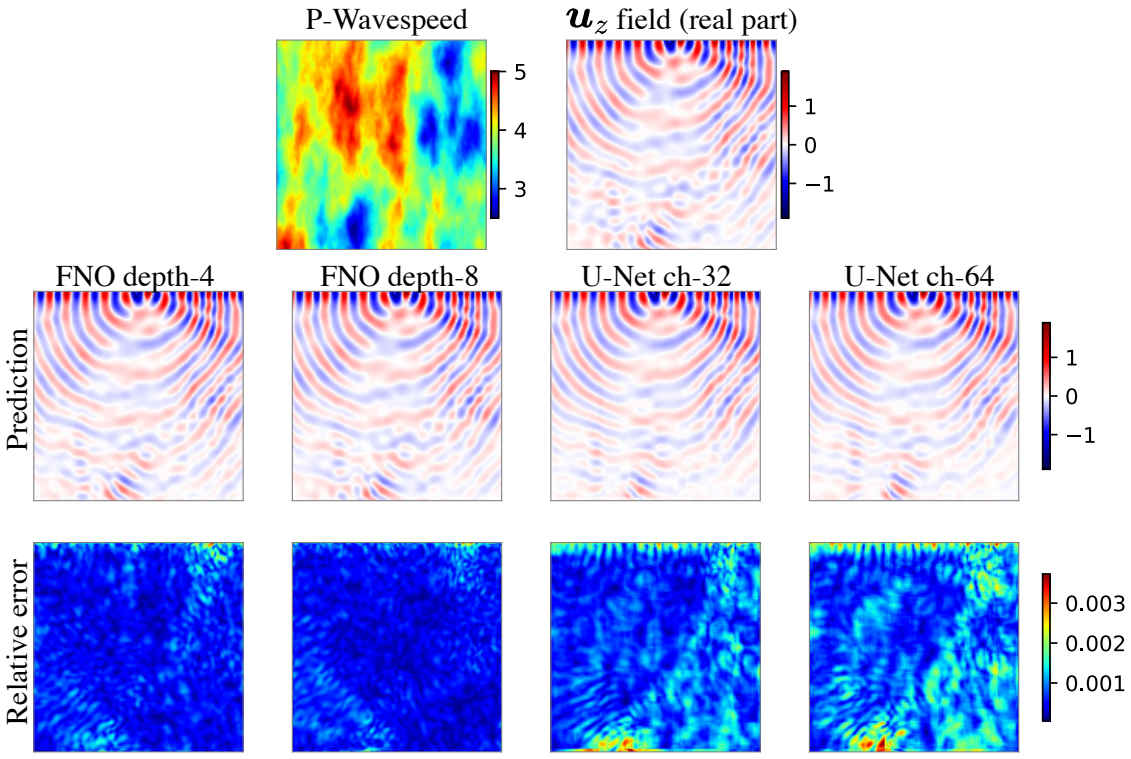

Figure 23: In-distribution test performance of models on an elastic time-harmonic dataset. The time-harmonic dataset is configured with anisotropic GRF wavespeed and frequency $\omega/2\pi = 15$Hz.

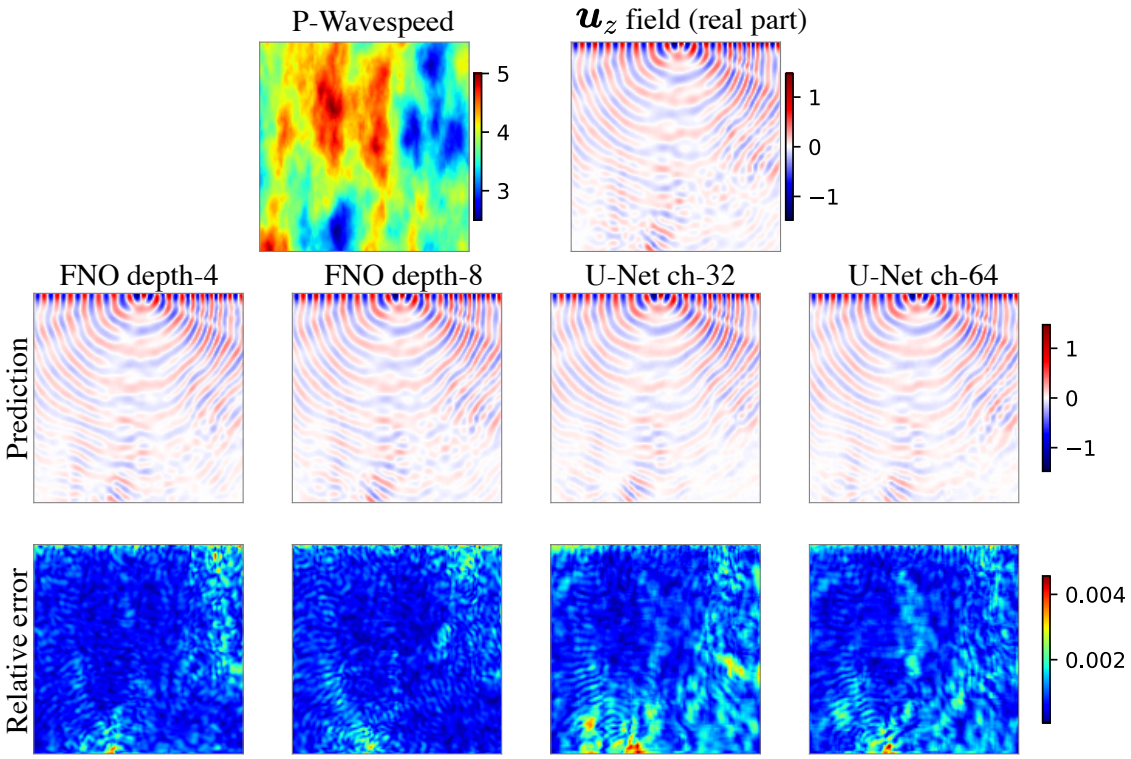

Figure 24: In-distribution test performance of models on an elastic time-harmonic dataset. The time-harmonic dataset is configured with anisotropic GRF wavespeed and frequency $\omega/2\pi = 20$Hz.

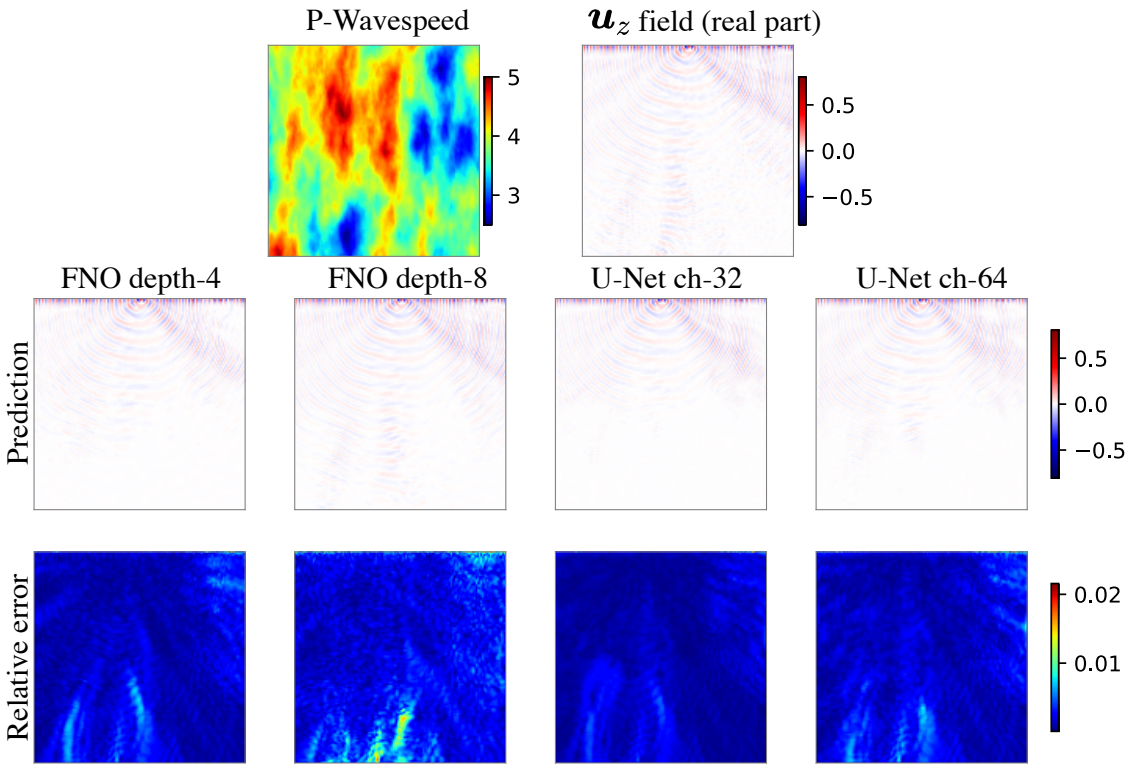

Figure 25: In-distribution test performance of models on an elastic time-harmonic dataset. The time-harmonic dataset is configured with anisotropic GRF wavespeed and frequency $\omega/2\pi = 40$Hz.

E Full experimental results of the time-varying datasets

Problem	Wavespeed	Init. press.	FNO-depth-4	FNO-depth-8	U-Net-ch-32	U-Net-ch-64	UNO-modes-12	UNO-modes-16
RTC	Gaussian lens	Thick lines	0.421	0.381	0.393	0.371	0.466	0.433
		MNIST	0.461	0.410	0.525	0.520	0.478	0.434
	Iso. GRF	Thick lines	0.365	0.329	0.451	0.430	0.411	0.378
		MNIST	0.342	0.349	0.535	0.517	0.337	0.329
	Aniso. GRF	Thick lines	0.348	0.308	0.432	0.414	0.395	0.362
		MNIST	0.378	0.377	0.469	0.491	0.364	0.362
IS	Gaussian lens	Thick lines	0.550	0.500	0.446	0.443	0.551	0.545
		MNIST	0.695	0.686	0.629	0.627	0.652	0.667
	Iso. GRF	Thick lines	0.436	0.380	0.383	0.356	0.462	0.441
		MNIST	0.489	0.479	0.511	0.511	0.463	0.465
	Aniso. GRF	Thick lines	0.415	0.351	0.359	0.327	0.436	0.419
		MNIST	0.471	0.449	0.537	0.514	0.413	0.414

Table 7: **Performance comparison of models on the test folds of the time-varying datasets.** The error metric is the relative L2 error $\|p - \widehat{p}\|_{L^2} / \|p\|_{L^2}$ between the ground-truth p and prediction $\widehat{p}$.

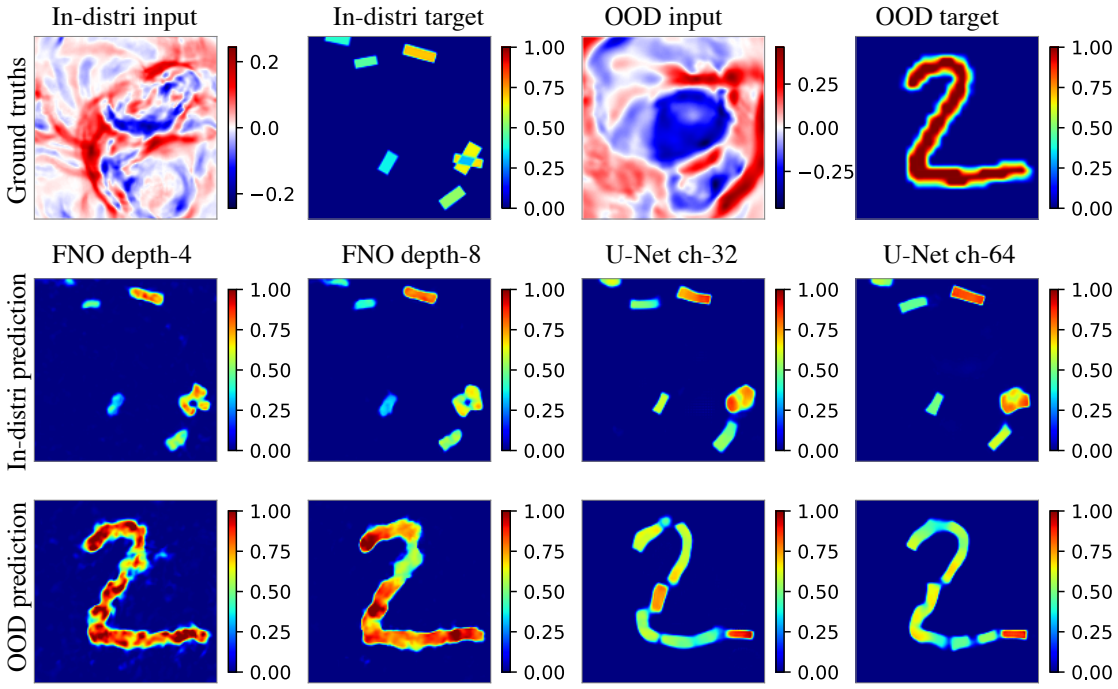

Figure 26: Test performance of models on the time-varying RTC dataset with isotropic GRF wavespeed. The first row shows the input and target samples of the RTC dataset; they can be either in-distribution or OOD. The second row shows the model predictions on the in-distribution sample. The second row shows the model predictions on the OOD sample.

F Case study: Gradually challenging the OOD generalization

We have seen the limitation of PDE surrogates on OOD samples. We note that the comparison "in-distribution vs OOD" is a simplified, binary notion, as it implies that a sample is either within a distribution or outside of it. More fine-grained OOD notions are helpful, as intuitively the performance of models on a sample may depend on the degree to which that sample differs from those in the training distribution.

In this section, we present a case study where we vary wavespeeds from near in-distribution samples ("less OOD") to distant ones ("more OOD"). See Figure 31 for visualization. We employ neural style transfer (NST) (Gatys et al., 2016) to create '0' digits of different OOD levels. These images are then used as wavespeeds in time-varying problems.

Recall that the NST algorithm separates and recombines the content and the style of images. Our content of interest is a '0' digit and the style of interest box-like strokes. The style and content of an image are balanced by a content weighting factor; see Gatys et al. (2016) for details. By adjusting the content weighting factor (referred to as OOD weight in Figure 31) from small to large, we generate images that resemble in-distribution thick lines (left panels of Figure 31(A)) and OOD MNIST (right panels of Figure 31(A)). We use the open-source NST implementation[1] for data generation.

Using different OOD degrees for wavespeed samples, we assess various PDE surrogates on RTC and IS tasks. Figure 31(B) displays the numerical results. As anticipated, higher OOD degrees' wavespeeds are harder to recover, giving higher relative errors.

[1] https://github.com/crowsonkb/style-transfer-pytorch

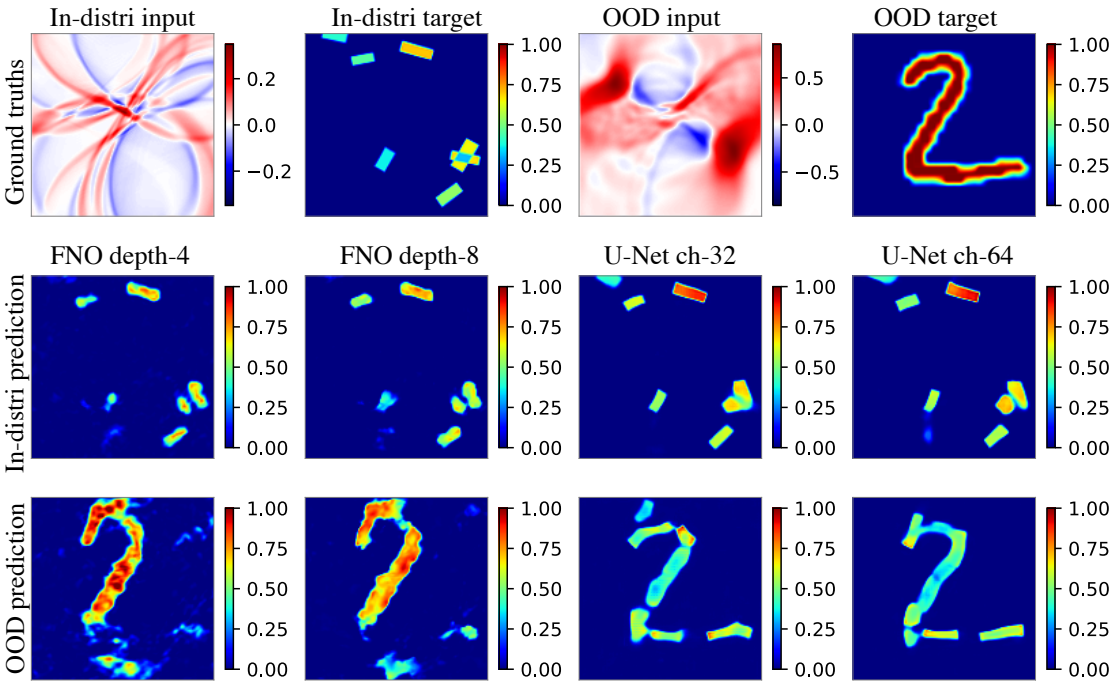

Figure 27: Test performance of models on the time-varying RTC dataset with Gaussian lens wavespeed. The figure layout is the same as Figure 26.

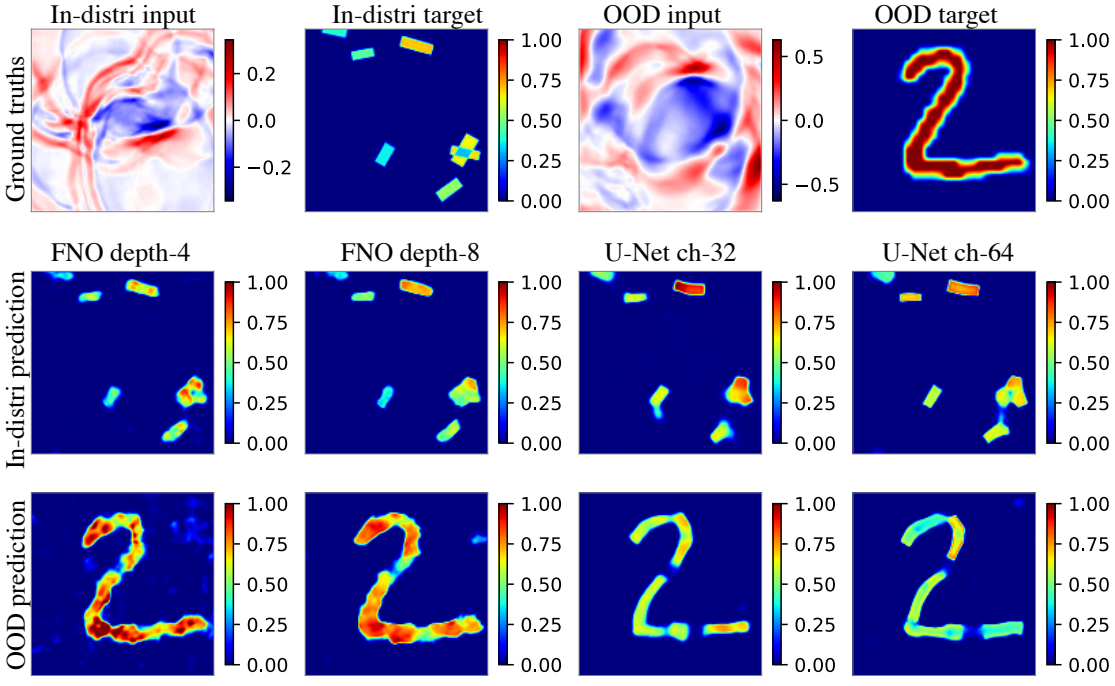

Figure 28: Test performance of models on the time-varying RTC dataset with anisotropic GRF wavespeed. The figure layout is the same as Figure 26.

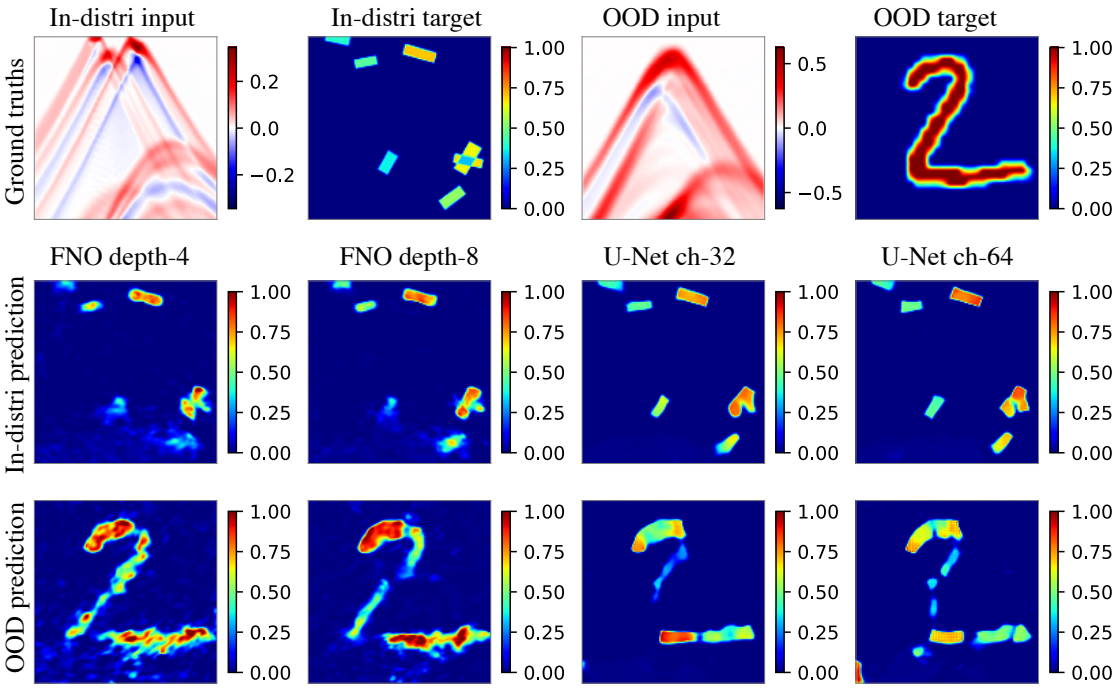

Figure 29: Test performance of models on the time-varying IS dataset with Gaussian lens wavespeed. The figure layout is the same as Figure 7.

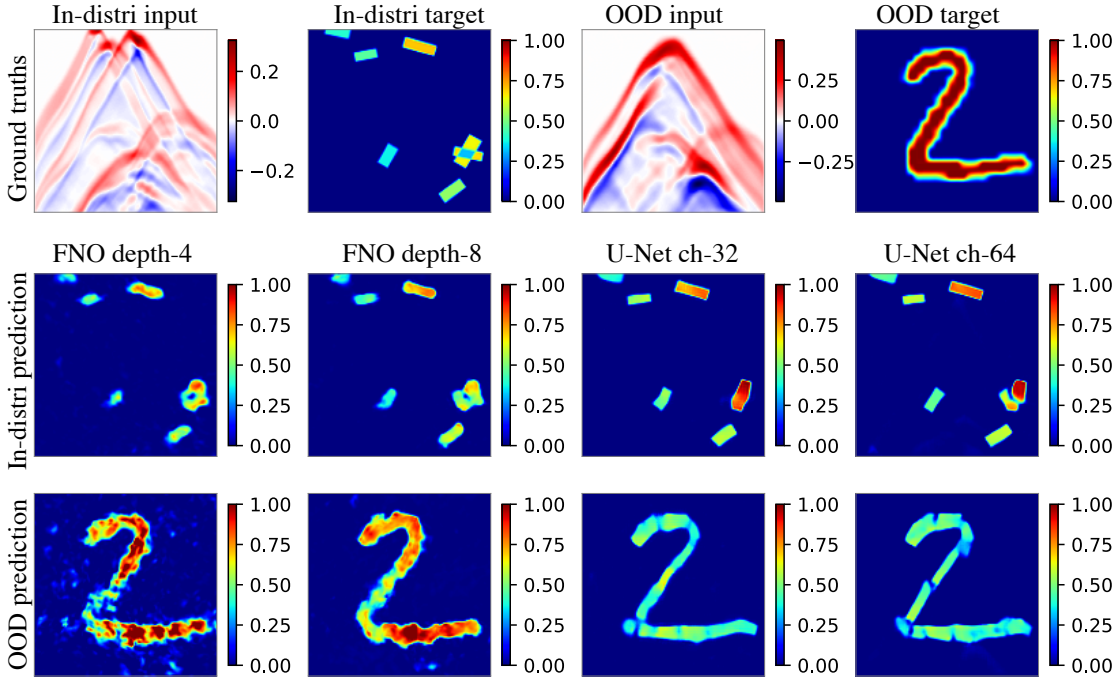

Figure 30: Test performance of models on the time-varying IS dataset with anisotropic GRF wavespeed. The figure layout is the same as Figure 7.

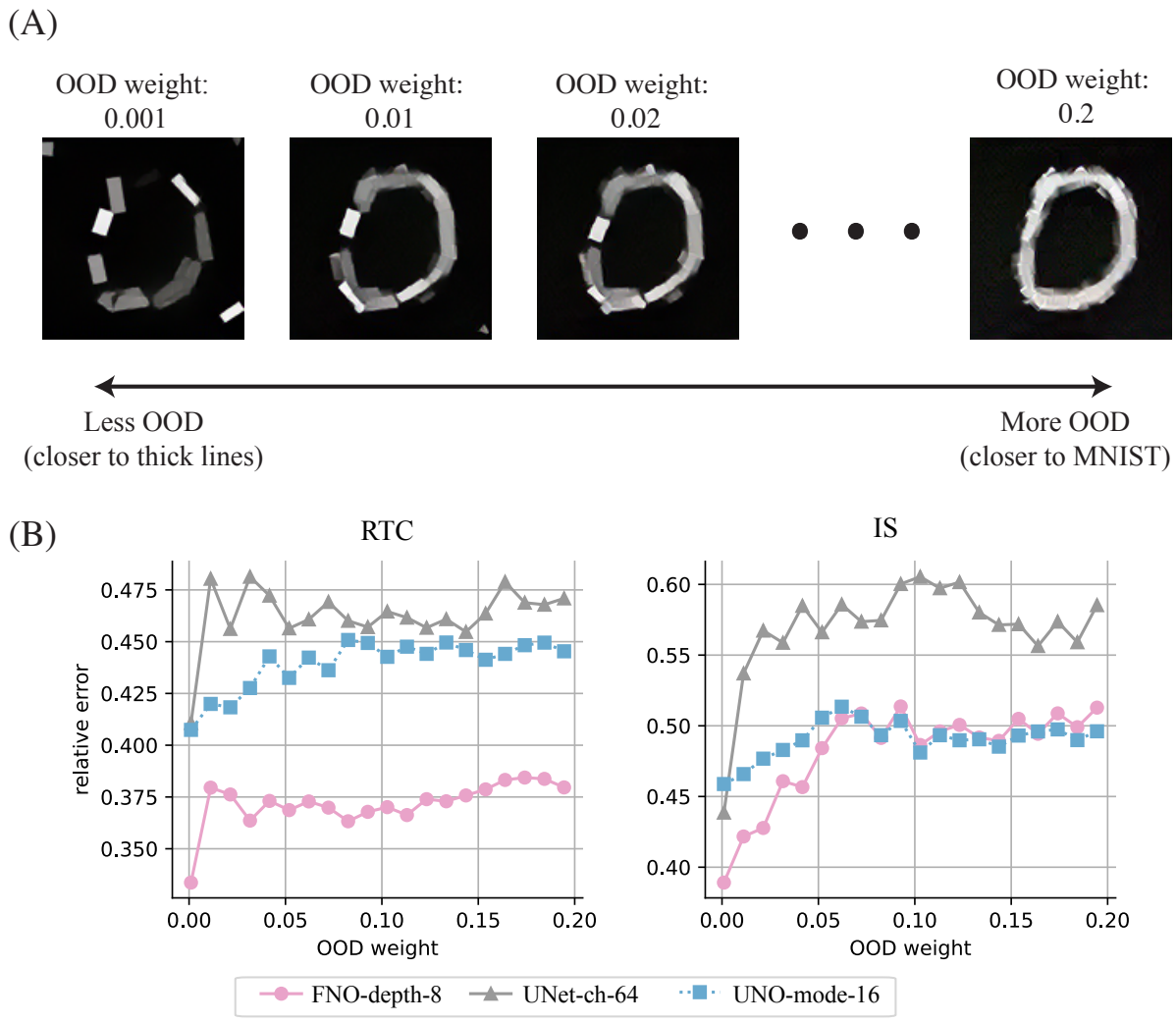

Figure 31: Case study: Model performance and OOD degree. Panel (A): Samples from neural style transfer, with content factor weights as 20 OOD weights ranging from 0.001 to 0.2. Panel (B): Models' performance on diverse wavespeed samples. More OOD wavespeed (closer to MNIST than thick lines) results in higher errors.