Towards Large Scientific Learning Models with In-Context Operator Networks (ICON)

Stanley J. Osher
Joint work with Liu Yang, Siting Liu, Tingwei Meng

Outline

- **Background: A Foundational Shift in AI**
- **In-Context Operator Networks (ICON)**
  - Motivation
  - Methodology
  - Numerical Results
    - One Model for a Wide Range of Scientific Machine Learning Problems
    - Comparison with Classic Operator Learning
    - Multi-Modal ICON (text + data)
    - A Study on Conservation Laws; Generalization to New PDEs
- **ICON and Conditional Generative Modeling**
- **Summary**

A Foundational Shift in AI

- **Previous AI paradigm**
  - Task-specific dataset/architecture/loss -> task-specific model
- **Large Language Models (LLMs), e.g., GPT**
  - One model for a broad array of tasks (foundation model)
- **In-Context Learning**
  - Unify multi-tasks with "next token prediction"
  - Specify the task in the "prompt", i.e. model input
  - Prompt: including instruction, backgrounds, examples, etc...
  - No training for each task
- **Benefits**
  - Efficient in marginal cost
  - More powerful for individual tasks, due to cross-task knowledge transfer
    - Train the model to write code helps logical deduction
  - Emergent capabilities when scaling up
    - Generalize to new tasks beyond training distribution, even beyond human's expectations.

Motivation

- Act 1: train a model to approximate the **solution function**
  - e.g., Deep Galerkin method, Deep Ritz method, Physics-Informed Neural Networks
  - vector -> model -> vector
- Act 2: train a model to approximate the **solution operator** (operator learning)
  - e.g., Fourier Neural Operator, DeepONet
  - function -> model -> function

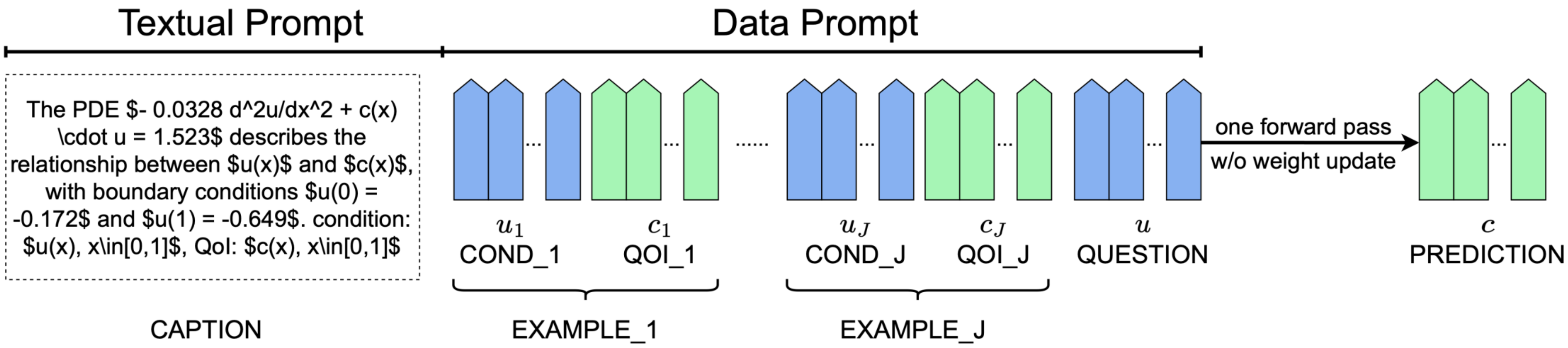
- These neural-network-based methods are **task-specific** and need **frequent retraining**.
- We need a model that adapts to new physical systems and tasks, just as a human would.

- Act 3: train a model as an **operator learner** (in-context operator learning)
  - In-Context Operator Networks (ICON)
  - A **single** model for a wide range of scientific learning problems
  - **Learn and apply operator in the forward propagation, without weight updates**
  - Generalization to new operators, even new equations

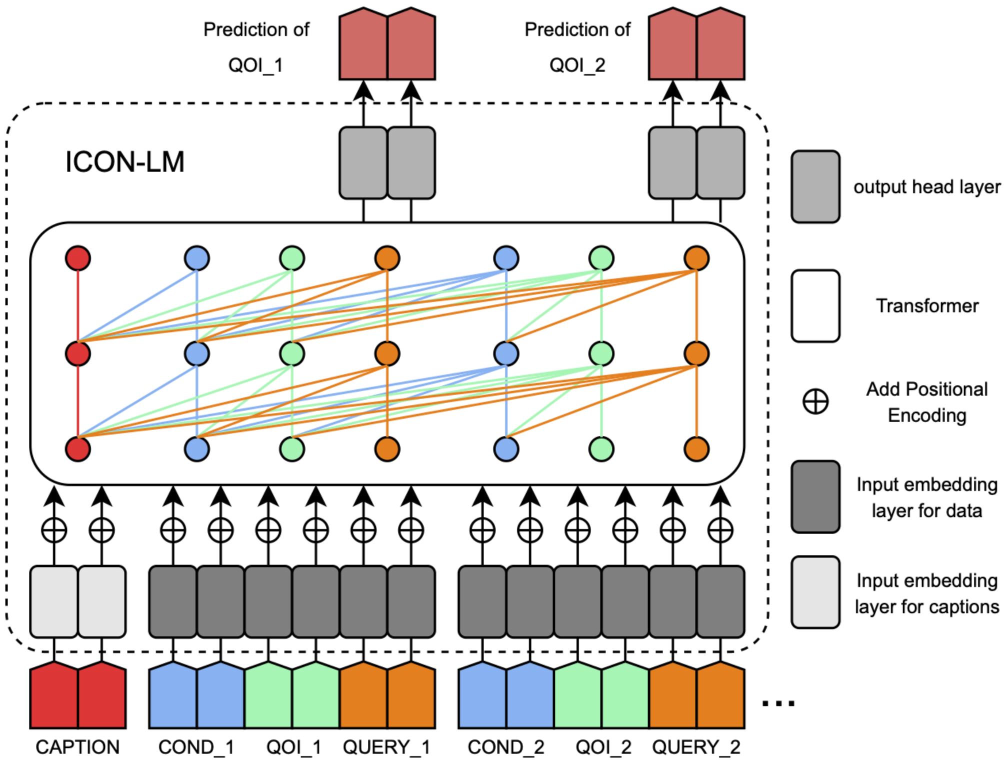
In-Context Operator Learning

- Learn from function examples, i.e. condition-QoI function pairs.
  - Condition/QoI (quantity of interest) denotes the operator input/output.
  - Each function is represented by multiple tokens.

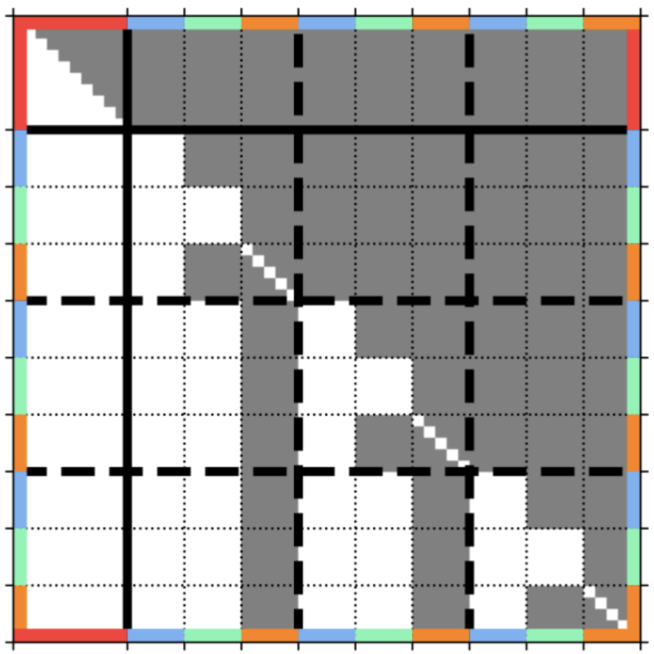
- Multi-modal: apart from function examples, optionally take "captions" as input.
  - Captions: texts that integrate human knowledge, in natural language and equations written in LaTeX.
  - A different approach for physics-informed models.
  - We will mostly focus on single-modal learning without captions.


Textual Prompt
Data Prompt
The PDE $- 0.0328 d^2u/dx^2 + c(x) \cdot u = 1.523$ describes the relationship between $u(x)$ and $c(x)$, with boundary conditions $u(0) = -0.172$ and $u(1) = -0.649$. condition: $u(x), x\in[0,1]$, QoI: $c(x), x\in[0,1]$
CAPTION
u_1
COND_1
c_1
QOI_1
EXAMPLE_1
u_J
COND_J
c_J
QOI_J
EXAMPLE_J
u
QUESTION
one forward pass
w/o weight update
c
PREDICTION

Architecture



attention mask

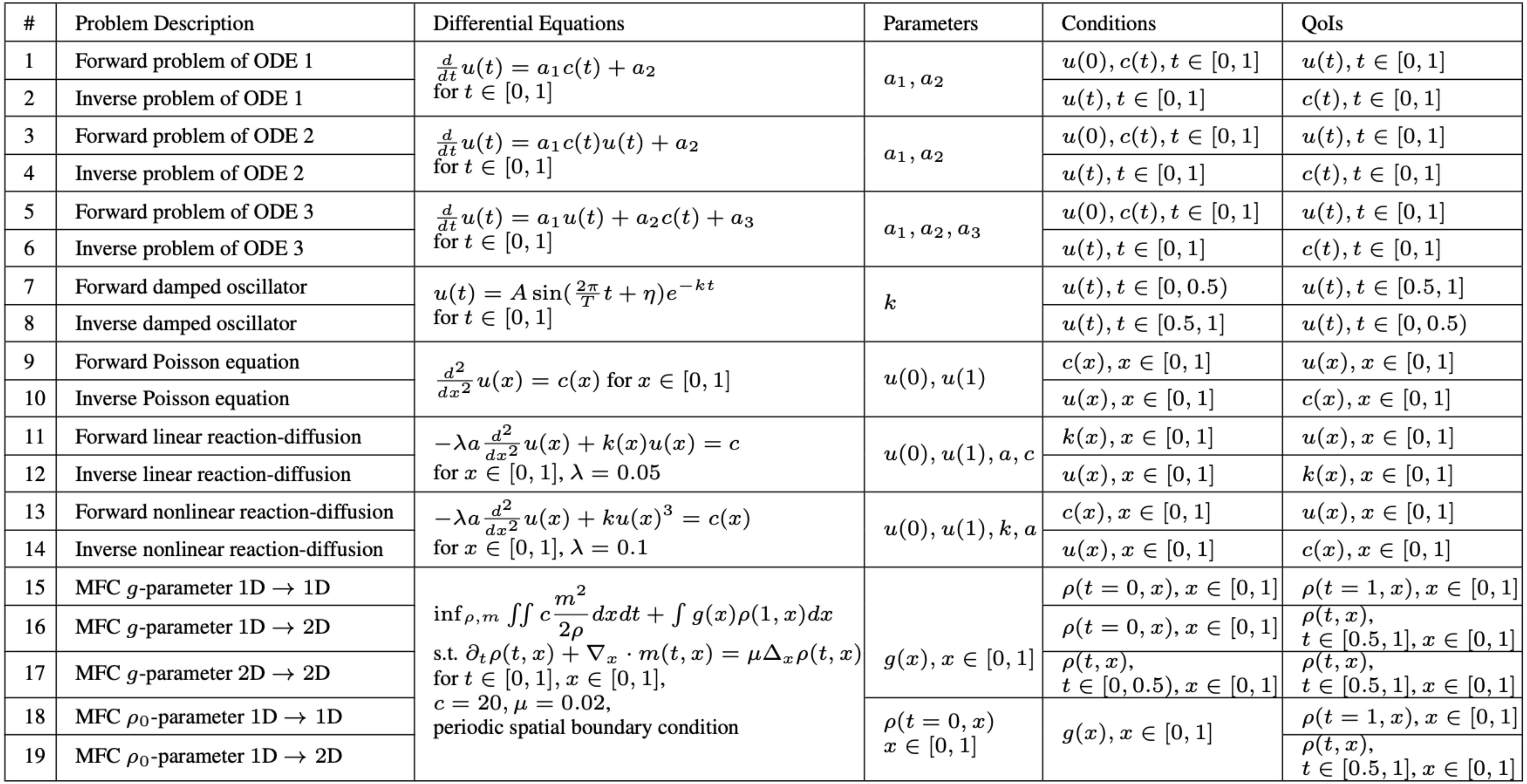


- Training: "Next function prediction", predict the QoI based on the caption, previous examples and the current condition.
- Inference: Flexible number of examples
- Flexible data points in each function
- Evaluate QoI anywhere, in parallel and independently

		condition	QoI	query
key	term	$0 \quad 0 \quad \dots \quad 0 \quad 1$	$0 \quad 0 \quad \dots \quad 0$	$0 \quad 0 \quad \dots \quad 0$
	time	$t_1 \quad t_2 \quad \dots \quad t_{n_j-1} \quad 0$	$\tau_1 \quad \tau_2 \quad \dots \quad \tau_{m_j}$	$\tau_1 \quad \tau_2 \quad \dots \quad \tau_{m_j}$
	space	$0 \quad 0 \quad \dots \quad 0 \quad 0$	$0 \quad 0 \quad \dots \quad 0$	$0 \quad 0 \quad \dots \quad 0$
value		$c(t_1) \quad c(t_2) \quad \dots \quad c(t_{n_j-1}) \quad u(0)$	$u(\tau_1) \quad u(\tau_2) \quad \dots \quad u(\tau_{m_j})$	$0 \quad 0 \quad \dots \quad 0$

Forward ODE problem

$$\frac{d}{dt}u(t) = a_1 c(t) u(t) + a_2$$



#	Problem Description	Differential Equations	Parameters	Conditions	QoIs
1	Forward problem of ODE 1	$\frac{d}{dt}u(t) = a_1 c(t) + a_2$ for $t \in [0,1]$	a_1, a_2	$u(0), c(t), t \in [0,1]$	$u(t), t \in [0,1]$
2	Inverse problem of ODE 1			$u(t), t \in [0,1]$	$c(t), t \in [0,1]$
3	Forward problem of ODE 2	$\frac{d}{dt}u(t) = a_1 c(t)u(t) + a_2$ for $t \in [0,1]$	a_1, a_2	$u(0), c(t), t \in [0,1]$	$u(t), t \in [0,1]$
4	Inverse problem of ODE 2			$u(t), t \in [0,1]$	$c(t), t \in [0,1]$
5	Forward problem of ODE 3	$\frac{d}{dt}u(t) = a_1 u(t) + a_2 c(t) + a_3$ for $t \in [0,1]$	a_1, a_2, a_3	$u(0), c(t), t \in [0,1]$	$u(t), t \in [0,1]$
6	Inverse problem of ODE 3			$u(t), t \in [0,1]$	$c(t), t \in [0,1]$
7	Forward damped oscillator	$u(t) = A\sin(\frac{2\pi}{T}t + \eta)e^{-kt}$ for $t \in [0,1]$	k	$u(t), t \in [0,0.5)$	$u(t), t \in [0.5,1]$
8	Inverse damped oscillator			$u(t), t \in [0.5,1]$	$u(t), t \in [0,0.5)$
9	Forward Poisson equation	$\frac{d^2}{dx^2}u(x) = c(x)$ for $x \in [0,1]$	$u(0), u(1)$	$c(x), x \in [0,1]$	$u(x), x \in [0,1]$
10	Inverse Poisson equation			$u(x), x \in [0,1]$	$c(x), x \in [0,1]$
11	Forward linear reaction-diffusion	$-\lambda a \frac{d^2}{dx^2}u(x) + k(x)u(x) = c$ for $x \in [0,1]$, $\lambda = 0.05$	$u(0), u(1), a, c$	$k(x), x \in [0,1]$	$u(x), x \in [0,1]$
12	Inverse linear reaction-diffusion			$u(x), x \in [0,1]$	$k(x), x \in [0,1]$
13	Forward nonlinear reaction-diffusion	$-\lambda a \frac{d^2}{dx^2}u(x) + ku(x)^3 = c(x)$ for $x \in [0,1]$, $\lambda = 0.1$	$u(0), u(1), k, a$	$c(x), x \in [0,1]$	$u(x), x \in [0,1]$
14	Inverse nonlinear reaction-diffusion			$u(x), x \in [0,1]$	$c(x), x \in [0,1]$
15	MFC g-parameter 1D → 1D	$\inf_{\rho,m} \iint c\frac{m^2}{2\rho}dxdt + \int g(x)\rho(1,x)dx$ s.t. $\partial_t \rho(t,x) + \nabla_x \cdot m(t,x) = \mu\Delta_x \rho(t,x)$ for $t \in [0,1], x \in [0,1]$, $c = 20, \mu = 0.02$, periodic spatial boundary condition	$g(x), x \in [0,1]$	$\rho(t=0,x), x \in [0,1]$	$\rho(t=1,x), x \in [0,1]$
16	MFC g-parameter 1D → 2D			$\rho(t=0,x), x \in [0,1]$	$\rho(t,x)$, $t \in [0.5,1], x \in [0,1]$
17	MFC g-parameter 2D → 2D			$\rho(t,x)$, $t \in [0,0.5), x \in [0,1]$	$\rho(t,x)$, $t \in [0.5,1], x \in [0,1]$
18	MFC ρ_0-parameter 1D → 1D		$\rho(t=0,x)$ $x \in [0,1]$	$g(x), x \in [0,1]$	$\rho(t=1,x), x \in [0,1]$
19	MFC ρ_0-parameter 1D → 2D				$\rho(t,x)$, $t \in [0.5,1], x \in [0,1]$

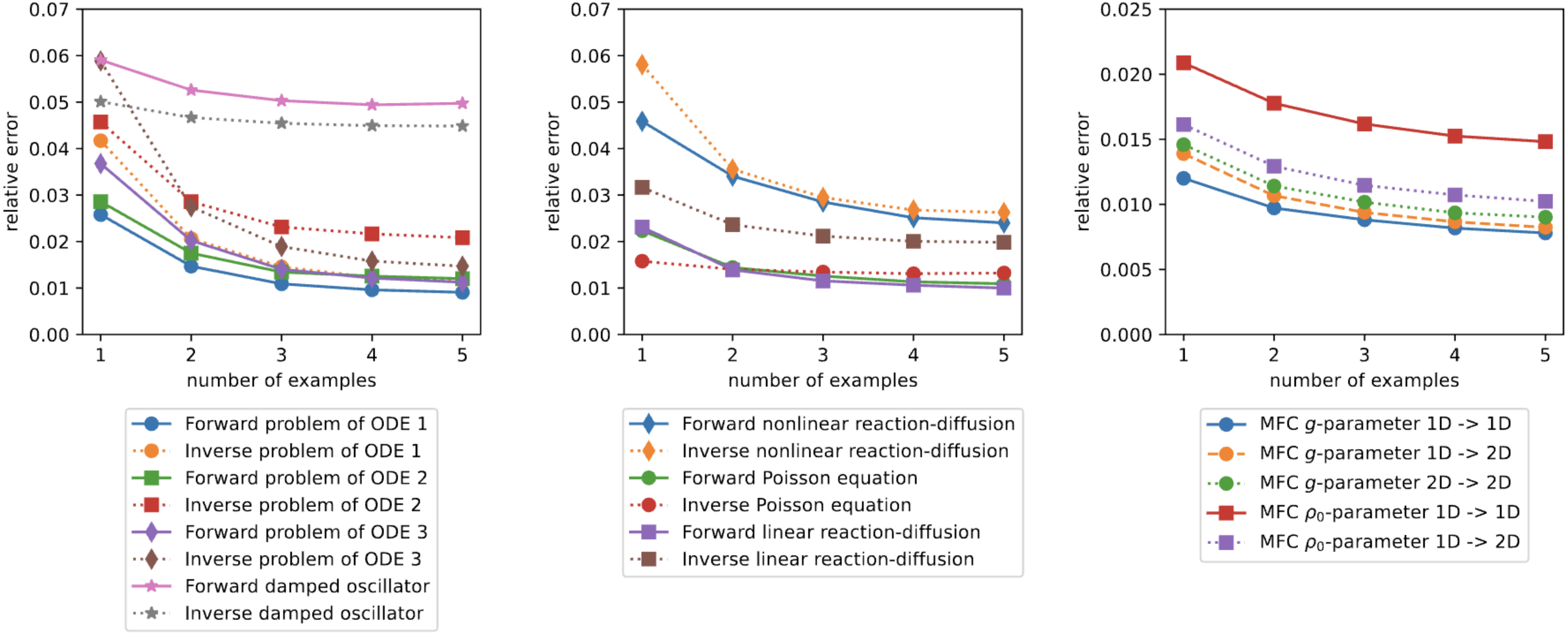
List of the 19 types of problems, including forward and inverse ODE, PDE, and mean-field control problems.

Solved with a single model.

Training: 1000 operators for each problem type, each with 100 examples.

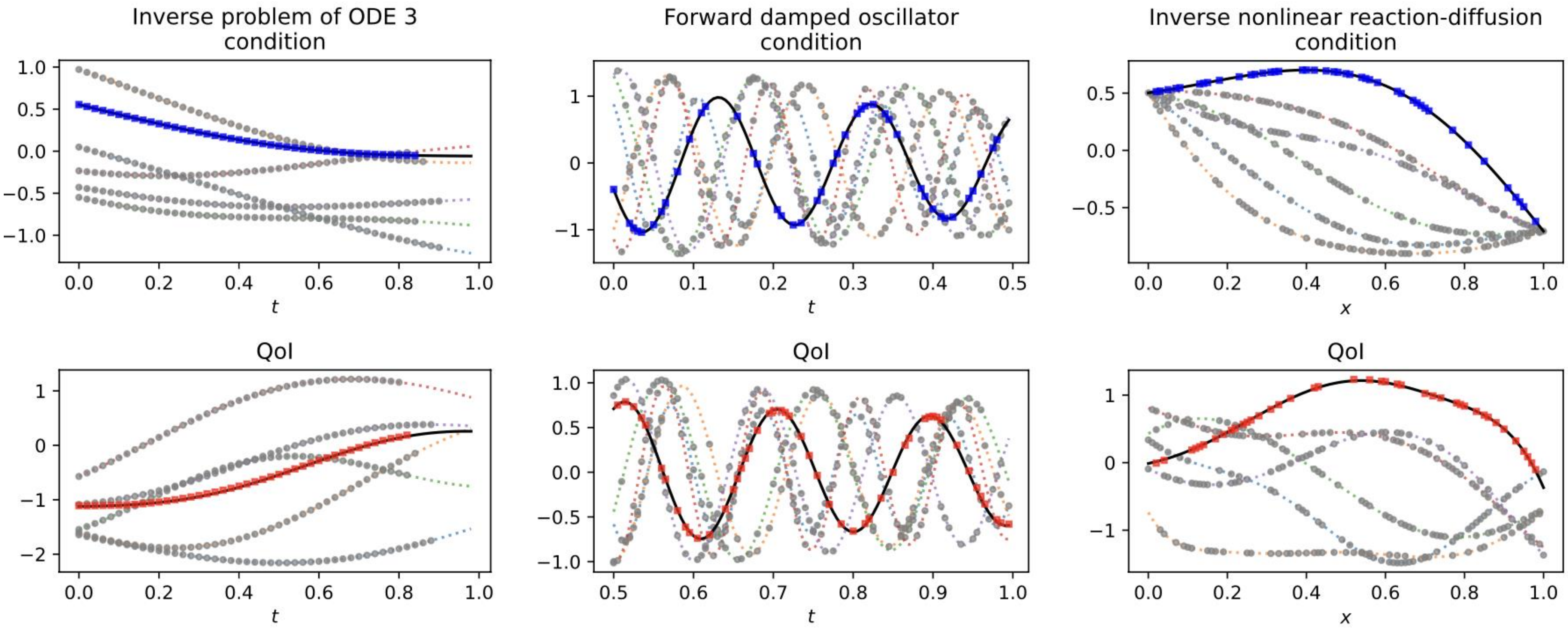
Testing: other operators, only having at most 5 examples in the prompt.

One Model for 19 Types of Problems



With only **five** examples, no captions, the relative error goes down to about **1%-2%** for most cases.

A Glance of ICON for 1D ODE and PDE Problems



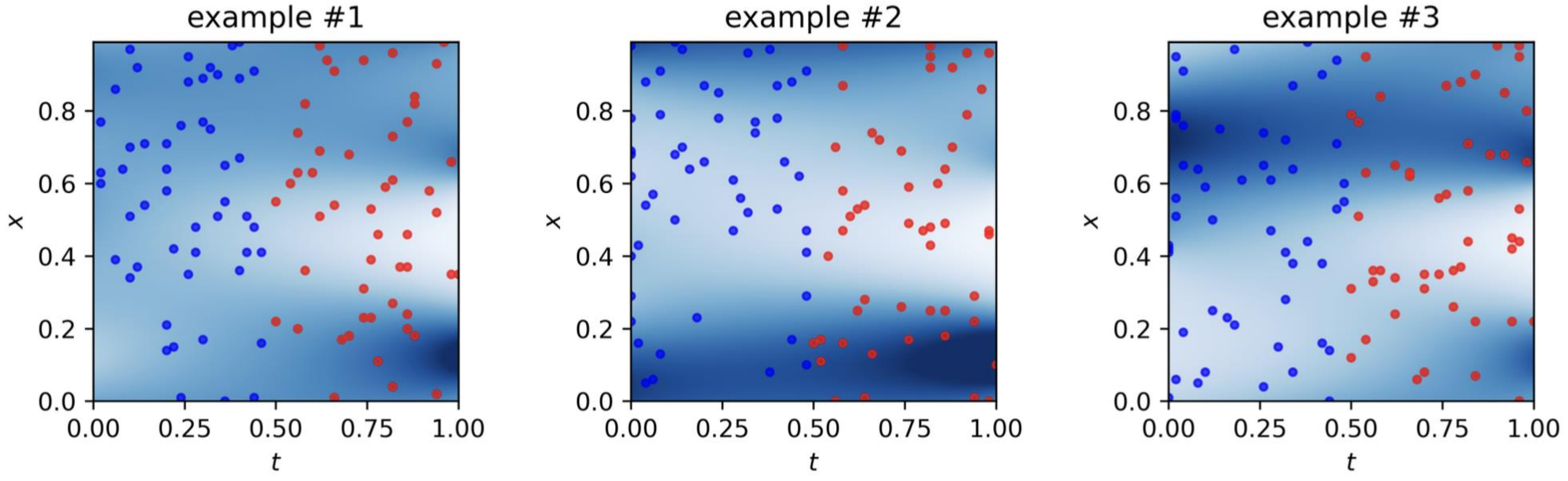
Grey dots: data of the examples in the prompts.

Blue dots: data in the question conditions.

Red dots: prediction of the question QoI.

Solid black lines: ground truth. (overlap with prediction)

Mean-Field Control Problem (Problem #17)



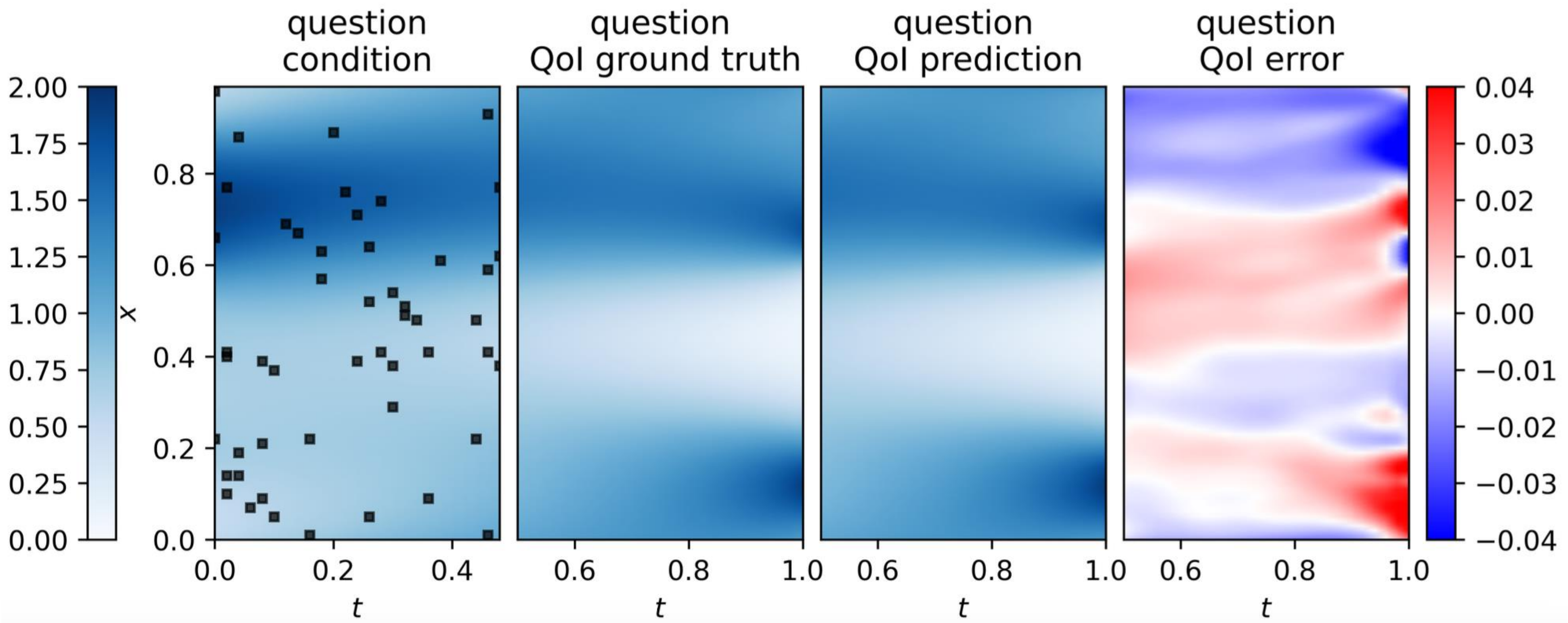
$$\inf_{\rho, m} \iint c\frac{m^2}{2\rho} dxdt + \int g(x)\rho(1,x)dx$$

s.t. $\partial_t \rho(t,x) + \nabla_x \cdot m(t,x) = \mu\Delta_x \rho(t,x)$

for $t \in [0,1], x \in [0,1]$,

$c = 20, \mu = 0.02$,

periodic spatial boundary condition



terminal cost g(x) as the hidden parameter

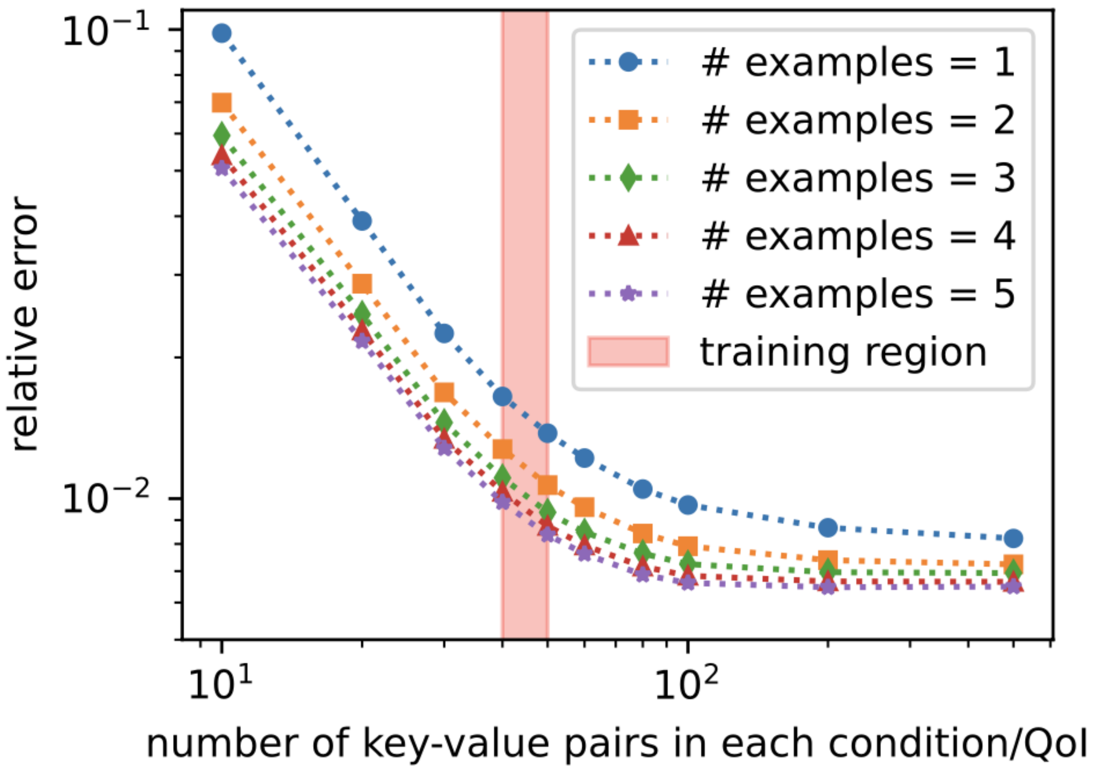
density field in temporal-spatial domain

Blue dots: data for example condition

Red dots: data for example QoI

Black dots: data for question condition

More/Less Data Points (Super/Sub-Resolution)

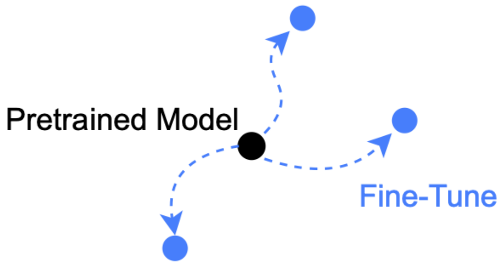


Still the mean-field control problem (Problem #17)

ICON is trained using 41 to 50 data points in each function, represented by the narrow **red region**.

But during inference, the number of data points in each function is rather flexible.

Comparison with Classic Operator Learning

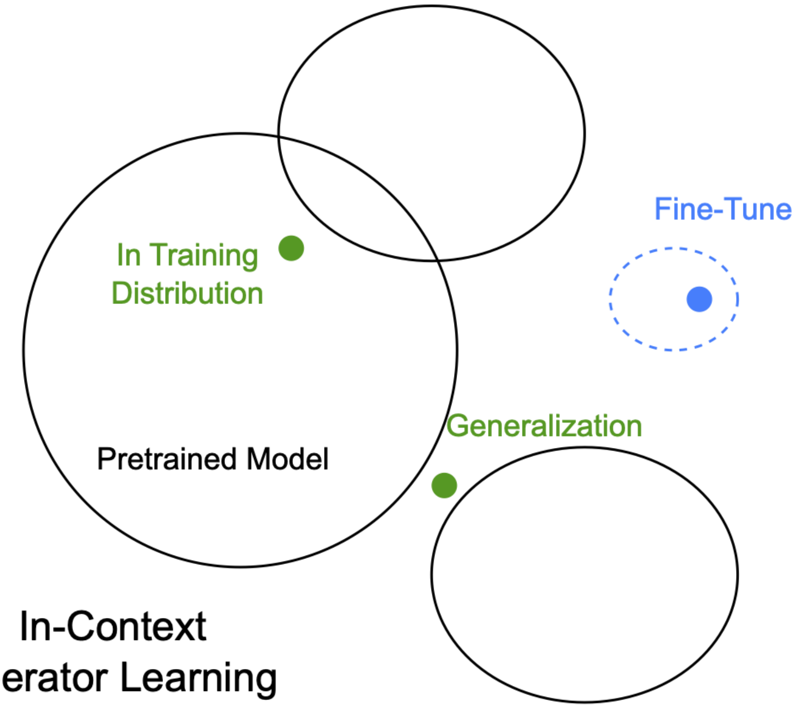


Pretrained Model

Fine-Tune

Operator Learning

Pretrain + Finetune



In Training Distribution

Pretrained Model

Generalization

Fine-Tune

In-Context Operator Learning

Comparison with Classic Operator Learning

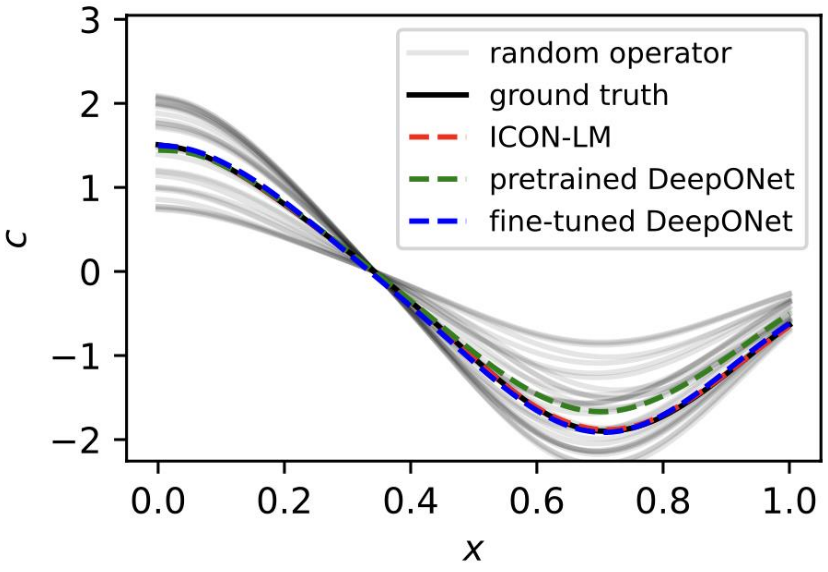
$$-a\frac{d^2}{dx^2}u(x) + ku(x)^3 = c(x),$$
$$a \sim U(0.05, 0.15), k \sim U(0.5, 1.5)$$

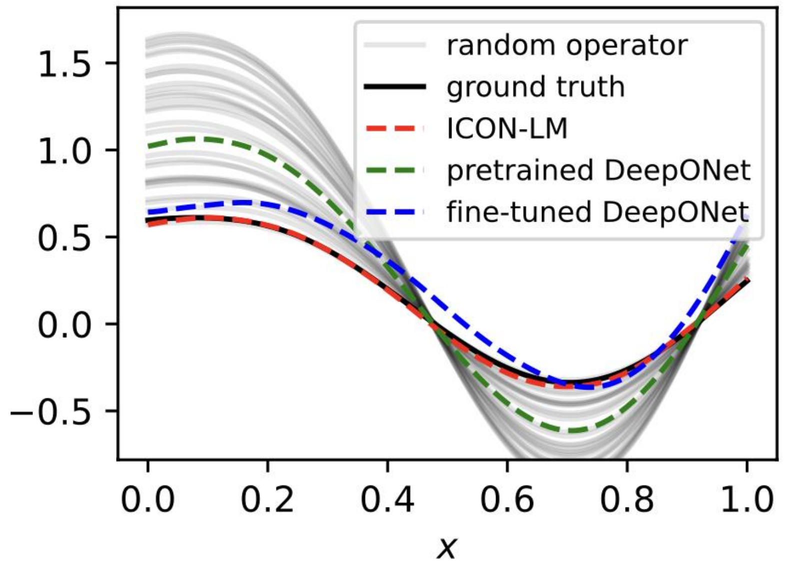
Condition u(x), QoI c(x)

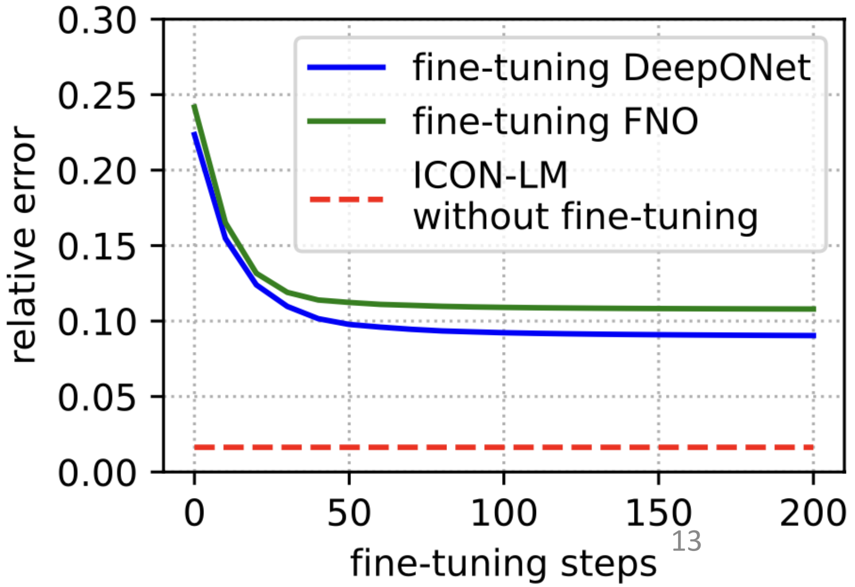
Pretrain the FNO and DeepONet on the distribution of operators -> approximate the mean operator
Fine-tune the pretrained models using **five examples** corresponding to the testing operator.

- Fine-tuning works well when the testing operator is close to the mean operator (left figure)
- Fine-tuning fails when the testing operator is not close to the mean operator (middle figure)

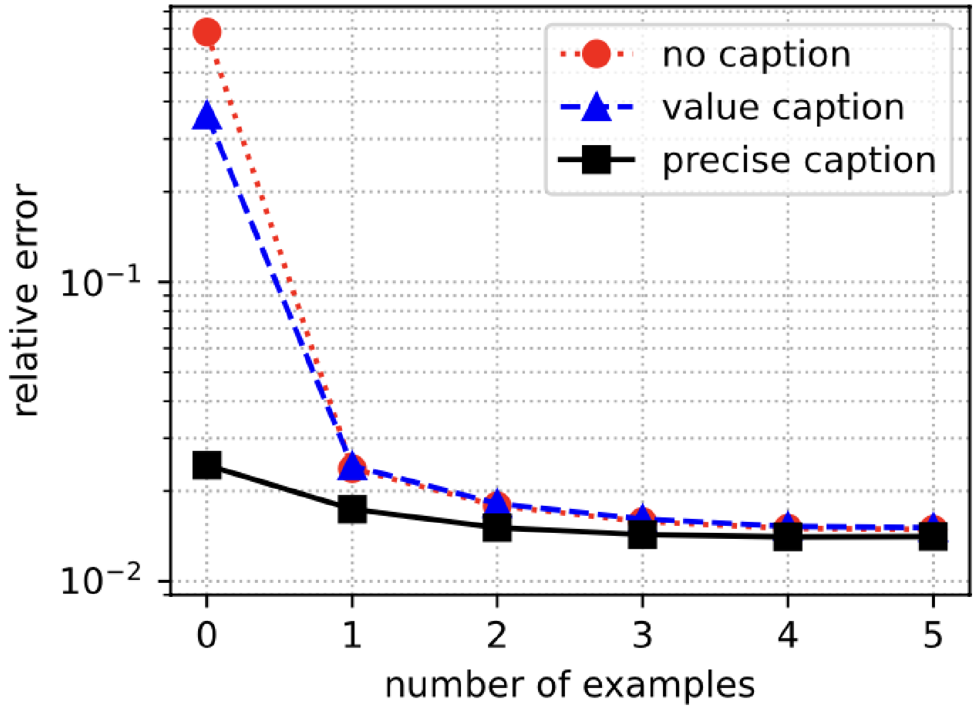
ICON consistently outperforms classic operator learning, even without finetuning.







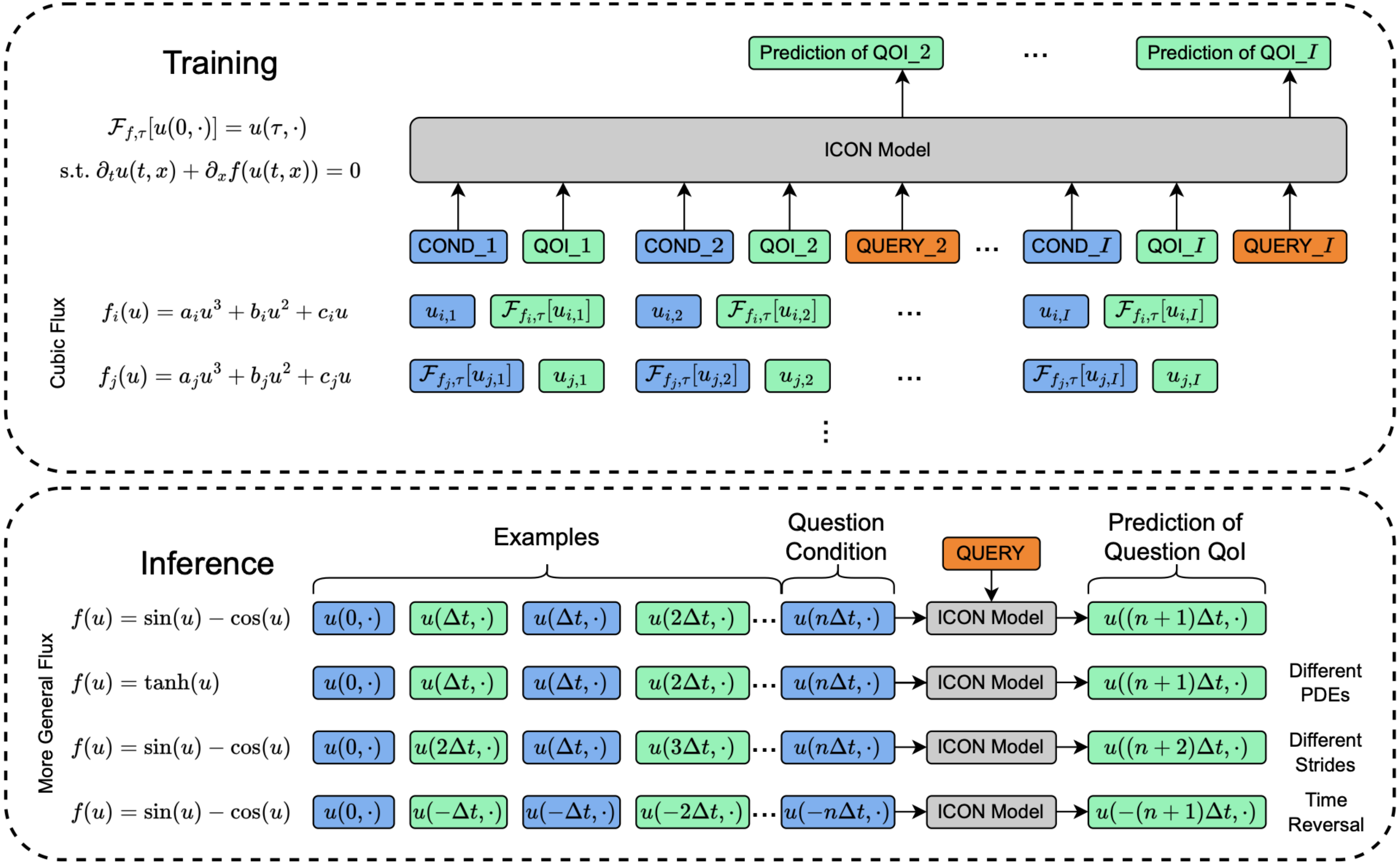
Fine-Tune GPT-2 for Multi-Modal Learning



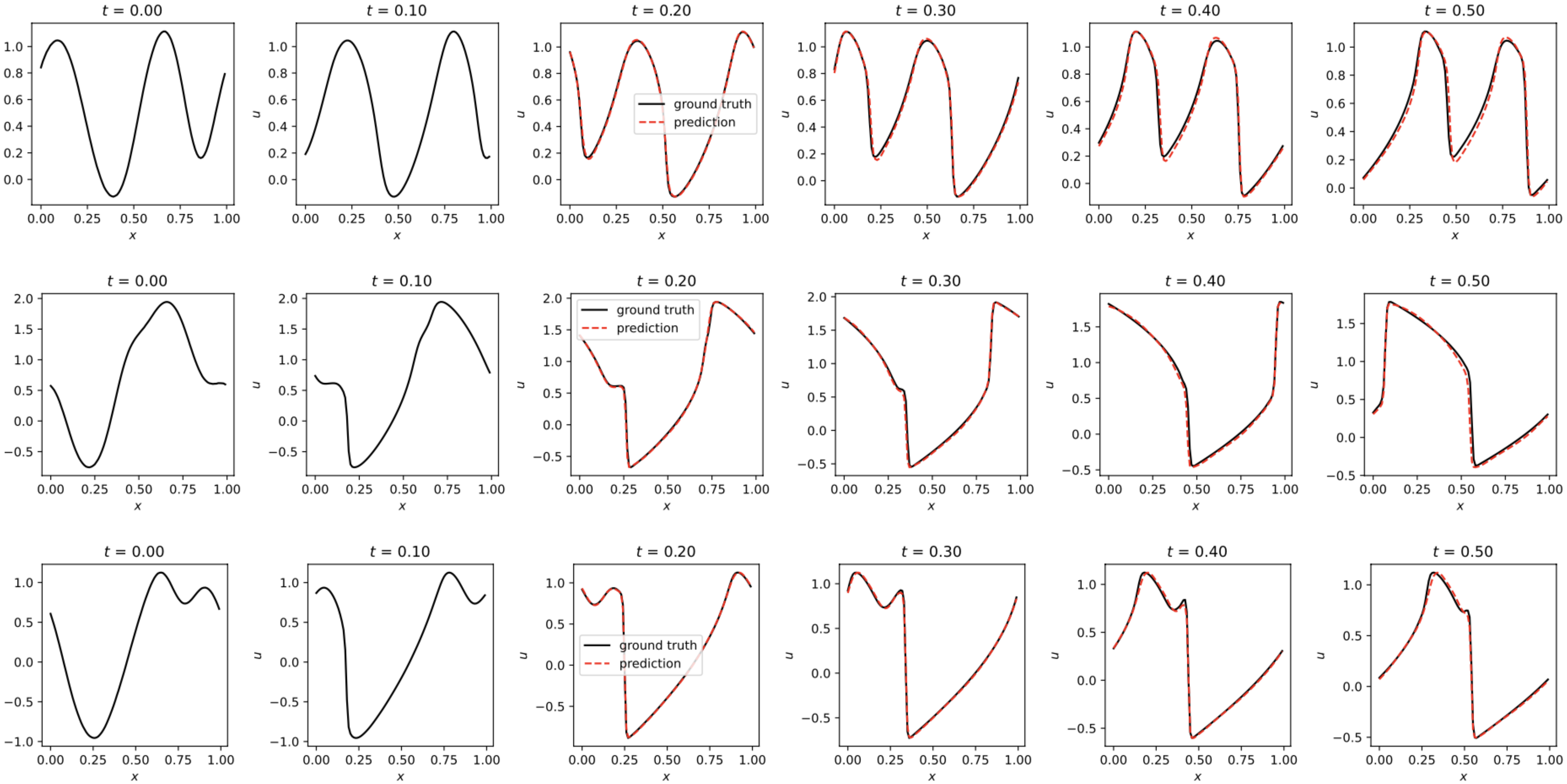
Vague caption (without numbers): The rate of change of \$u(t)\$ over time is given by the equation \$du(t)/dt = a_1 \cdot u(t) + a_2 \cdot c(t) + a_3\$. Condition: \$u(0)\$ and \$c(t), t\in[0,1]\$, QoI: \$u(t), t\in[0,1]\$

Precise caption (with numbers): The relationship between \$u(t)\$ and \$c(t)\$ is governed by the equation \$du(t)/dt = **0.48** \cdot u(t) + **1.06** \cdot c(t) + **0.691**\$. Condition: \$u(0)\$ and \$c(t), t\in[0,1]\$, QoI: \$u(t), t\in [0,1]\$

PDE Prediction with ICON (Conservation Laws)

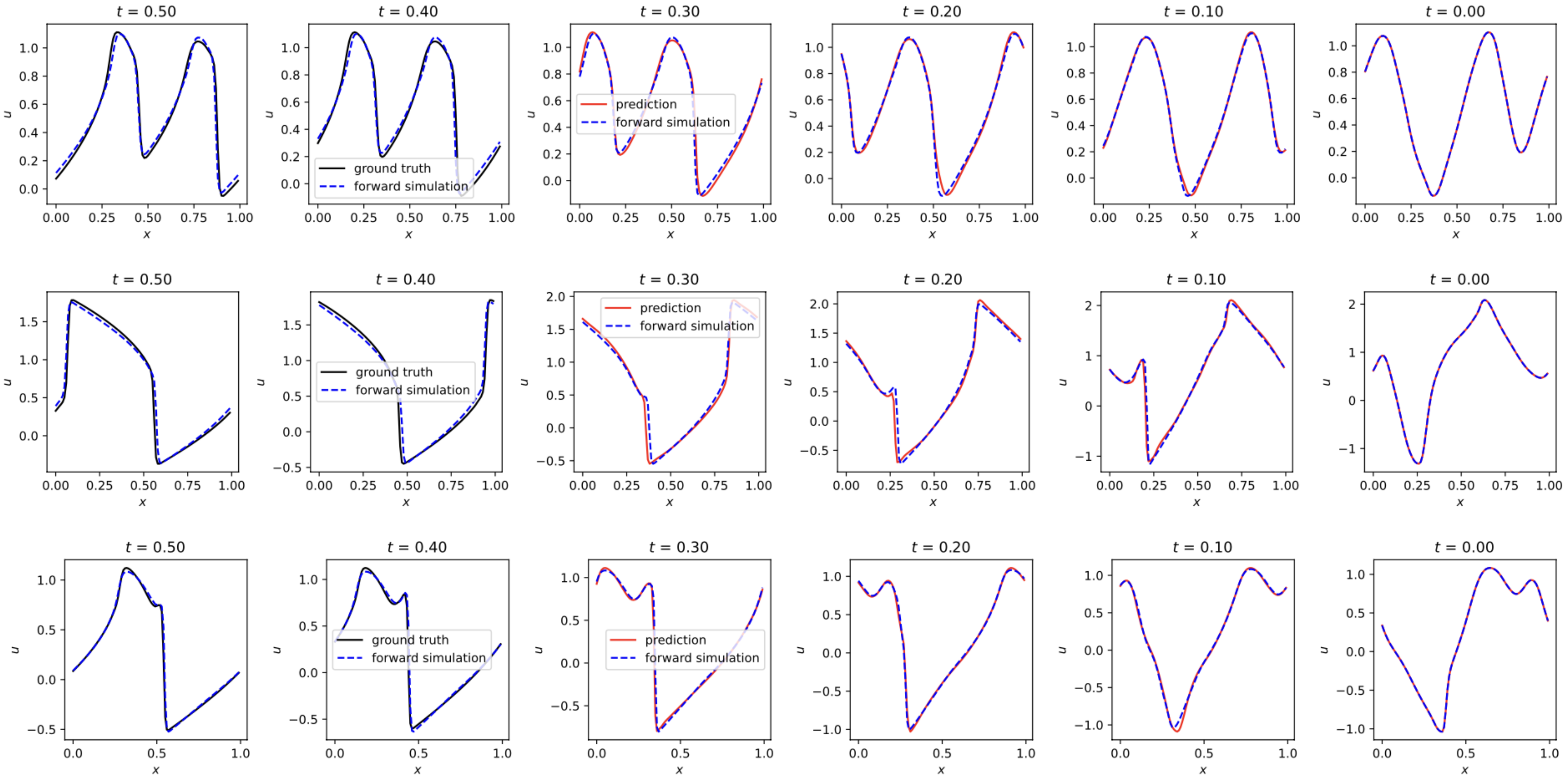


Testing on New PDE (Forward)



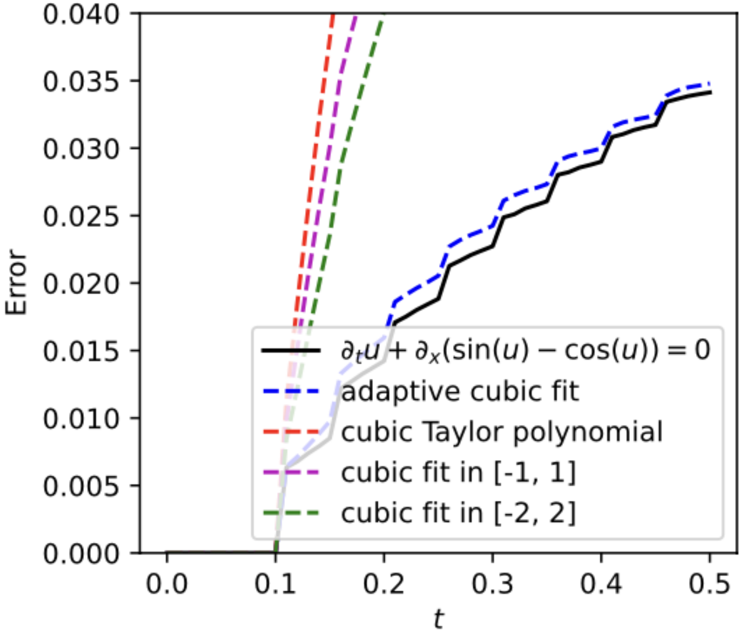
$$\partial_t u + \partial_x(\sin(u) - \cos(u)) = 0$$

Testing on New PDE (Reverse)

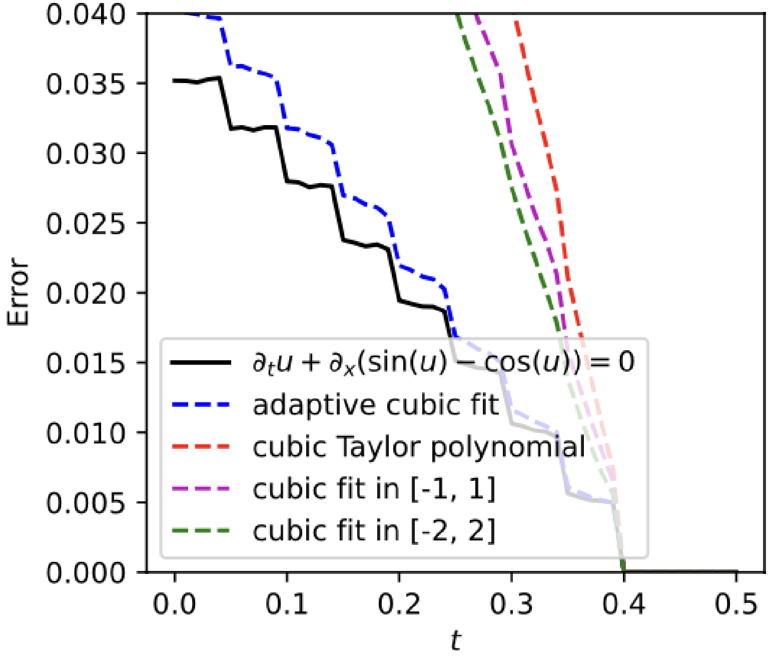


$$\partial_t u + \partial_x(\sin(u) - \cos(u)) = 0$$

A Closer Look at Generalization



Forward



Reverse

Did the ICON model simply memorize the cubic flux functions and approximate the new flux with the closest cubic function? We make predictions with examples coming from f(u) = sin(u)-cos(u) and "similar cubic functions", and compare the errors between the predictions and the ground truth corresponding to f = sin(u)-cos(u).

- The cubic Taylor polynomial
- The best cubic fit in [-1,1] or [-2,2]
- The best cubic fit in [u_min,u_max] (adaptive cubic fit)

ICON and Conditional Generative Modeling

Markos Katsoulakis, Benjamin Zhang

- Given data points $(x_i, y_i) \sim p(x, y)$
  - Create a generative model for the joint distribution $q_\theta(x, y) \approx p(x, y)$
  - Learn how to generate from the **conditional** distribution: $q_\theta(y|x)$
  - ICON is Learning the joint obviates the need to re-train: just sample/generate instead
- Given a choice of probability divergence/metric, find best model: solve variational problem

$$\min_\theta D(p(\cdot), q_\theta(\cdot))$$

- For example, given the KL divergence, variational problem equivalent to

$$\min_\theta \mathbb{E}_{p(x,y)} \left[\log \frac{p(y|x)p(x)}{q_\theta(y|x)p(x)} \right] \iff \max_\theta \mathbb{E}_{p(x,y)} \left[\log q_\theta(y|x) \right]$$

- If In LLM-type implementations, sampling means finding the MAP point:

$$f(x_i) = \arg\max_y q_\theta(y|X = x_i) \text{ equivalent to conditional mean if } q_\theta(y|x) \text{ is Gaussian}$$

Summary

- Drawing inspiration from LLMs, we proposed In-Context Operator Learning and In-Context Operator Networks (ICON).
- The model can learn and apply operator in the forward propagation, instead of approximating a specific operator.
- A single ICON model can handle a wide range of scientific learning problems.
- ICON showed advantage compared with classic operator learning (pretrain + finetune).
- ICON showed generalization to new PDEs, not just memorizing the training PDEs.
- Multi-modal ICON provides a different approach for physics-informed models.

Reference

- Liu Yang, Siting Liu, Tingwei Meng, and Stanley J. Osher. "In-Context Operator Learning With Data Prompts for Differential Equation Problems" Proceedings of the National Academy of Sciences 120.39 (2023): e2310142120.
- Liu Yang, Siting Liu, and Stanley J. Osher. "Fine-Tune Language Models as Multi-Modal Differential Equation Solvers'' arXiv:2308.05061 (2023).
- Liu Yang, and Stanley J. Osher. "PDE Generalization of In-Context Operator Networks: A Study on 1D Scalar Nonlinear Conservation Laws'' arXiv:2401.07364 (2024).