

VICON:
Vision In-Context Learning for Multi-Physics Fluid Dynamics

Yadi Cao
Assistant Professor, Computer Science
University of Central Florida

Successful Scaling of General AI

Strong AI models

 Claude



 Gemini

 Qwen

Deploy
→

Accelerate

Generate
←

Train

AI assisting tools



Copilot



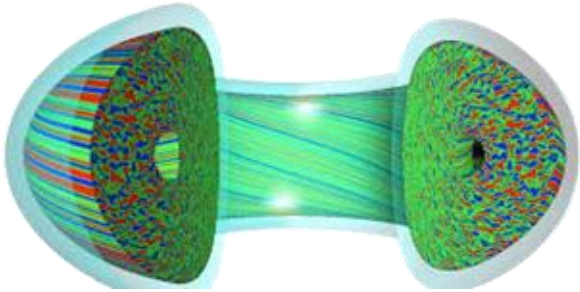
Antigravity



Challenges of Scaling AI 4 Sci/Eng

Expensive experiments/simulations for downstream

6-D gyrokinetic simulation in Tokamak



Princeton Plasma Lab

- 2 million samples
- 2 GPU hrs/sample

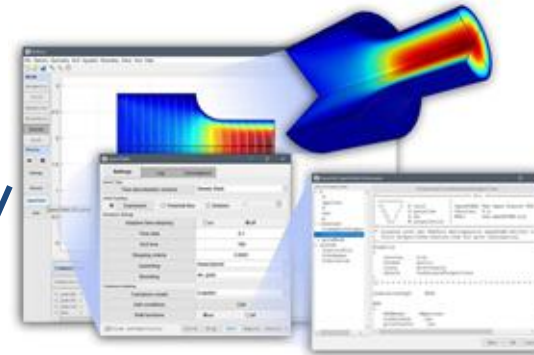
Long hrs

Job sent



Credit: Nova HPC

Expensive expert labor



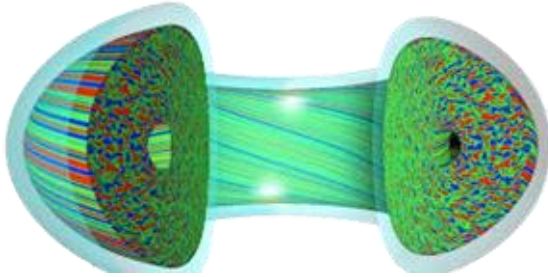
Years of training



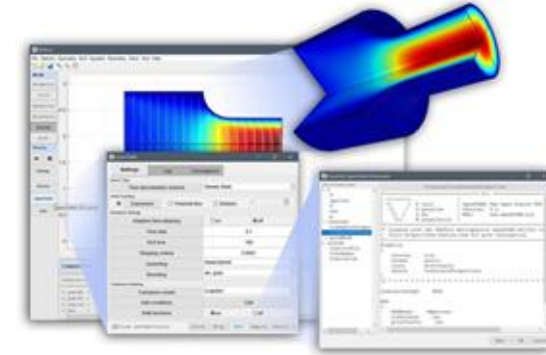
If above solved:

- Scaled, automatic explorations
- Accelerated discovery & deployment

My Lab's Directions



How to make experiments affordable?



How to reduce human labor?

Realistic Digital Twins

Respect data cost

- AI for Fusion [1]



Deploy



Accelerate

General geometry

- BSMS-GNN [2]



Multi-physics/condition

- VICON [3]



Today

Realistic AI Engineers

Automated workflows

- FoamAgent [4]



Benchmarks that matter

- SimulCost [5]



Transferable sol

- Adapting-While-Learning [6]

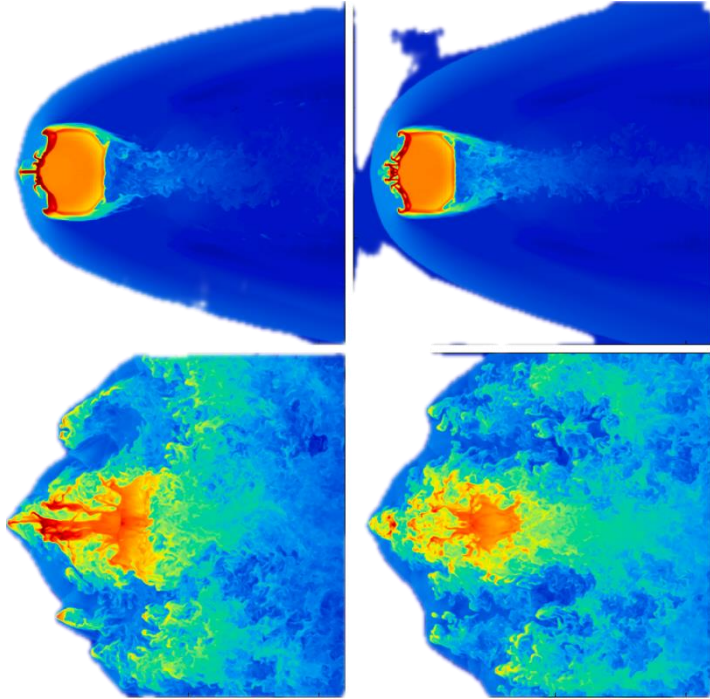


Generate

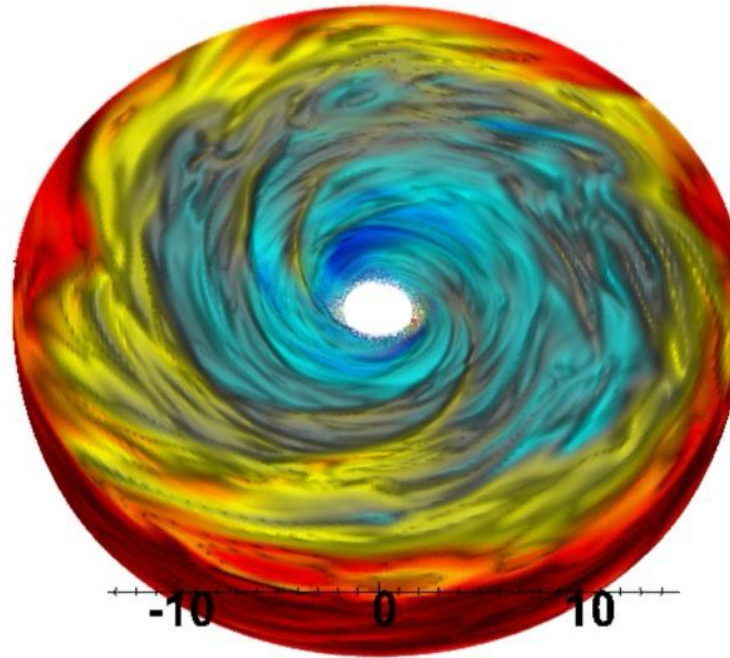


Train

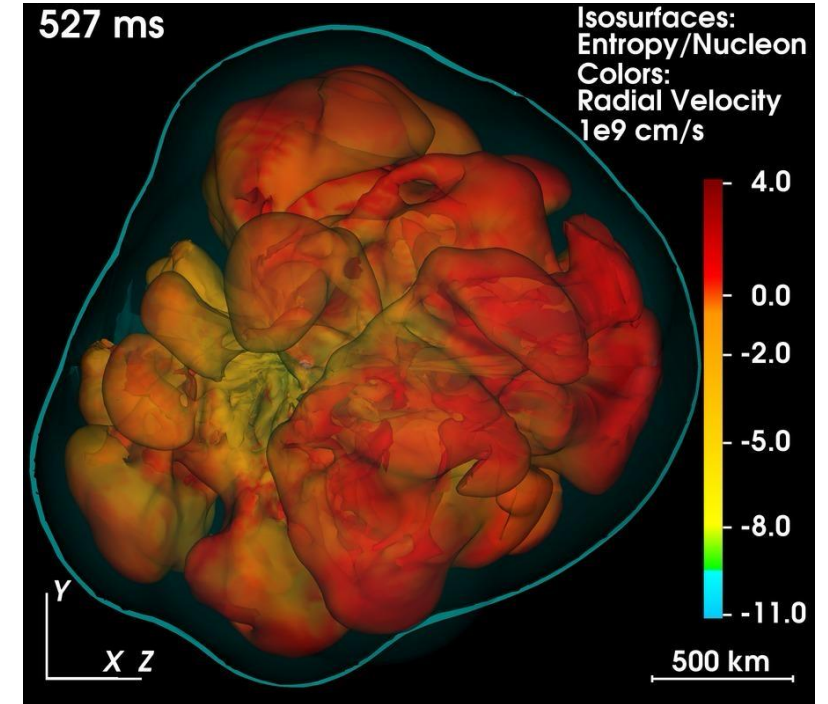
Different Fluid Like Systems in Astrophysics



[HD \[NASA Cygnus Loop\]](#)



[MHD \[MPA / MHD turbulence\]](#)



[W/ General Relativity \[Bernuzzi/Dietrich et al.\]](#)

Digital Twin for Multi-Physics

Train for each system:

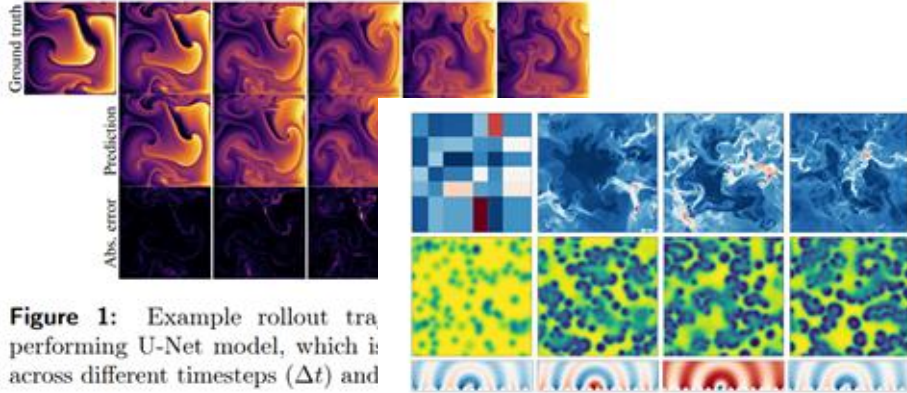


Figure 1: Example rollout trajectory performing U-Net model, which is across different timesteps (Δt) and

[PDEArena](#)

[Wells Multi-Physics Benchmark](#)

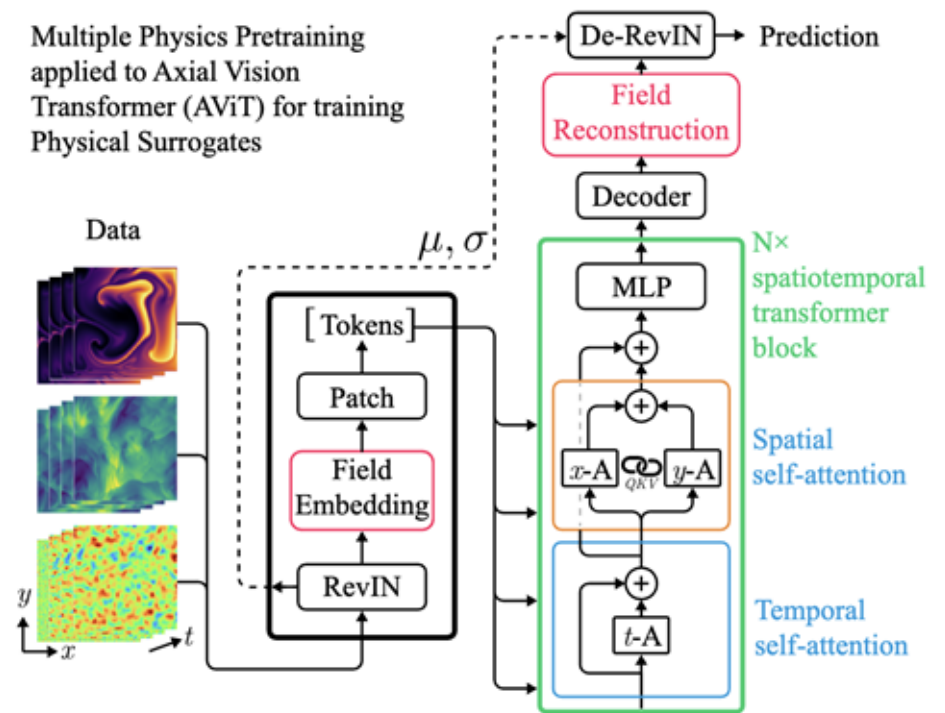
Train for this

Train another

1 model for 1 physics + 1 working condition

No Data Efficiency

Train with 1 seq2seq model:



[MPP \[McCabe et al, 2024\]](#)

Finetuning needed for each new deployment
Where is the data?

No Flexibility for online deployment

Problem Setup: One Model, Many Systems

Given: one sequence, unknown system

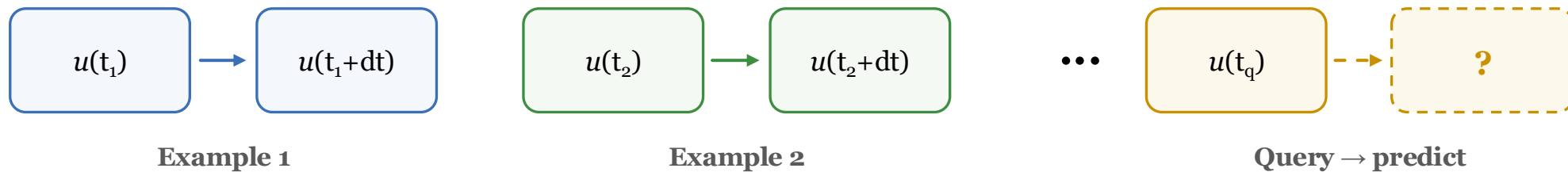
- Frames $u(t_1), \dots, u(t_n)$: uneven, some missing
- Physics & working condition: **unknown**
- Stride $dt \rightarrow$ pairs $(u(t_i), u(t_i+dt))$

Wanted: one model, no retraining

$$\hat{u}(t_q+dt) = G(\{(u(t_i), u(t_i+dt))\}_{i=1..N}, u(t_q))$$

- One set of weights, *all* systems
- Adapt **in the forward pass** — no gradient step
- dt picked at inference, not at training

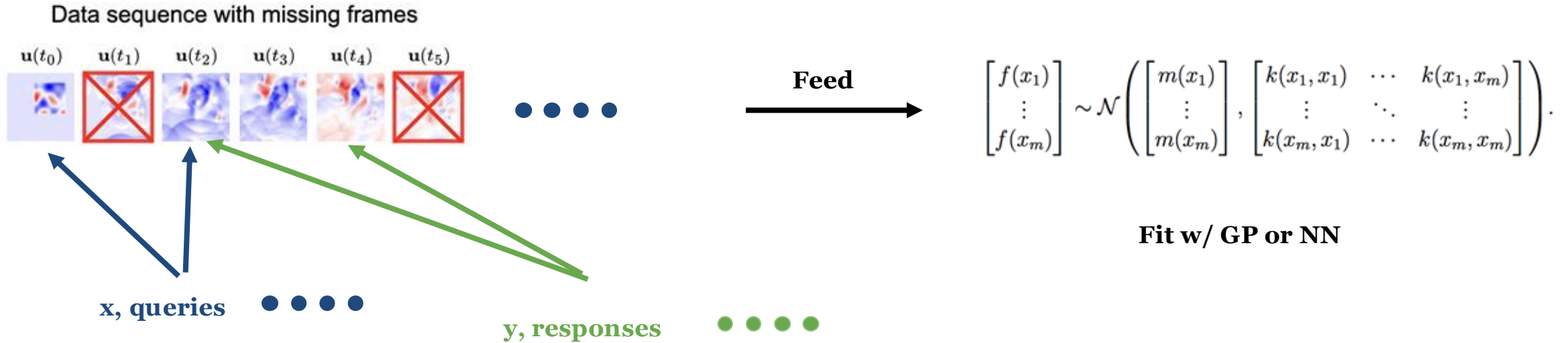
The prompt: N example pairs + 1 query frame



Closes both gaps: 1 training run for all systems + online adaptation, no finetuning

Simple Alternative to Seq2Seq Models

One Markovian system, but
1,5,8 frames missing



Form pairs w/ $dt=2$:

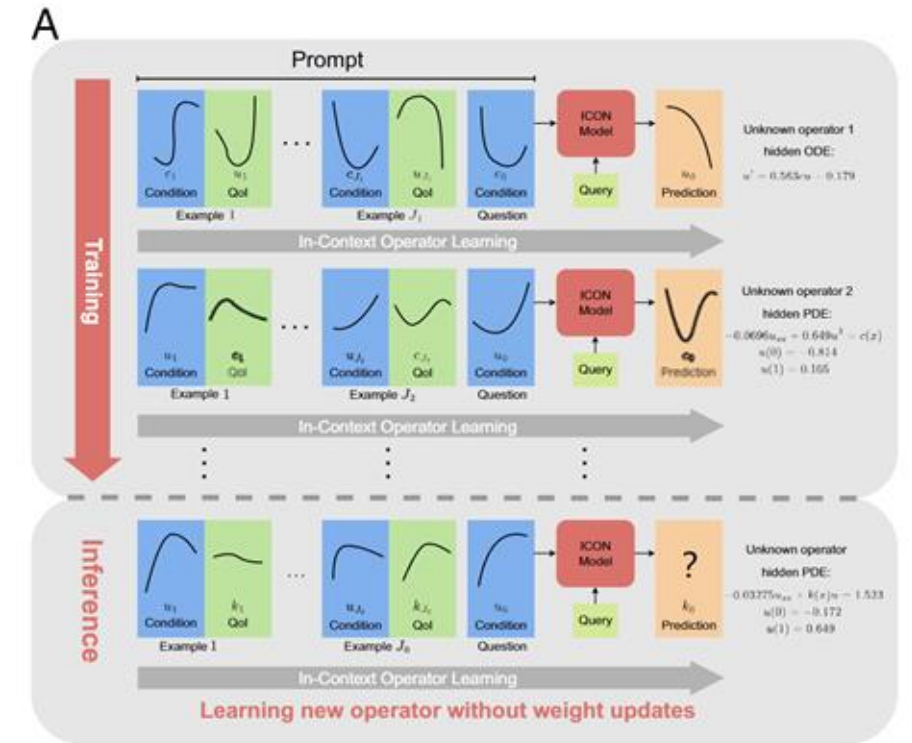
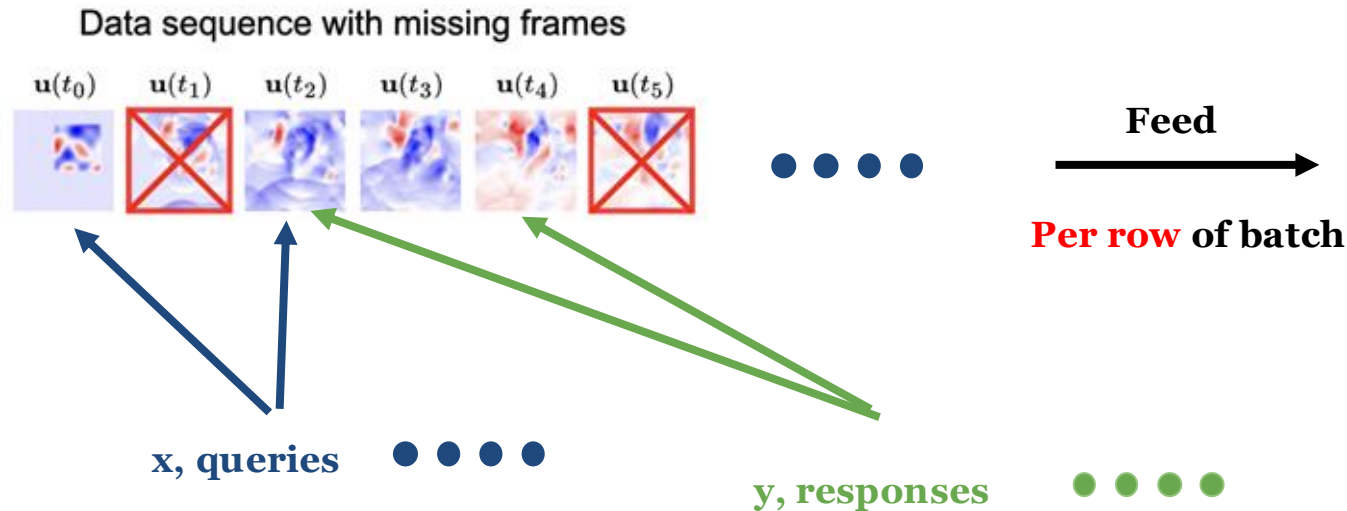
- (0→2),(2→4),(4→6),(7→9)
- States before/after an operator ($dt = 2$)

In-Context Operator Neural Network

Many Markovian systems

Pairs w/ **different** dts

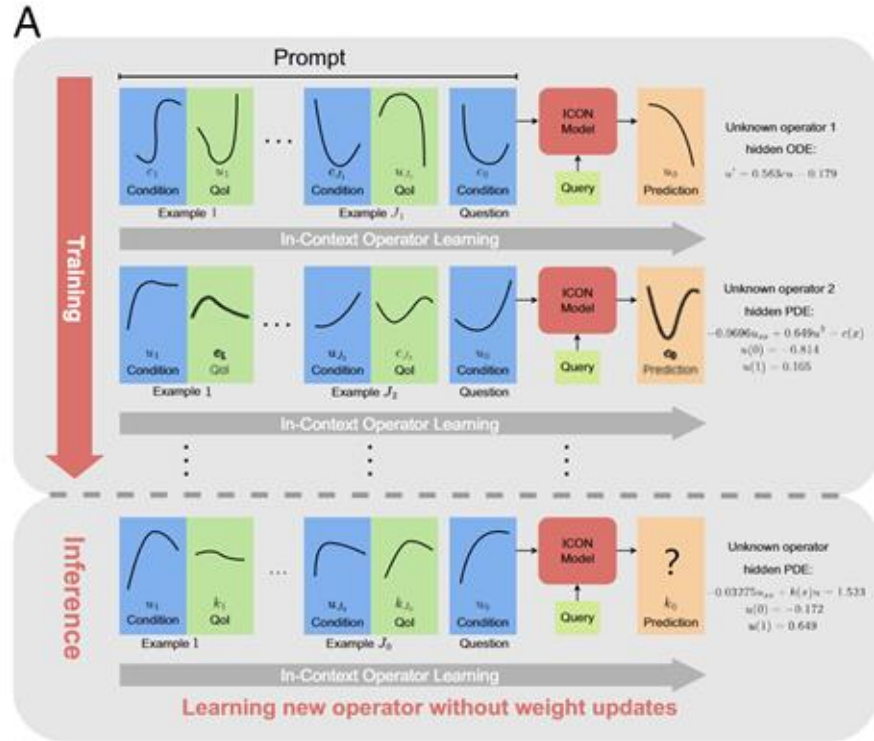
- Choose a (system, dt) combo
- Eg, (NS w/ Re=100, dt = 2)
- (0->2),(2->4),(4->6),(7->9)



[ICON \[Yang et al 2024\]](#)

Fit w/ transformer

Curse of Dimensionality -> Sol: VICON [3]

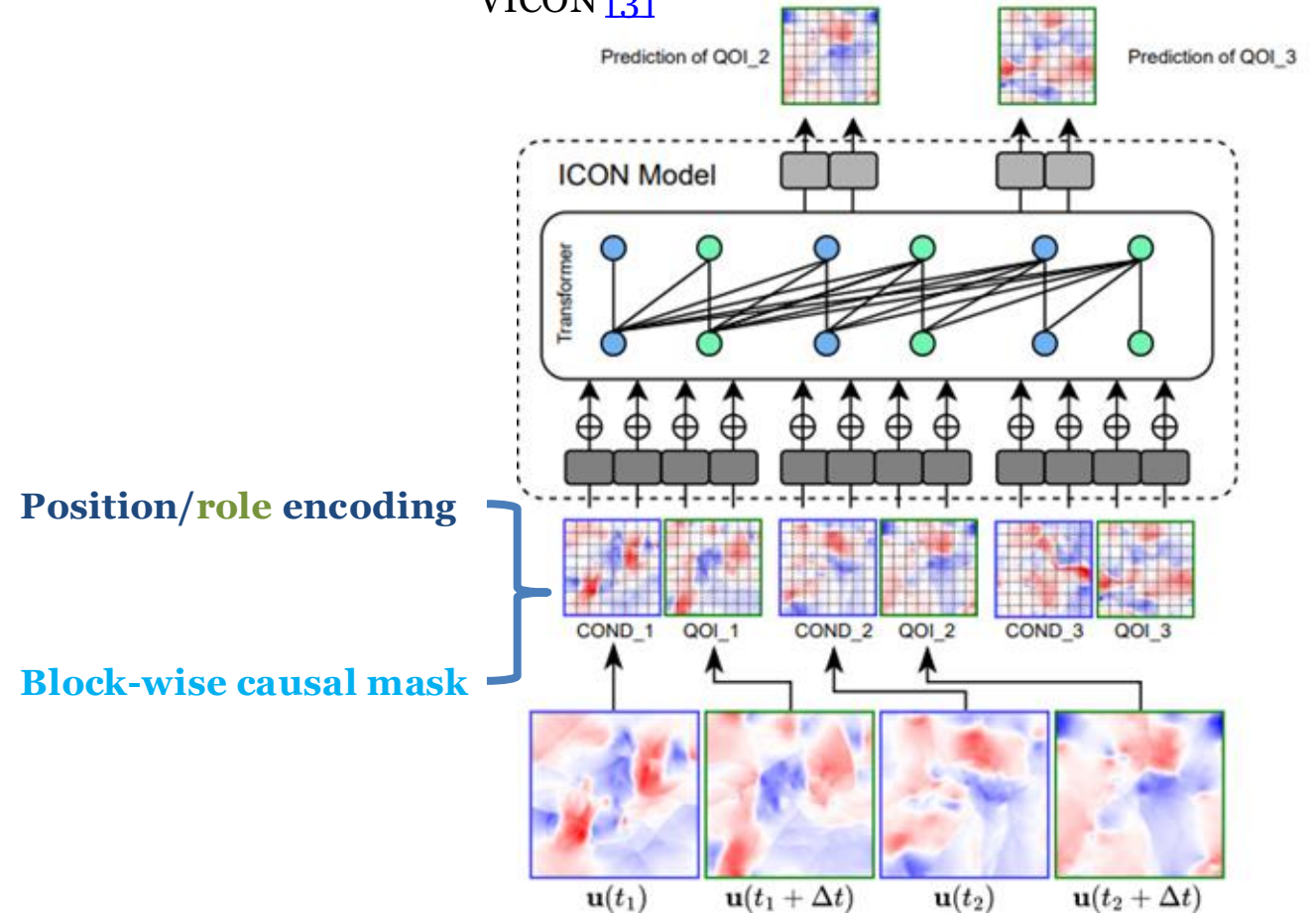


ICON [Yang et al 2024]

ICON mainly works in 1D:

- 1 point/pixel = 1 token
- In 2D, 10 pairs of (x/y) have:
 - $2 * 10 * 128^2 = 3.2E5$

VICON [3]



Patchify, insights from ViT:

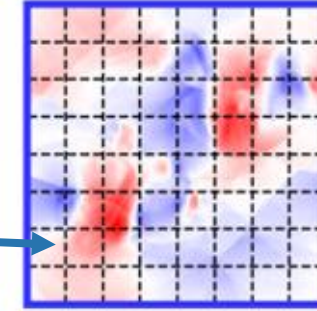
- 1 patch = 1 token, p.s. = 16
 - $2 * 10 * (128/16)^2 = 1280$

Positional & Functional Encodings

Positional:

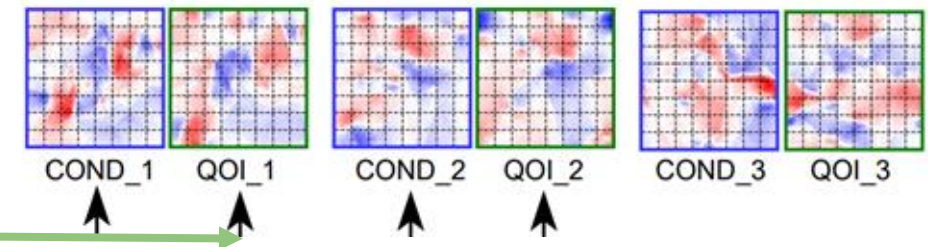
- Relative spatial locations
- eg, (1st, 2nd) patch in (x, y) direction

Assign to all (1,2)s



Functional:

- which example pair? **query** or **response**?
- eg, 1st pair's **response (QOI_1)**

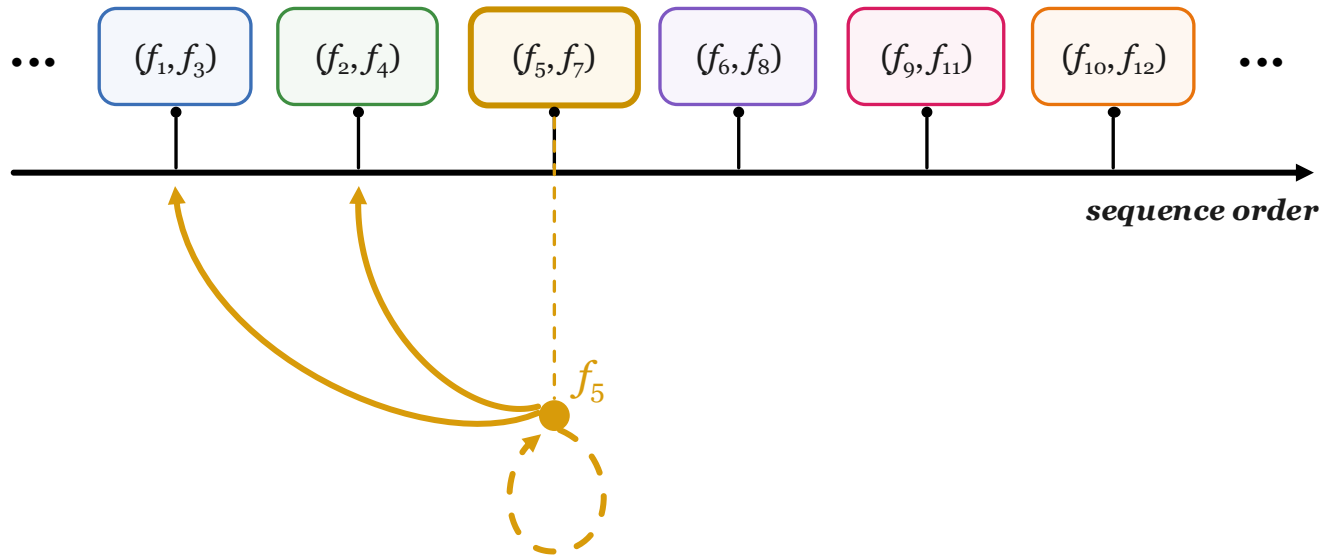


Repeat to all patches in **QOI_1**

Block-wise Diagonal Causal Mask

A prediction only “sees”:

- Previous example pairs, and
- the current input “frame”



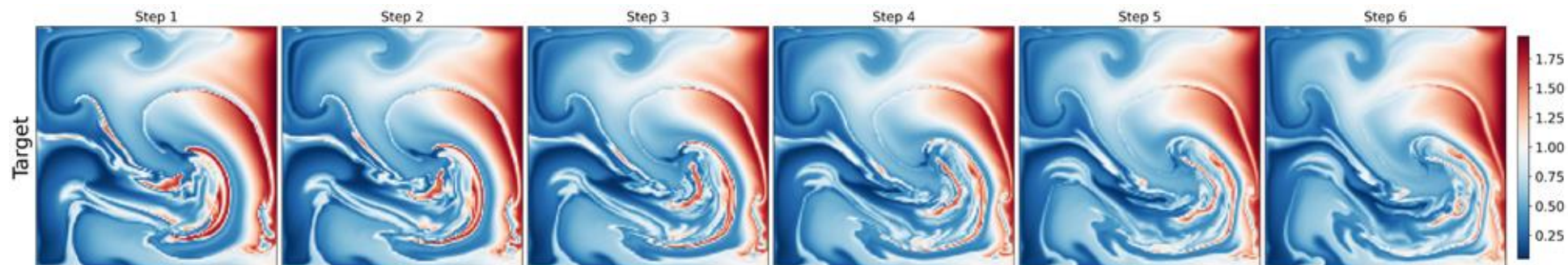
In matrix form:

$$\mathbf{M} = \begin{bmatrix} \mathbf{1}_{N_c, N_c} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{N_q, N_c} & \mathbf{1}_{N_q, N_q} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_{N_c, N_c} & \mathbf{1}_{N_c, N_q} & \cdots & \mathbf{1}_{N_c, N_c} & \mathbf{0} \\ \mathbf{1}_{N_q, N_c} & \mathbf{1}_{N_q, N_q} & \cdots & \mathbf{1}_{N_q, N_c} & \mathbf{1}_{N_q, N_q} \end{bmatrix} \quad (6)$$

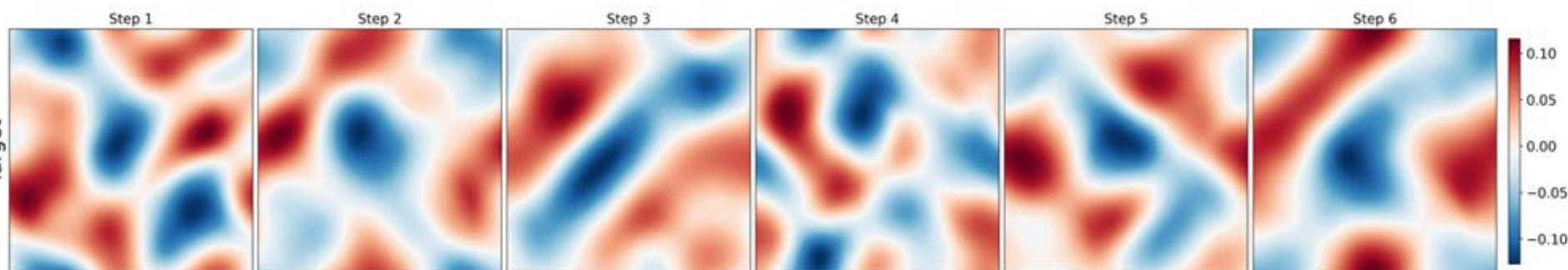
ie, a block-wise diag matrix

Benchmarks

Incompressible NS

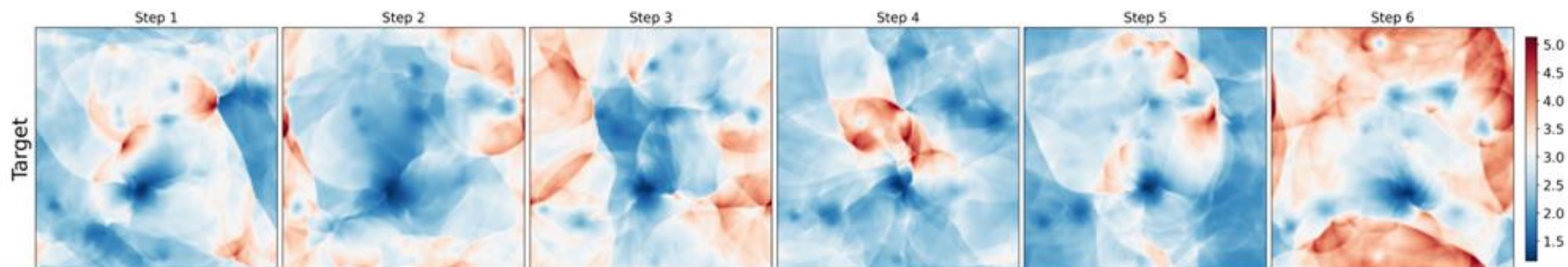


Target

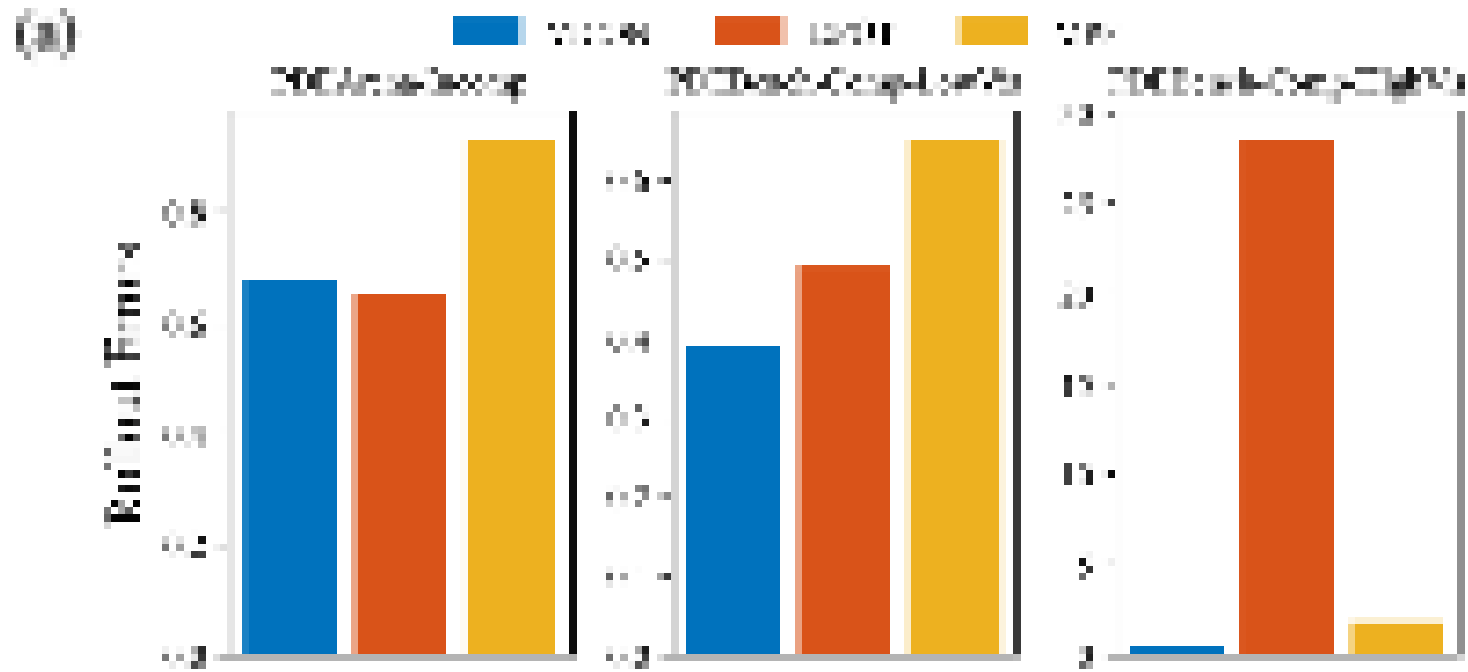


Compressible NS
1e-2 viscosity

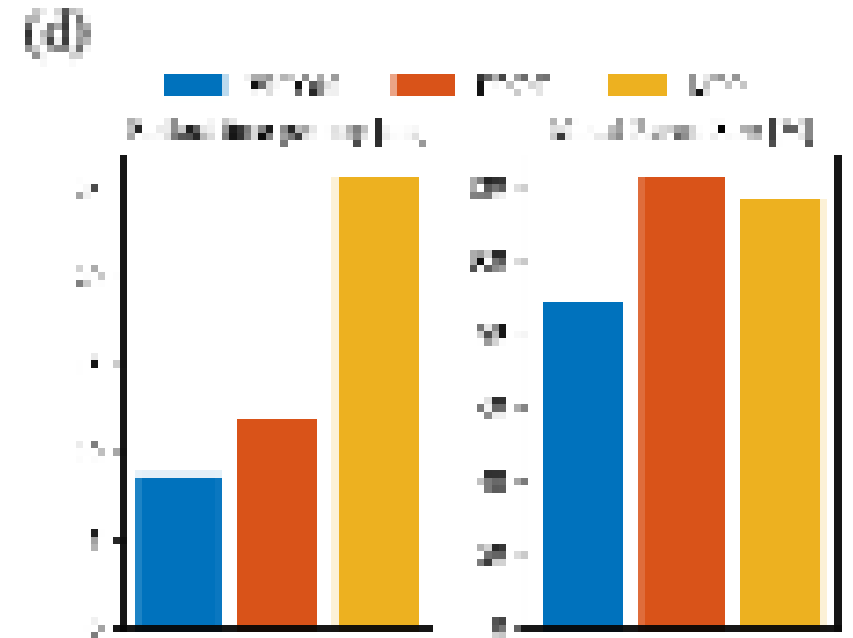
Compressible NS
1e-8 viscosity



SOTA Performance w/out Finetuning



Smallest error on all but 1 subset



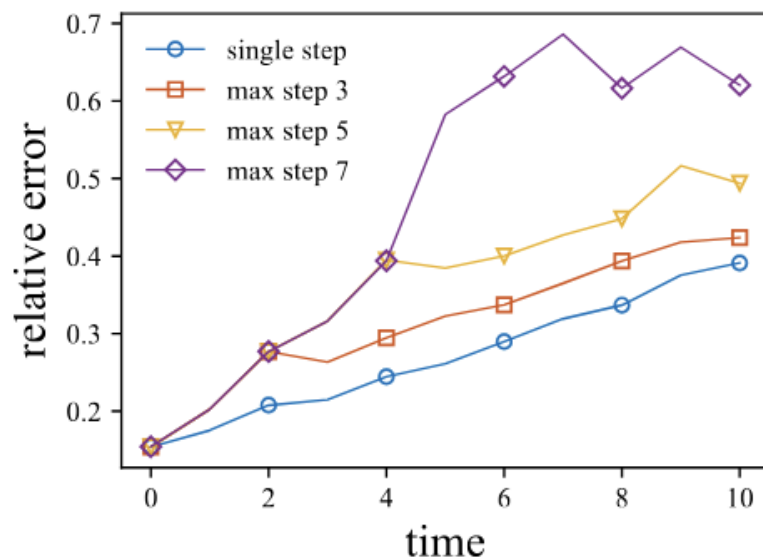
Fair parameter counts and size

Flexible Rollout Strategy

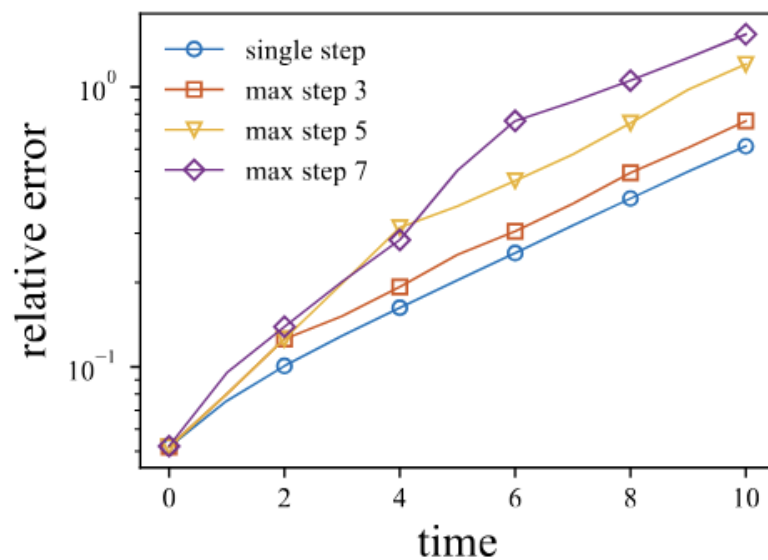
Trade-offs for auto-regressive forecasting:

- Smaller Δt , simpler prediction, but more steps (error accumulation)
- Larger Δt , fewer steps, but each forecast harder

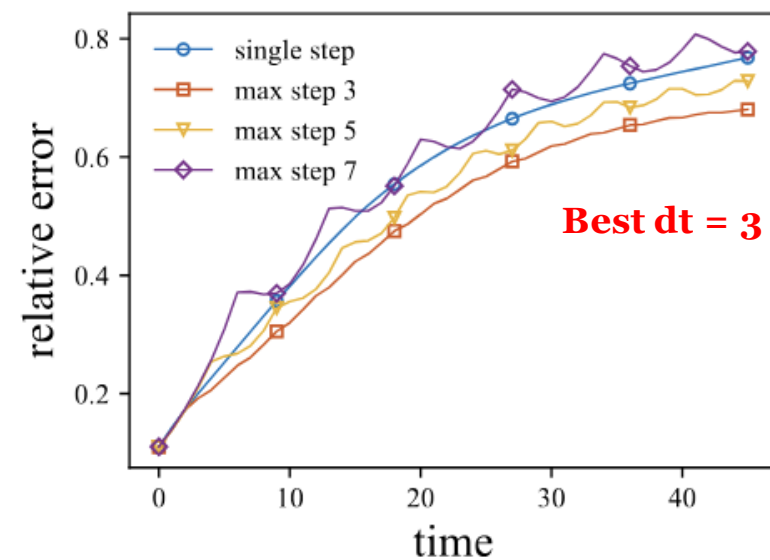
Pilot study -> select best Δt



(a) PDEBench-Comp-LowVis

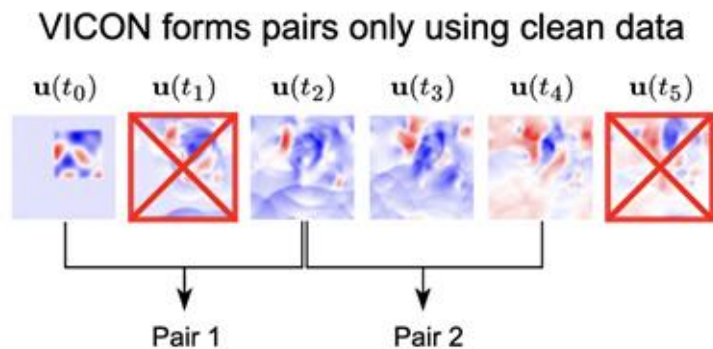
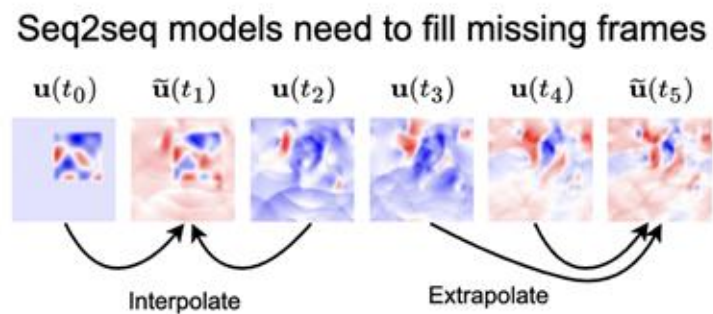
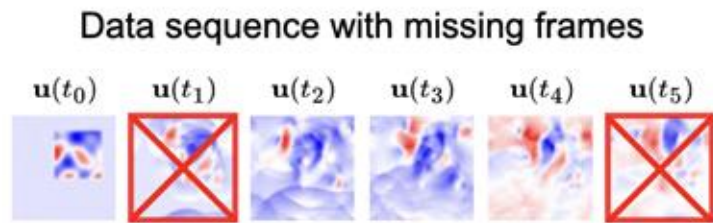


(b) PDEBench-Comp-HighVis



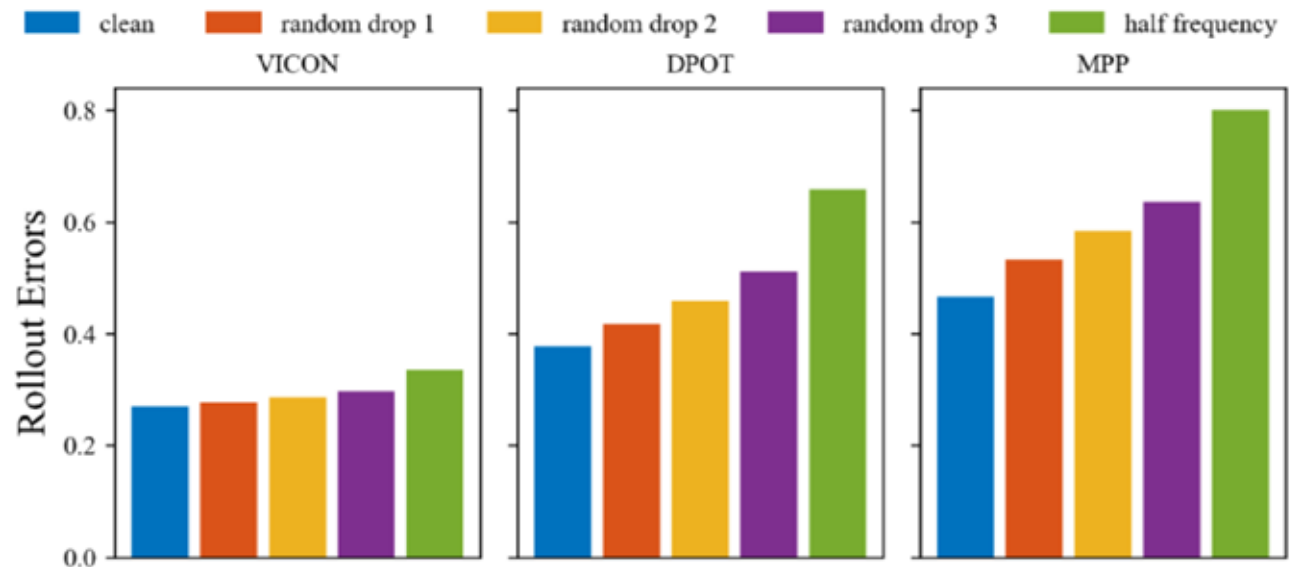
(c) PDEArena-Incomp

Robust to Irregular Sampling



(c) Imperfect Measurement

(c)



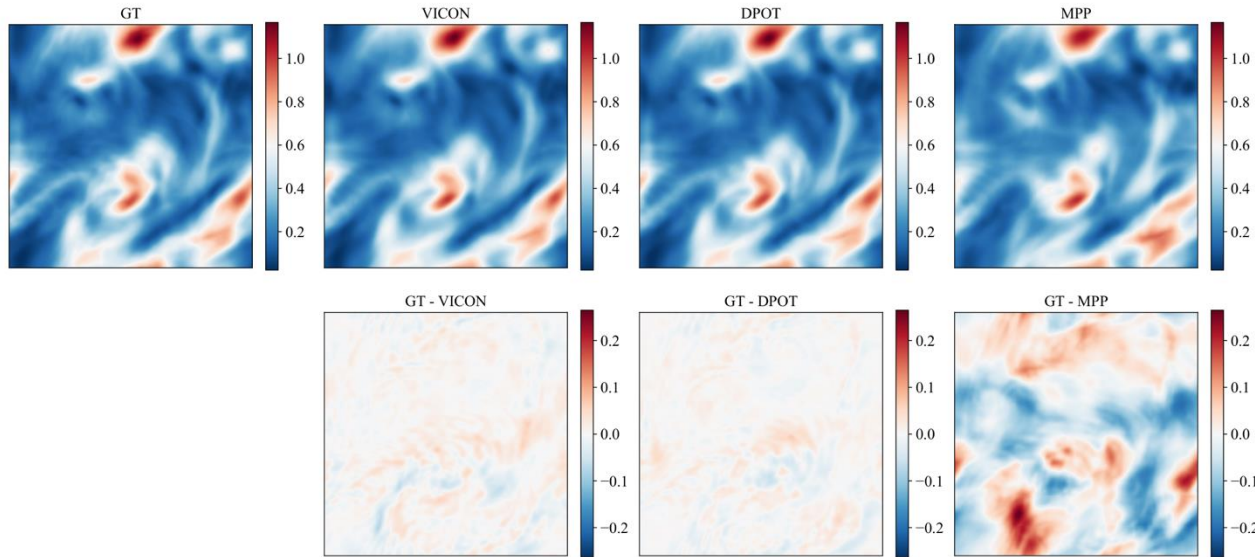
Error increase due to imperfection:

- VICON, 21%
- MPP/DPOT, 75%

TKE & Better Scaling

Turbulent Kinetic Energy:

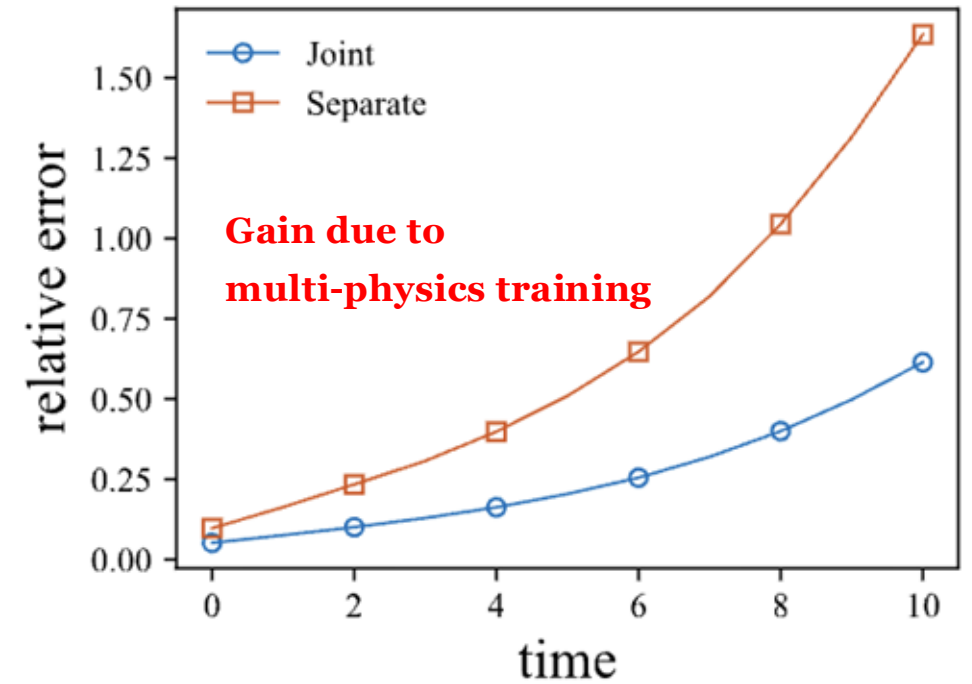
- Measured for fully developed subset



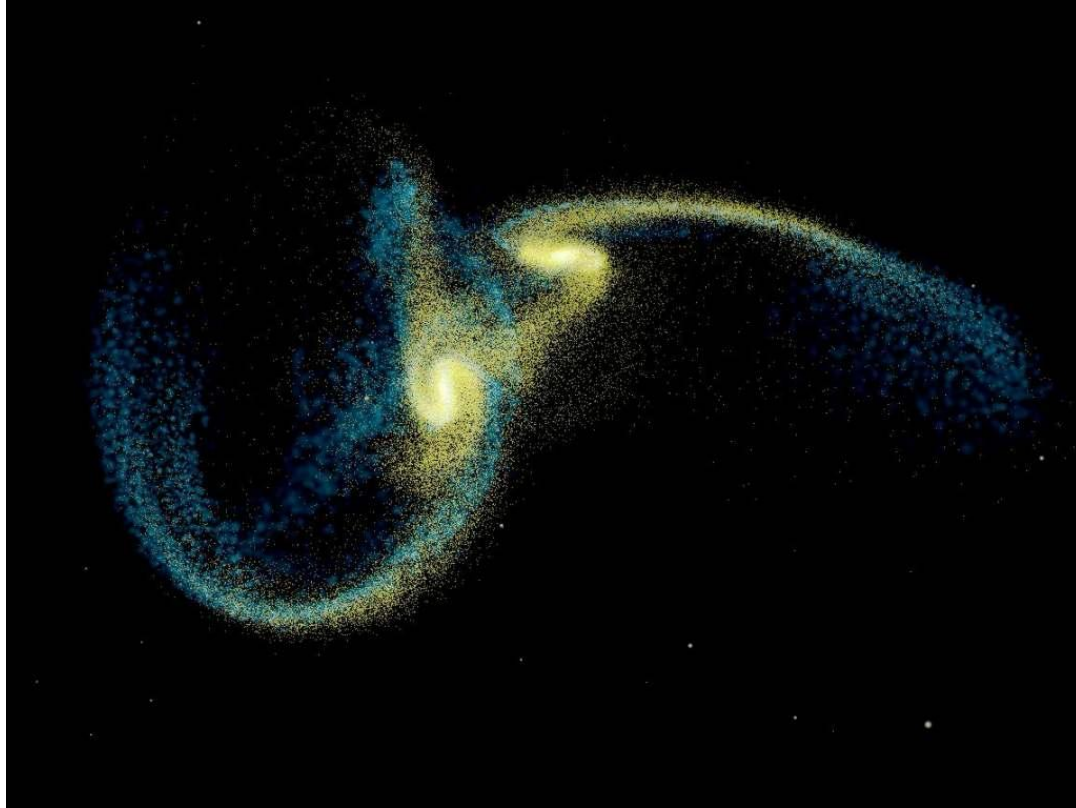
Scaling law:

- w/ the same training cost
- Train separately vs jointly on all datasets

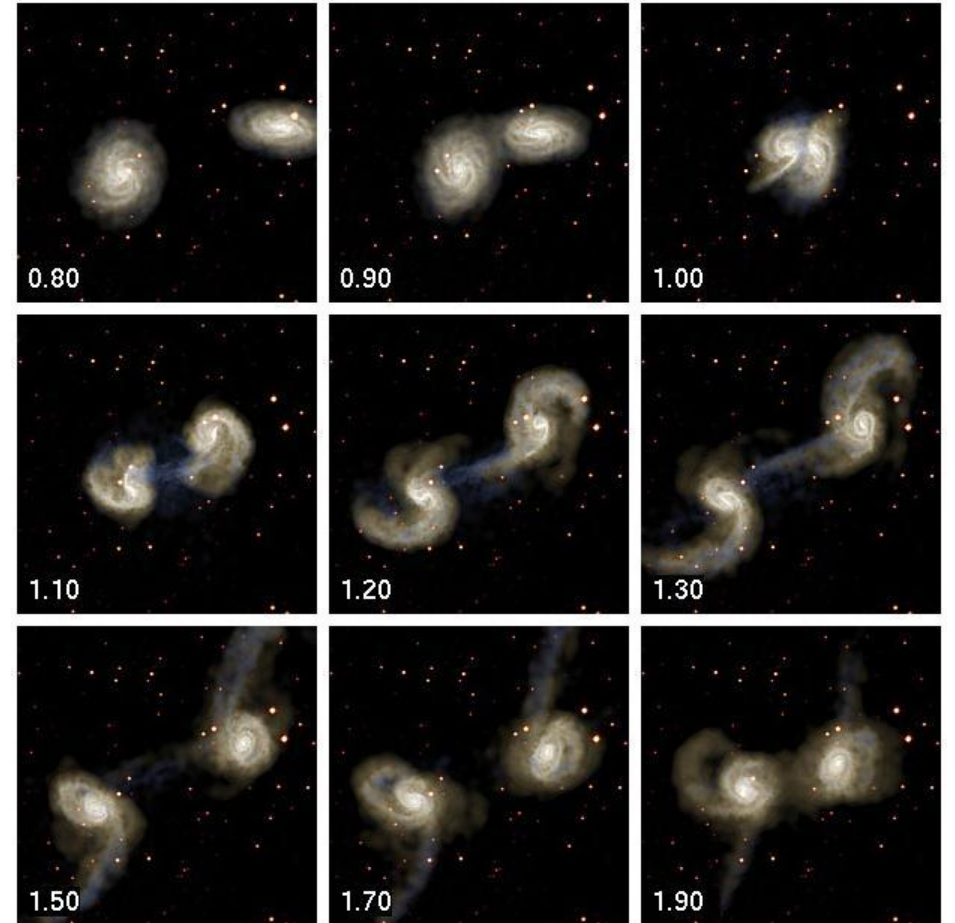
Potential as foundation model



Outlooks: Point Cloud & Lagrangian Systems



[OSU SPH astrophysics image gallery](#)

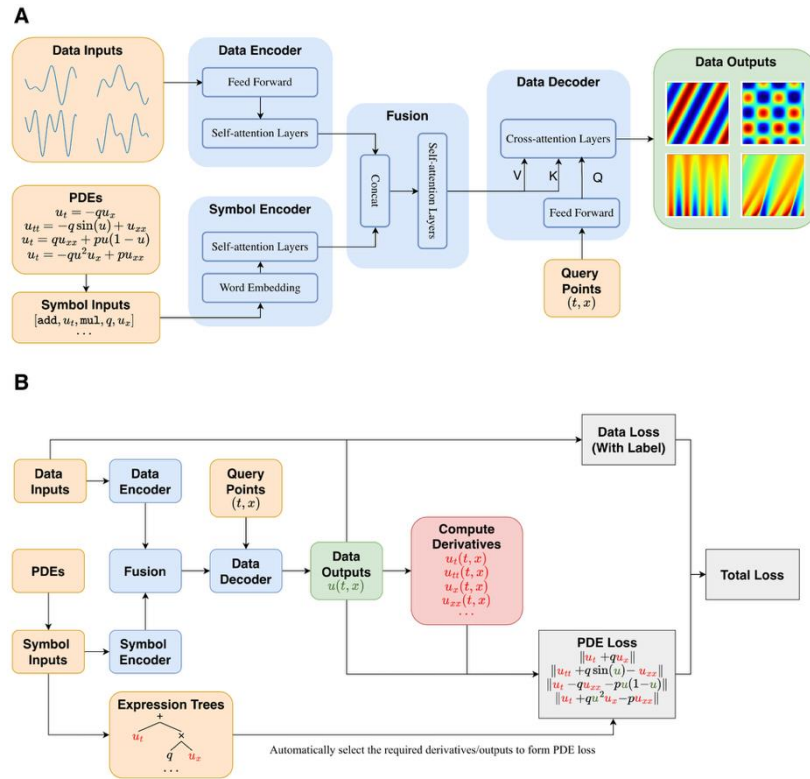


[SPLASH](#)

Point cloud + VICON

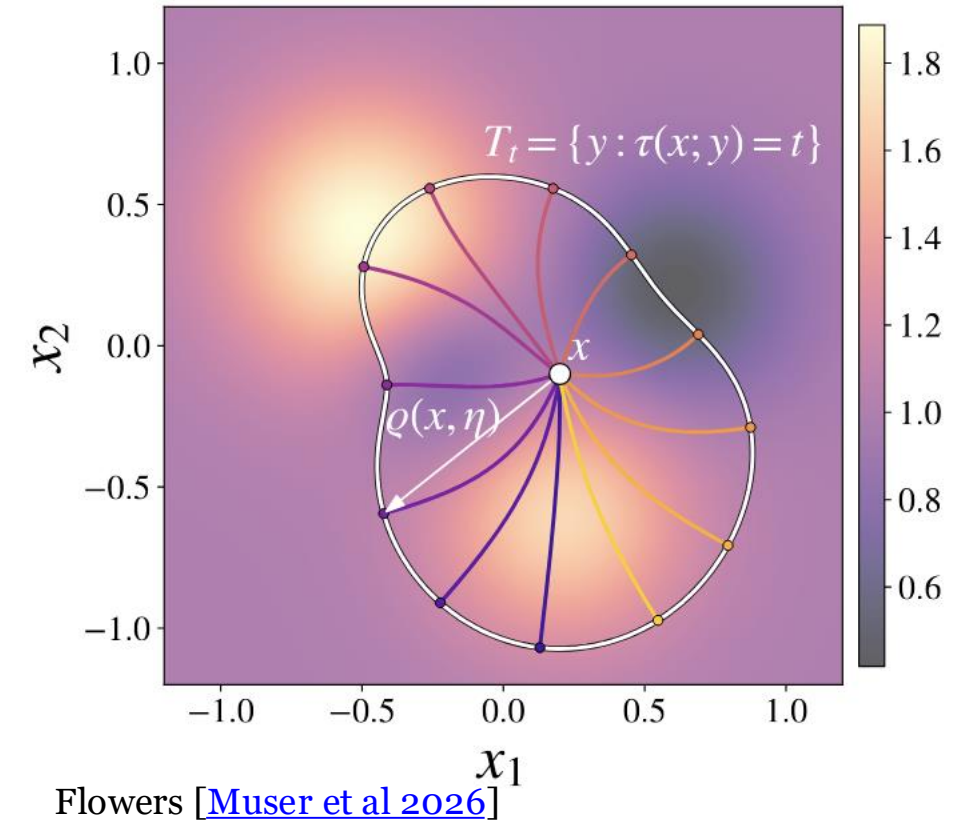
Outlooks: Embed Abstract Structures

Concrete, exact info



Unclear system?

Abstract info



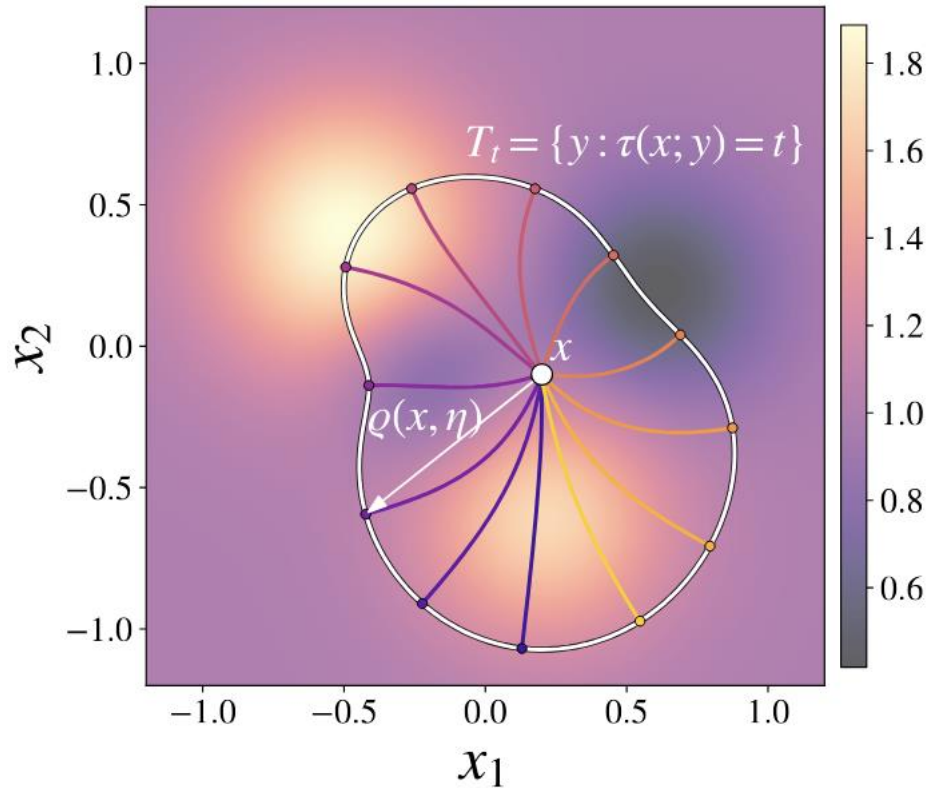
[Sun et al 2026]

- Data-efficient, but
- Requires exact info:
- Eqs, LaTeX formulas

- Abstract structure: eg, wavefront
- General within 1 structure, but
- No transferability across structures

Outlooks: Embed Abstract Structures

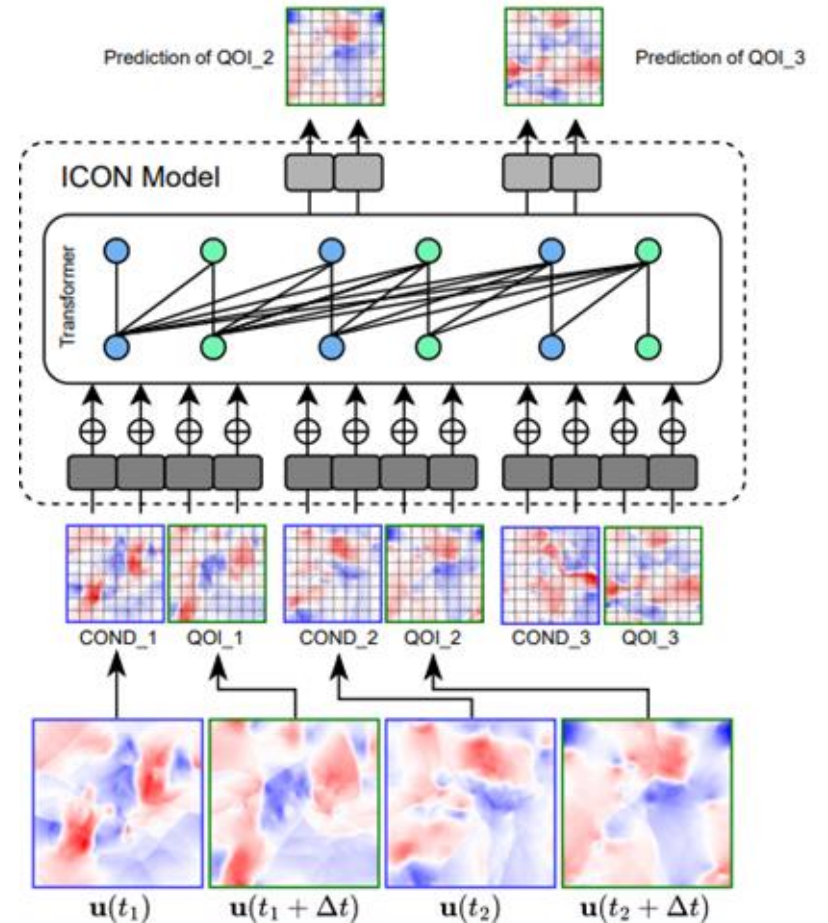
- 1 abstract physical structure: eg, learning the wavefront
- **No transferability across structures**



Flowers [Muser et al 2026]

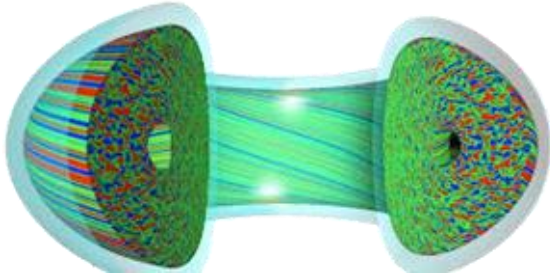
Enhanced by

VICON [3]

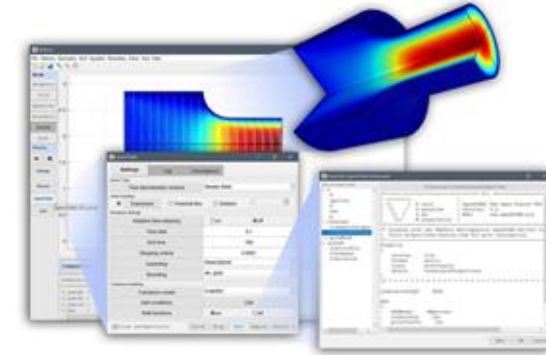


In-context learning enables multi-structures

My Lab's Directions



How to make experiments affordable?



How to reduce human labor?

Realistic Digital Twins

Respect data cost

- AI for Fusion [1]



Overview

General geometry

- BSMS-GNN [2]



Multi-physics/condition

- VICON [3]



Deploy



Accelerate

Realistic AI Engineers

Automated workflows

- FoamAgent [4]



Benchmarks that matter

- SimulCost [5]



Transferable sol

- Adapting-While-Learning [6]



Generate



Train

Digital Twin for Fusion [1]

AL saves 75% training data, on 1 GPU:

- 22 days, low fidelity
- **217 years, 1.12 M USD, high fidelity**

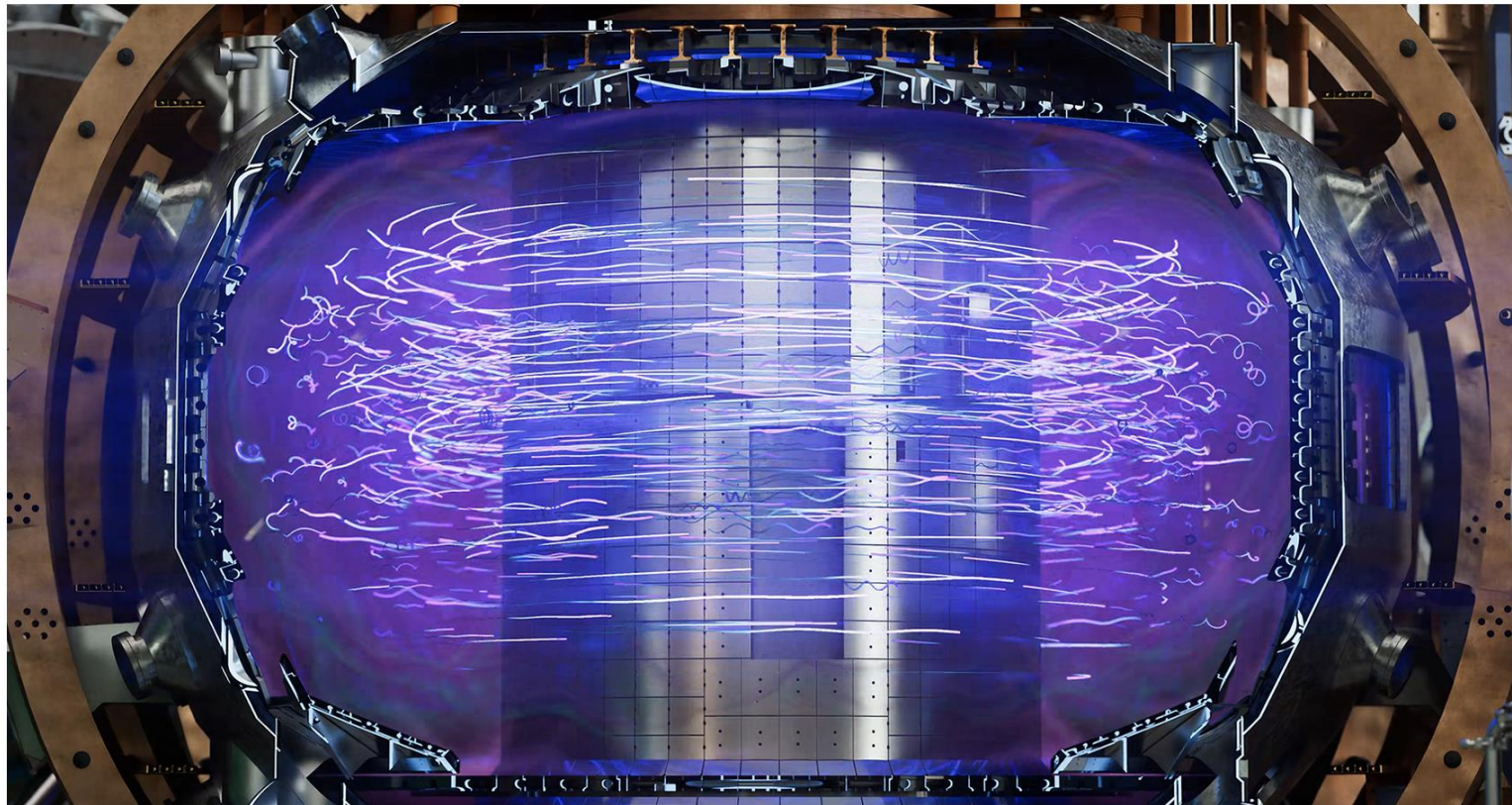
Speedup 45x whole-device tokamak simulation:

- **2 months -> 0.125 hrs/flight**
- Deployed at DIII-D

NVIDIA and General Atomics Advance Commercial Fusion Energy

NVIDIA and General Atomics, with support from UC San Diego, Argonne and NERSC, deliver a high-fidelity digital twin for fusion energy research.

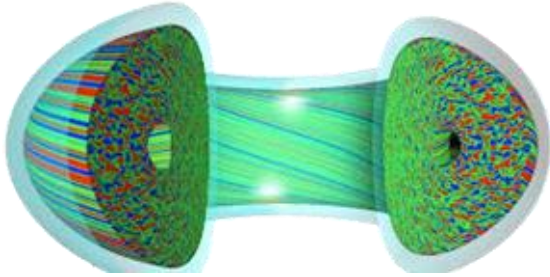
October 28, 2025 by [Chris Porter](#)



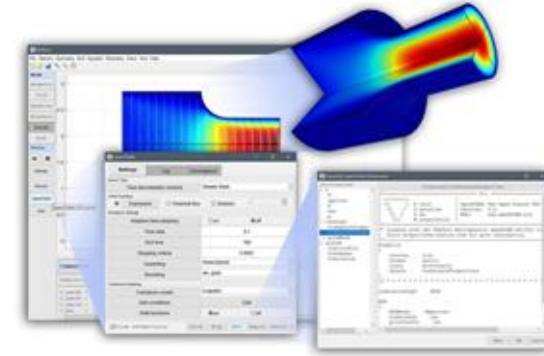
[Nvidia GTC 2025 keynote](#)



My Lab's Directions



How to make experiments affordable?



How to reduce human labor?

Realistic Digital Twins

Respect data cost

- AI for Fusion [1]



Deploy



Accelerate

General geometry

- BSMS-GNN [2]



Overview

Multi-physics/condition

- VICON [3]



Generate



Train

Realistic AI Engineers

Automated workflows

- FoamAgent [4]



Benchmarks that matter

- SimulCost [5]



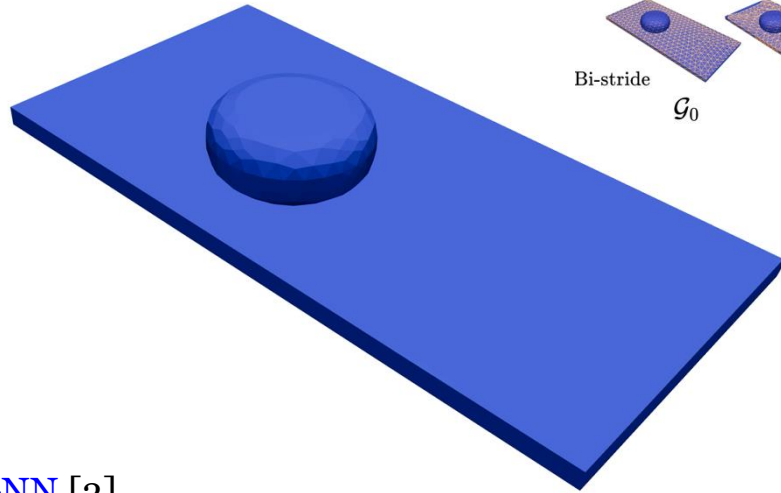
Transferable sol

- Adapting-While-Learning [6]

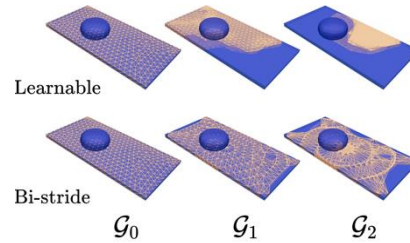
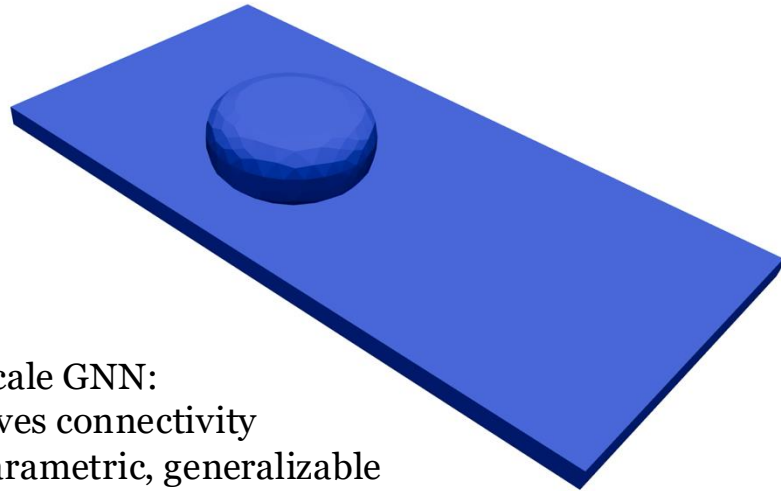


Digital Twins for Complex Geometry: BSMS-GNN [2]

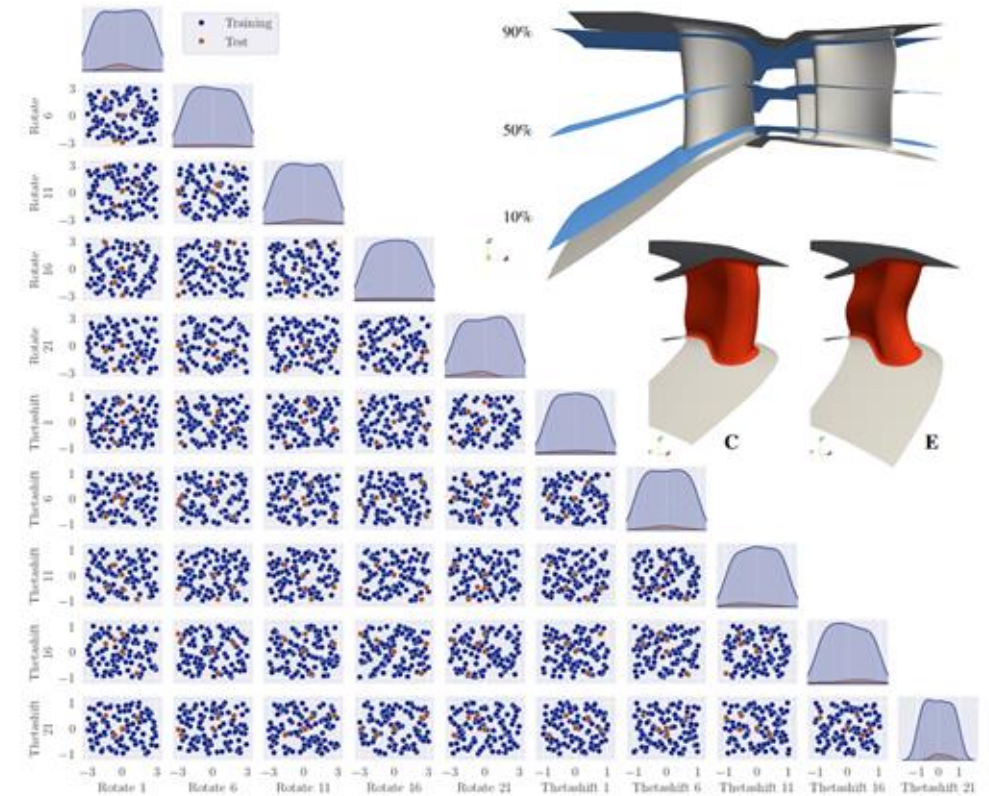
[Gao et al, 2019]



BSMS-GNN [2]



[Stronisch et al, 2025]



1st Multi-Scale GNN:

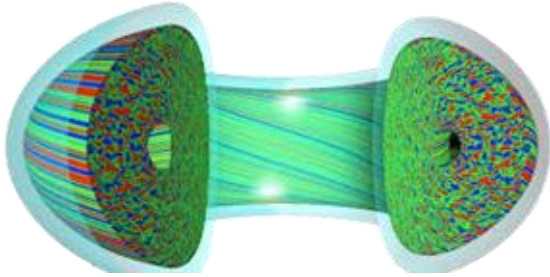
- Conserves connectivity
- Non-parametric, generalizable
- Parameter efficient

Deployed at Rolls-Royce

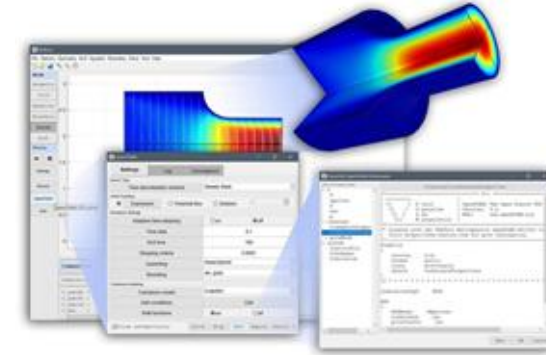
- Realistic turbomachinery
- **w/ 2.5 million nodes**



My Lab's Directions



How to make experiments affordable?



How to reduce human labor?

Realistic Digital Twins

Respect data cost

- AI for Fusion [1]



Deploy



Accelerate

General geometry

- BSMS-GNN [2]



Multi-physics/condition

- VICON [3]



Generate



Train

Realistic AI Engineers

Automated workflows

- FoamAgent [4]

Computer
Methods
in Applied
Mechanics and
Engineering

Overview

Benchmarks that matter

- SimulCost [5]

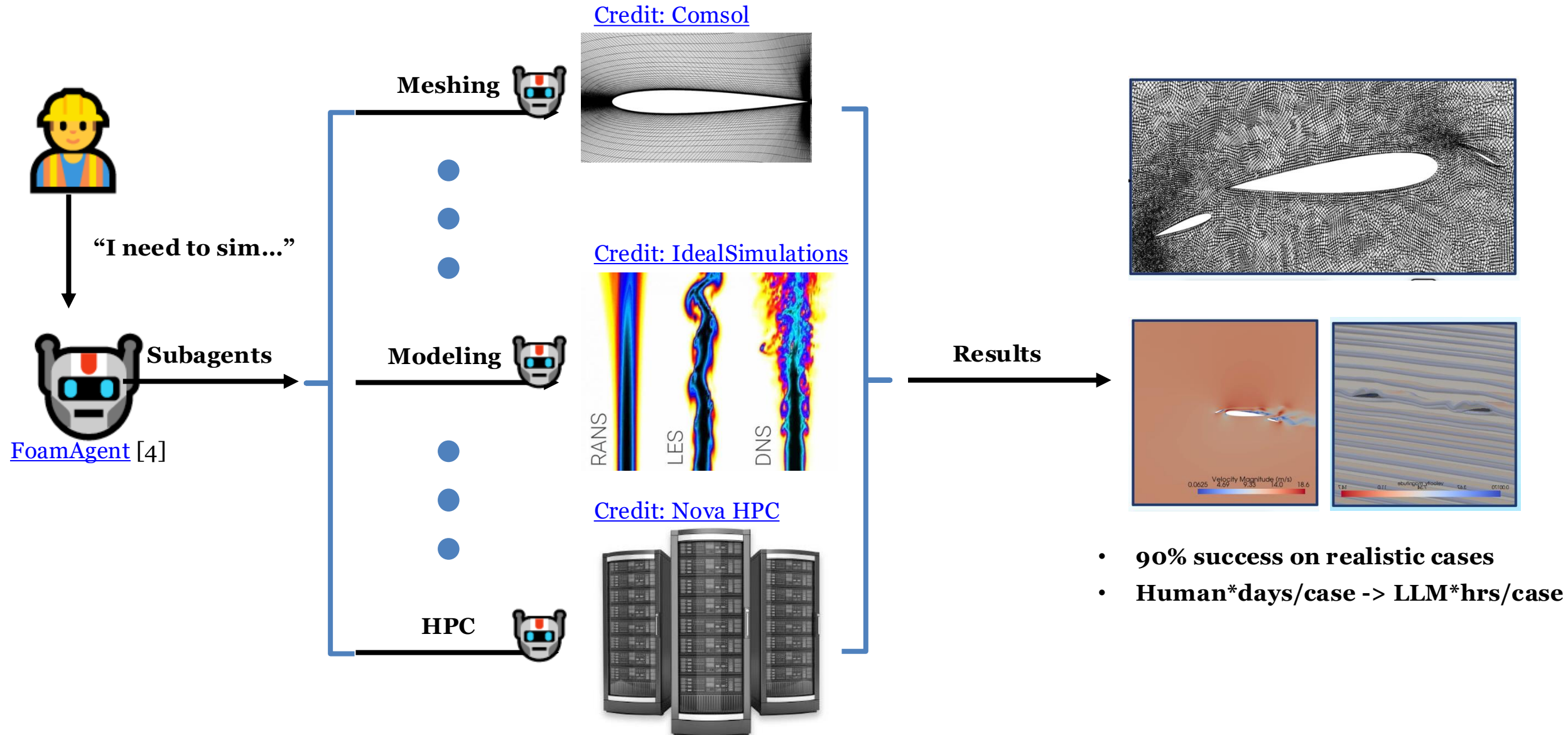


Transferable sol

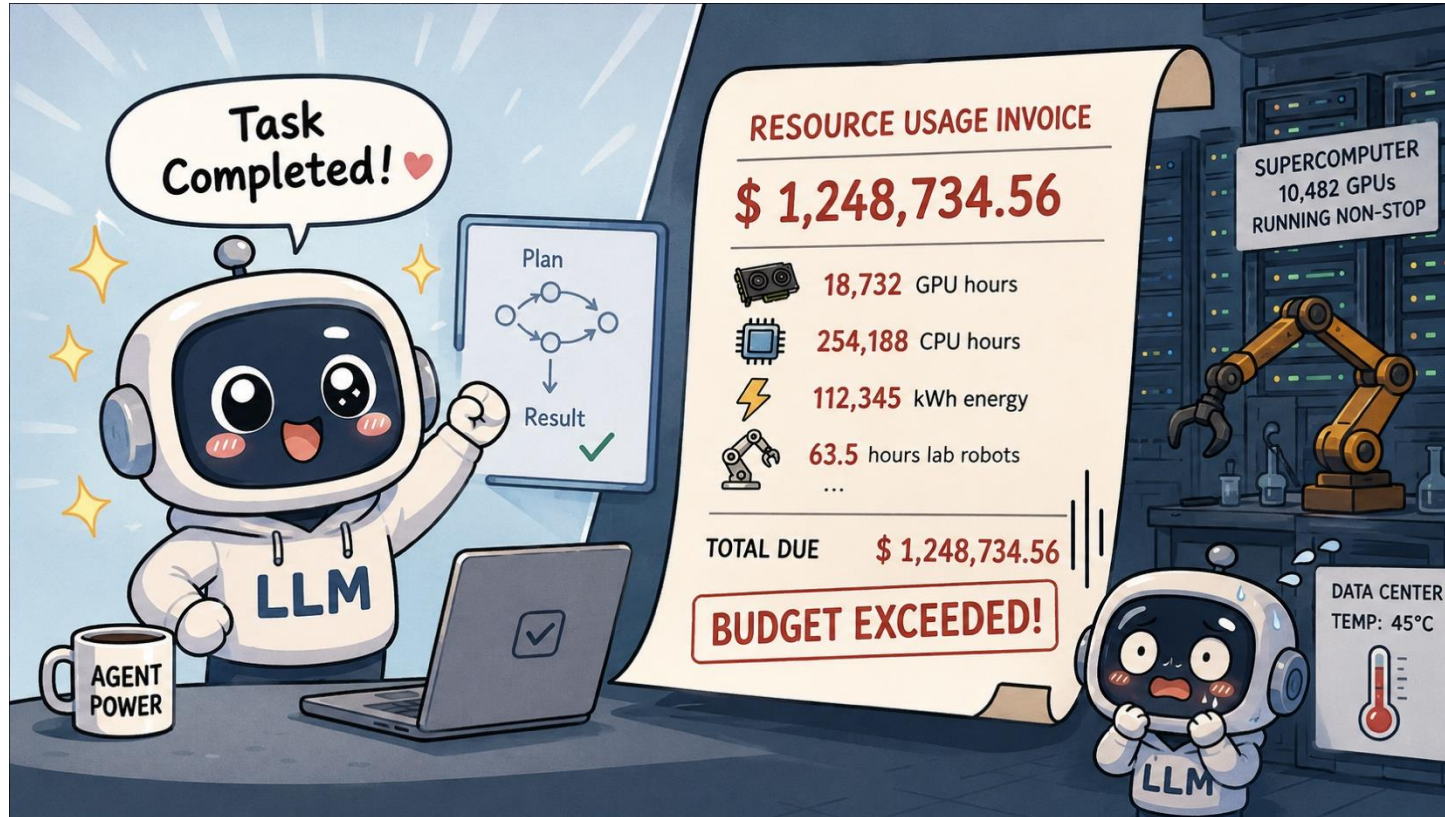
- Adapting-While-Learning [6]



FoamAgent [4] Automates CFD Configuration



Resource Matters



Created by GPT 5.6 Sol

<https://resource-aware-workshop.github.io/>

NeurIPS 2026

CALL FOR PAPERS

Scan for the site

Dec 12–13, 2026
Day set by NeurIPS

Atlanta, Georgia
United States

Aug 29, 2026
Deadline, AoE

Submit on OpenReview
Up to 8 pages · non-archival

Resource-Aware Agentic AI

Building smarter AI agents that do more with less. A NeurIPS 2026 workshop on designing agents that are both **efficient** and **effective**, from training to deployment.

About the Workshop

Motivation

With the rapid advances in AI, we have witnessed a wave of breakthroughs in autonomous AI agents. Today's agents write and debug code, solve competition-level mathematical problems, and conduct deep research.

Designing **resource-aware agents** that reason about, account for, and adapt to their own resource budgets is becoming increasingly important.

However, most are designed to maximize task performance. They are largely unaware of the resources they consume: compute, energy, memory, latency, tool cost, and data.

Topics of Interest

Not limited to

- Resource-aware planning and reasoning
- Resource-aware multi-agent orchestration
- Resource-aware agent harness
- Memory and context management
- Economic and accessibility implications

- Model compression for agents
- Resource-aware training and post-training of agents
- On-device agent deployment
- Benchmarks, metrics, and evaluation of resource consumption
- Industry applications of efficient agent deployment

Key Dates

Double-blind · NeurIPS format

SUBMISSION, AOE
August 29, 2026

REVIEWS DUE
September 20, 2026

NOTIFICATION, AOE
September 29, 2026

WORKSHOP
December 12–13, 2026

Organizers

Academia & industry

Kai-Wei Chang
UCLA

Dakuo Wang
Northeastern

Luke Zettlemoyer
UW / Meta

Jiri Gesi
Microsoft

Weijia Shi
Meta / Cornell Tech

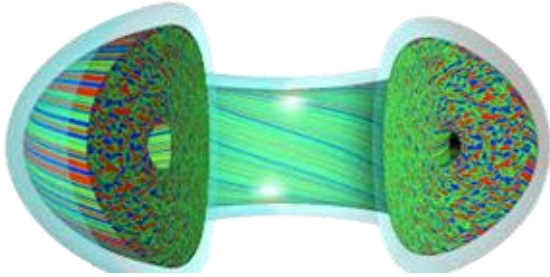
Jiaxin Pei
UT Austin / Stanford
HAI

Yuxuan Lu
Northeastern

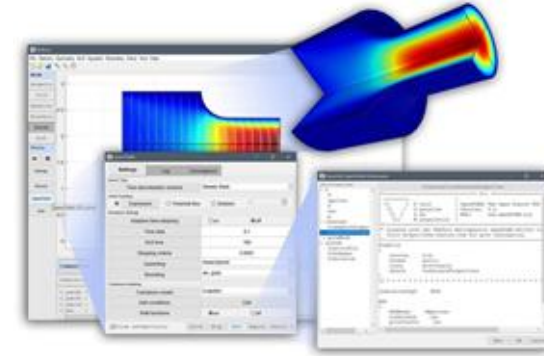
resource-aware-workshop@googlegroups.com

resource-aware-workshop.github.io

My Lab's Directions



How to make experiments affordable?



How to reduce human labor?

Realistic Digital Twins

Respect data cost

- AI for Fusion [1]



Deploy



Accelerate

General geometry

- BSMS-GNN [2]



Multi-physics/condition

- VICON [3]



Generate



Train

Realistic AI Engineers

Automated workflows

- FoamAgent [4]



Benchmarks that matter

- SimulCost [5]



Overview

Transferable sol

- Adapting-While-Learning [6]



Gap: Feasibility and Efficiency

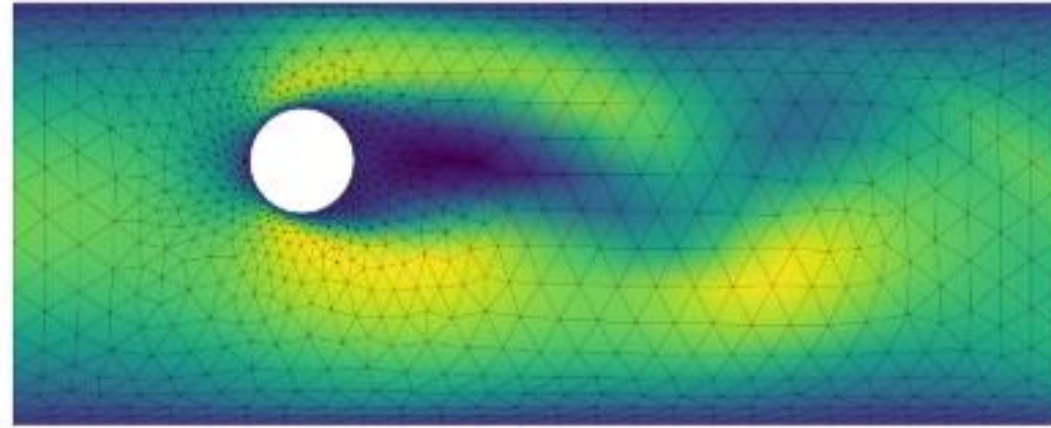
Cost-accuracy trade-off:

- Mesh resolution
- Time step
- Interpolation order
- ...



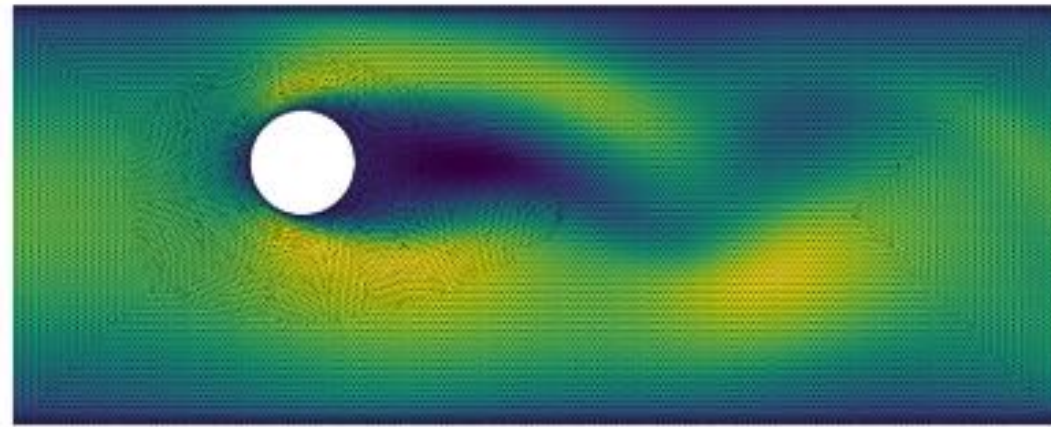
Example agent task:

- Tune cylinder flow's mesh resolution



Too coarse?

- **Large error**
- Hurts downstream



Too fine? solution good, but

- **High computing time**
- Hurts workflow efficiency

Strike balance

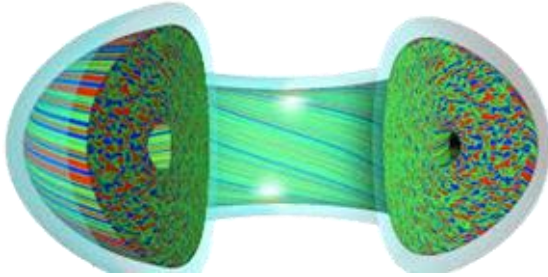
- Cost vs Accuracy

Is there a benchmark for accuracy-cost tradeoffs?

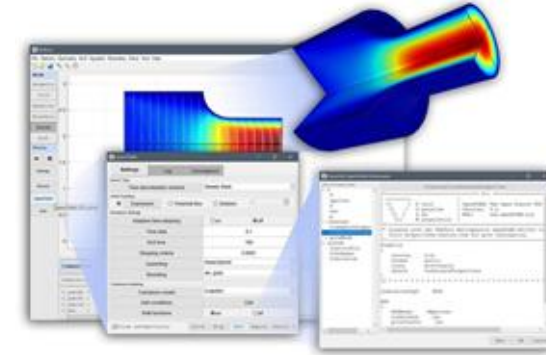
None! Worse:

Cost-ignorant metrics: pass@K (K=1024)

My Lab's Directions



How to make experiments affordable?



How to reduce human labor?

Realistic Digital Twins

Respect data cost

- AI for Fusion [1]



Deploy



Accelerate

General geometry

- BSMS-GNN [2]



Multi-physics/condition

- VICON [3]



Generate



Train

Realistic AI Engineers

Automated workflows

- FoamAgent [4]



Benchmarks that matter

- SimulCost [5]



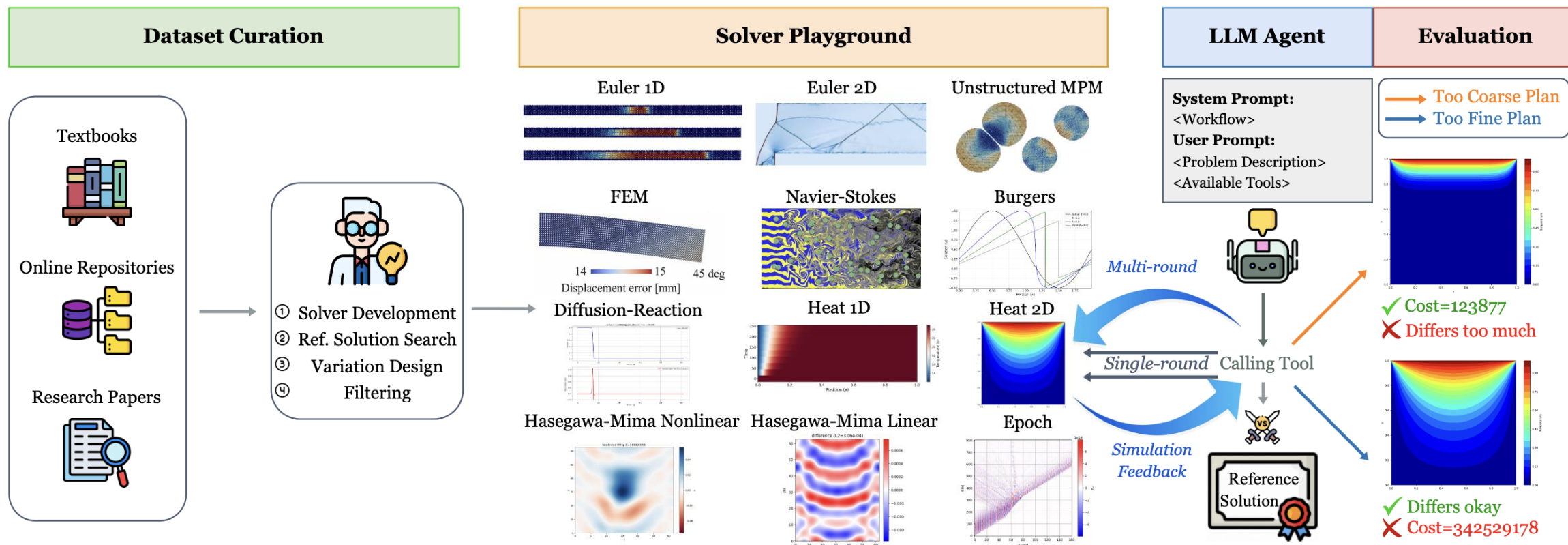
Overview

Transferable sol

- Adapting-While-Learning [6]



SimulCost [5]: 1st Benchmark for Accuracy-Cost Tradeoffs



Experts organize solvers

- Develop & debug
- Analytical cost definition (by FLOPs)

Scale up

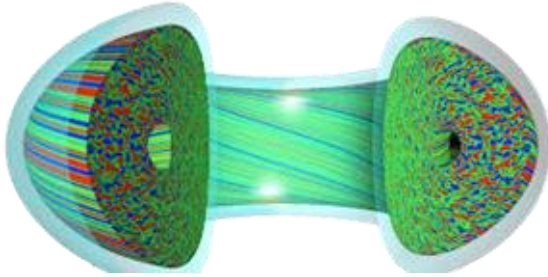
- Simulators & scenarios

Benchmarking

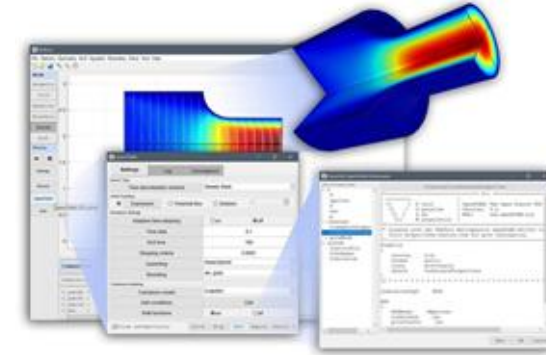
- LLM tuning software
- Measure accuracy + cost-efficiency

- **LLMs lack physics parameter intuition**
- **Its “tuning” is 2-3x more “costly”**

My Lab's Directions



How to make experiments affordable?



How to reduce human labor?

Realistic Digital Twins

Respect data cost

- AI for Fusion [1]



Deploy



Accelerate

General geometry

- BSMS-GNN [2]



Multi-physics/condition

- VICON [3]



Generate



Train

Realistic AI Engineers

Automated workflows

- FoamAgent [4]



Benchmarks that matter

- SimulCost [5]



Transferable sol

- Adapting-While-Learning [6]



Overview

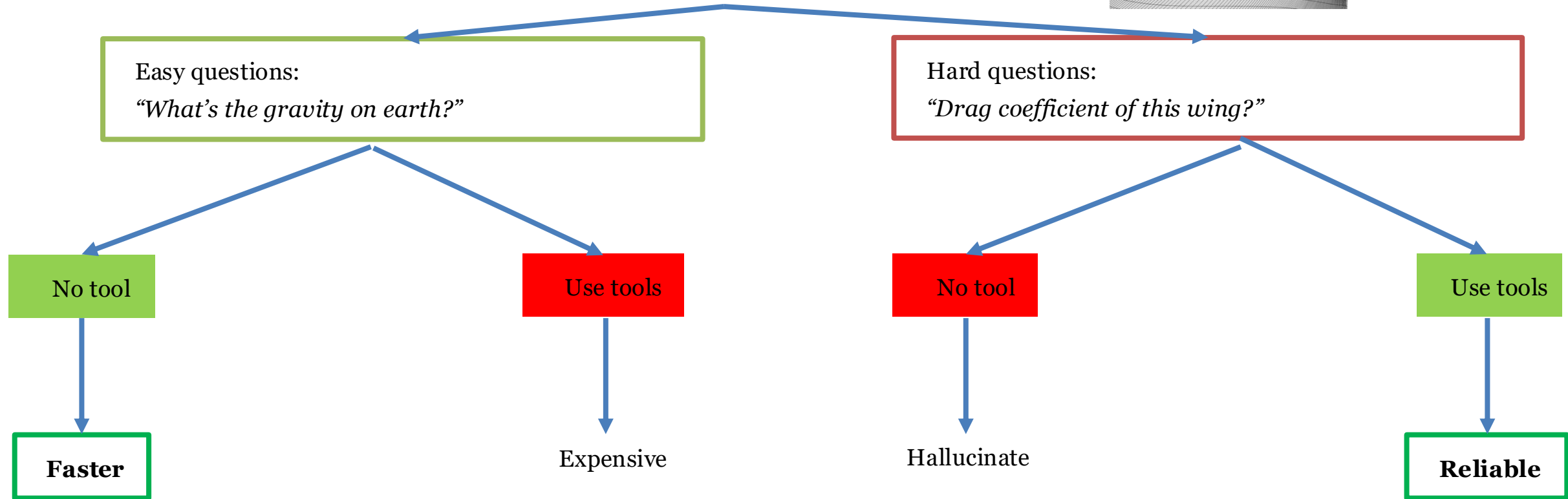
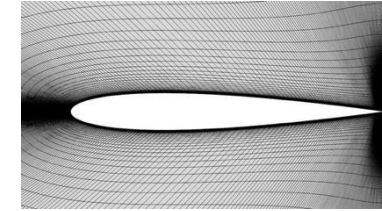
A Simplified Trade-off: when to use a tool?

Agent tasks are mixed:

- Conceptual + analytical + computing...



[Credit: Comsol](#)

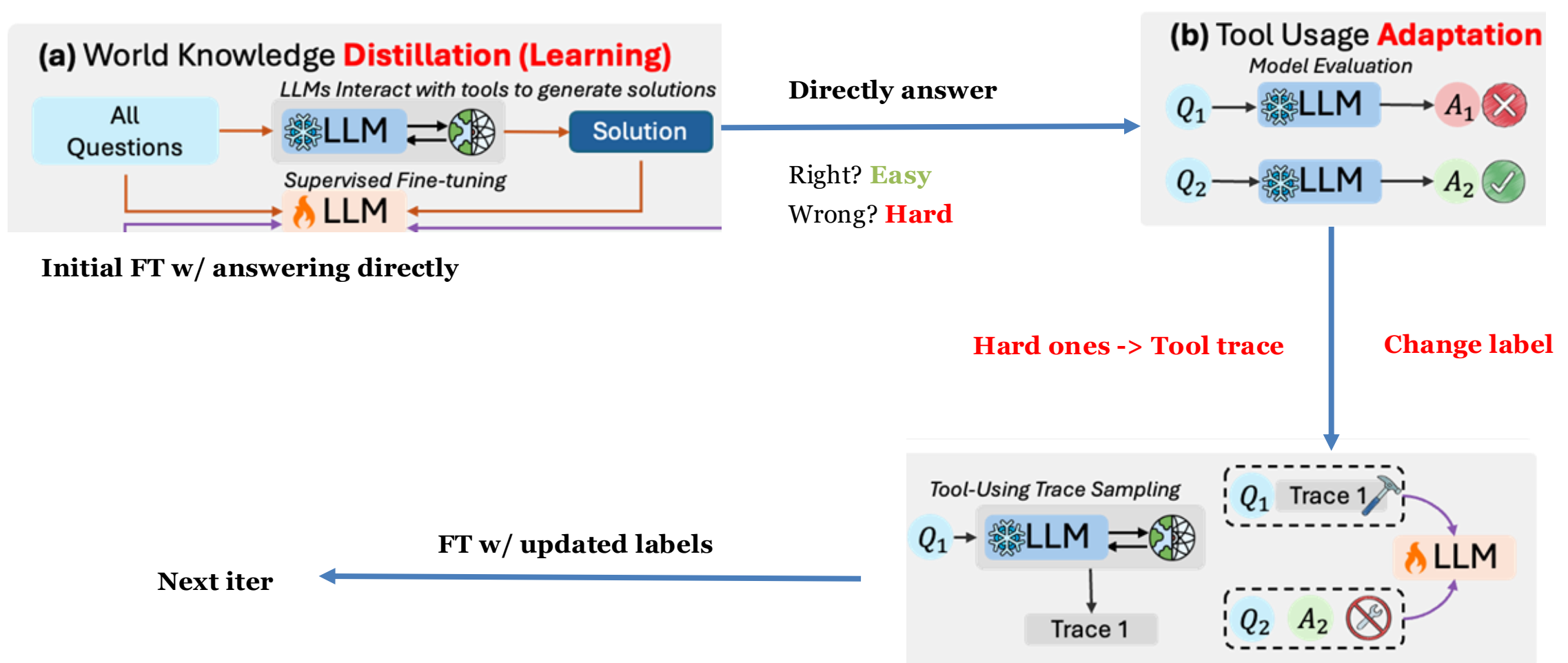


Finetune LLM to make the right choice?

- How to get labels, **at scale?**

AWL: Adapting While Learning

LLM labels easy/hard on its own

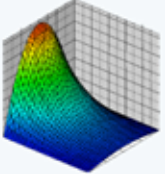


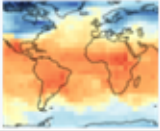
AWL + open models > Closed models
- Both for answer accuracy + tool-usage rate

AWL's datasets

(b) Demo Questions

Mujoco	
	In a physics laboratory, a double pendulum experiment is set up with the following parameters: $\{\{parameters\}\}$ How does the position of the first pendulum change over the 1-second observation period?

PDE
In a 2D population spread process with $\{\{initial\ condition\}\}$ and $\{\{boundary\ condition\}\}$, what is the population at $(x,y)=(59,2)$ km after 8 years? ($D=0.88$, $L_x=96$ km, $L_y=8$ km) 

Climate	
	How much will the temperature of Gikongoro in 2018 under ssp585 change if the emission of CO2 is decreased by -25%?

Epidemiology
$\{\{scenario\ description\}\}$ $\{\{initial\ states\}\}$ On which day will the number of hospitalized cases in California reach its maximum? 

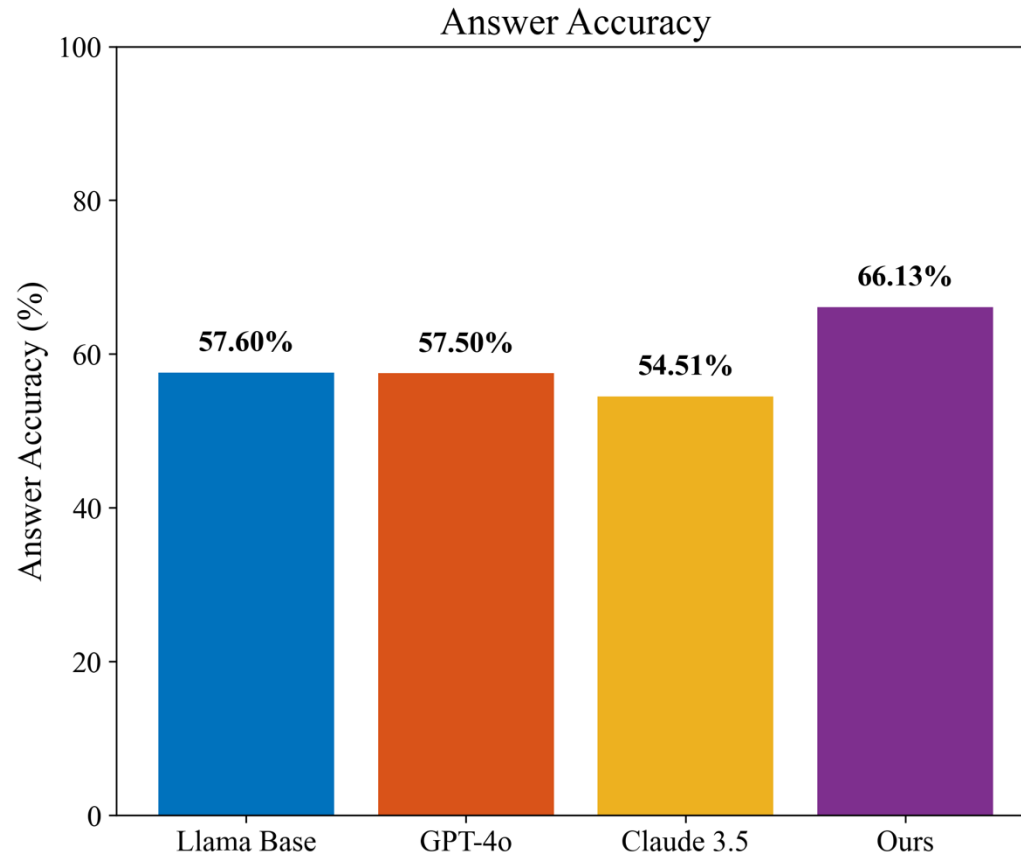
+ 2 public datasets:

MATH [\[Hendrycks et al, 2023\]](#), **SciBench** [\[Wang et al, 2023\]](#)

Better Knowledge Mem + Tool-Invoke Decision

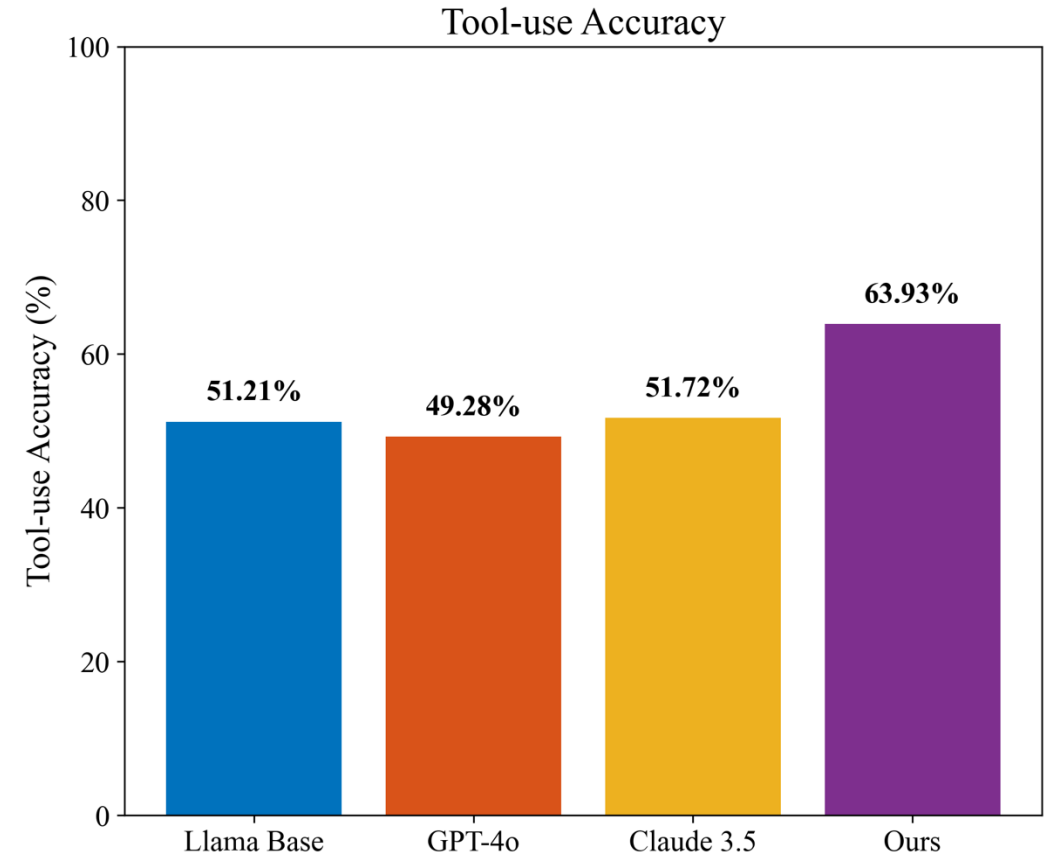
Answer Accuracy:

- Check results correctness



Tool-Usage Accuracy:

- Only invoke tools when the question is hard

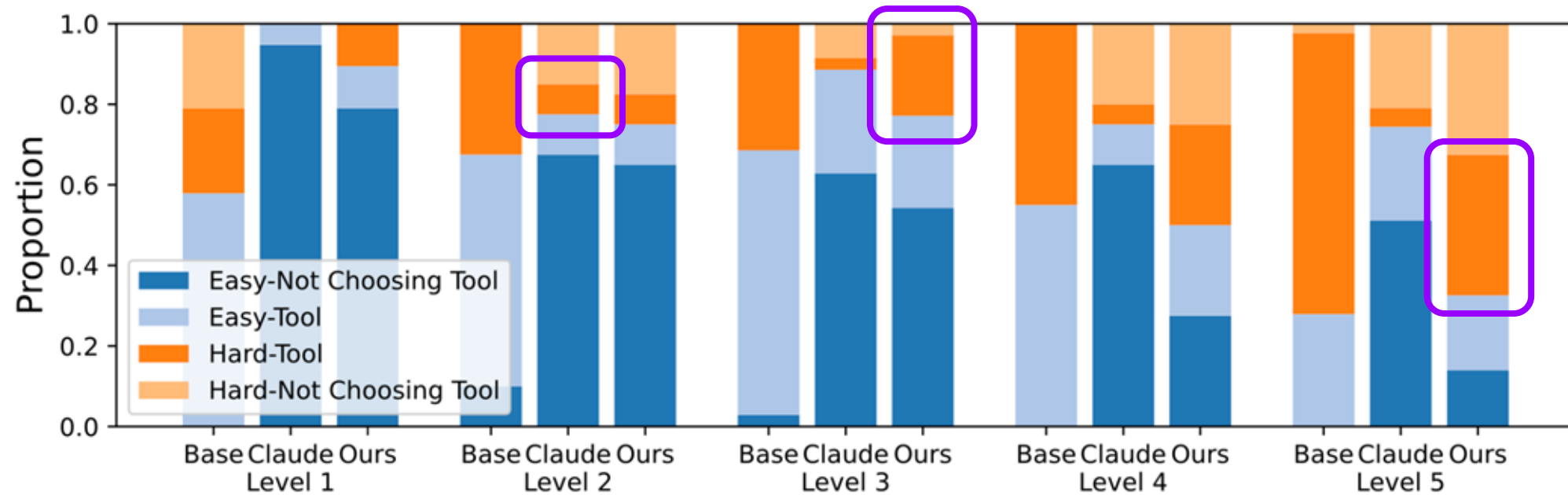


Opensource model + AWL > closed SOTA

Natural Handling of Difficulty Levels

MATH [\[Hendrycks et al, 2023\]](#) dataset:

- w/ labeled difficulty levels



AWL switches to tools for challenging questions

- Promising router for mixed task settings

Expanding Cost Efficiency to Broader Scientific Tasks

- From simulation-only to broad tasks
- Multi-“modality” of cost-tracking
- HCI (not anonymous) as baseline

SciTra: Cost-Aware Science Trajectory

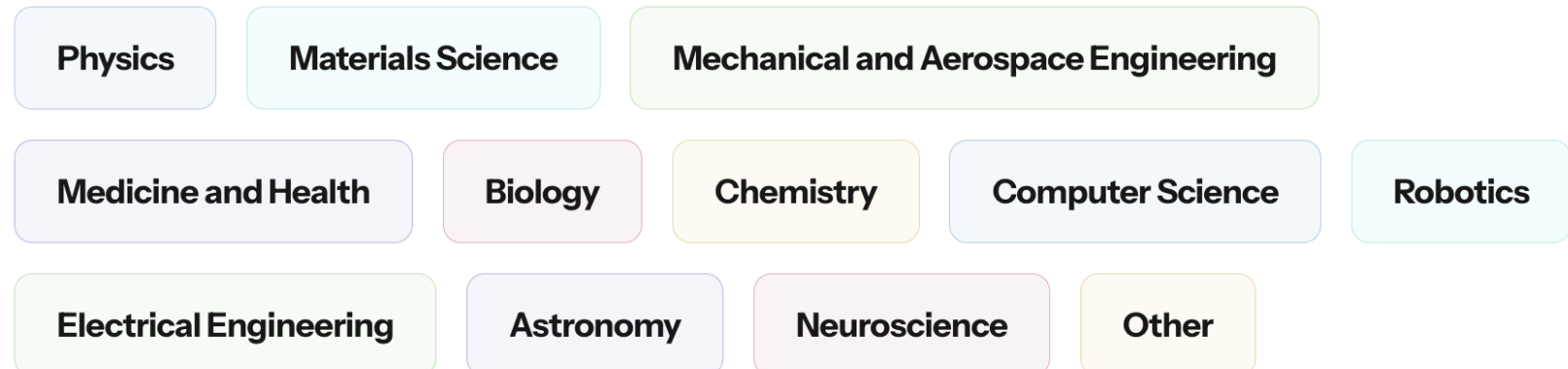
Evaluating the paths to scalable scientific discovery.

Scientific agents can reach the same outcome through very different sequences of decisions — and at very different costs. SciTra evaluates long-horizon scientific agents through their outcomes, resource use, and execution trajectories.



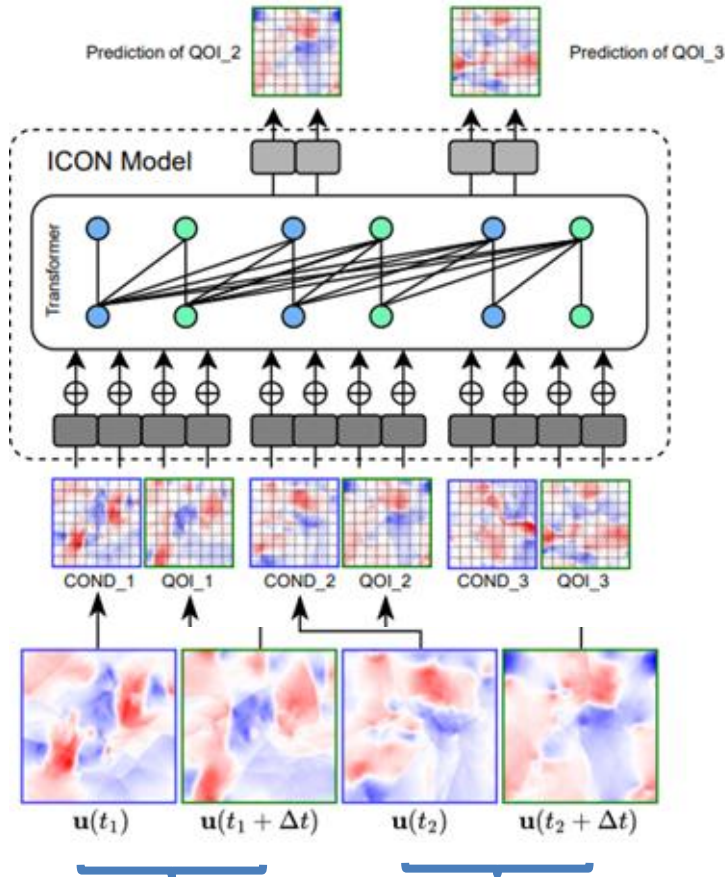
Across science and engineering

SciTra is being built across a growing range of scientific and engineering domains.



[Explore current coverage →](#)

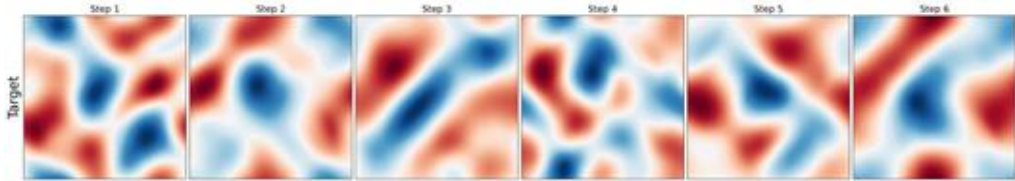
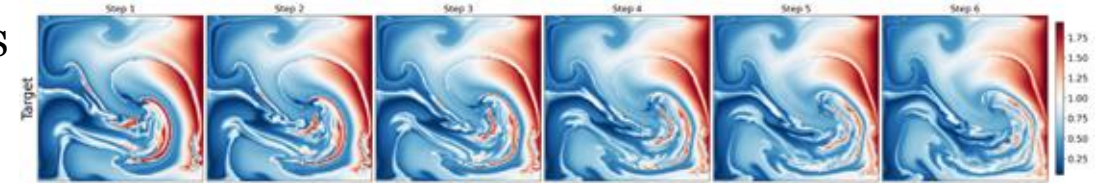
Recap of today: VICON



Consistent context pairs
Per row in batch

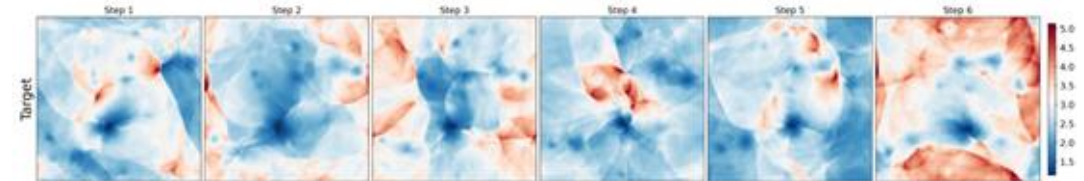
VICON learns
→
Extracting

Incompressible NS



Compressible NS
1e-2 viscosity

Compressible NS
1e-8 viscosity



Shuffle

Form

SOTA performance + No finetuning:

- Compared to MPP, DPOT

Adapt to shifted measurement frequencies, etc



Acknowledgements



End and QA?

Thank You!

If you have any questions, please email me at:
yadica095@gmail.com