
Fine-tuning a Multi-modal Generative Model for Inertial Confinement Fusion on NIF Experiments

Michael S. Jones
Lawrence Livermore National Lab

Noah Gonzalez
University of Wisconsin, Madison

Daniel Casey
Lawrence Livermore National Lab

Eugene Kur
Lawrence Livermore National Lab

Bogdan Kustowski
Lawrence Livermore National Lab

Kelli Humbird
Lawrence Livermore National Lab

Abstract

Inertial confinement fusion (ICF) experiments are expensive and sparse, while the simulations used to model them often exhibit discrepancies with experimental observations. Moreover, experiments provide only partial and noisy measurements across a heterogeneous suite of diagnostics, making adaptation of simulation-trained models challenging. We address this problem by fine-tuning a multi-modal generative diffusion model, pretrained on ICF simulations, using data from 48 National Ignition Facility (NIF) experiments. Our framework combines (1) imputation of missing experimental modalities, (2) simulation replay to maintain supervision over unobserved quantities, and (3) a round-trip objective that encourages self-consistency across inference directions. We demonstrate improved prediction of scalar quantities and neutron images on held-out experiments, as well as enhanced consistency under round-trip inference tests. Our method provides a general framework for adapting simulation-trained generative surrogates to sparse, heterogeneous experimental observations.

1 Introduction

Inertial Confinement Fusion (ICF) is currently the only approach to controlled nuclear fusion that has achieved ignition, with the energy gain ratio at the National Ignition Facility (NIF) steadily rising from 1.5 in 2022 to 4.2 in 2025 (1; 2). Recently, there has been significant investment in fusion technology with the hope of developing a next-generation energy source, however many challenges remain to robust and wall-plug efficient fusion power (3). In ICF, a primary challenge is the scale and resources required to perform deuterium-tritium experiments, meaning that physics-based simulations of various fidelities are integral to the design process, experimental data sets are sparse, and diagnostics must be prioritized and provide only a partial picture of the physics of interest (4; 5). It is therefore crucial to develop models that can leverage sparse experimental observations to correct simulation-trained surrogates and make predictions under uncertain and incomplete measurements.

Recently, a generative model based on a joint diffusion objective was shown to predict accurate multi-modal conditional distributions of 3D Rocket-Piston (RP) ICF simulations (6). The model is a strong surrogate for the simulator, however the simulator itself is approximate and does not capture all hydrodynamic effects present in higher order codes and experiments. NIF experiments have been used to fine-tune simulations before (7; 8; 9), but these efforts have generally been limited to particular sets of scalars and single images from <10 experiments. However, there are many more multi-modal quantities computed in simulations that are either diagnostically unreliable or infeasible

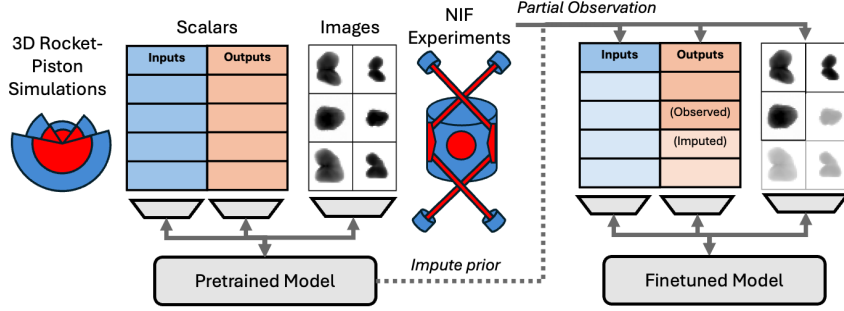


Figure 1: Simulation pre-training and experimental fine-tuning. A joint diffusion model is trained on scalar and image data from an ensemble of 3D RP simulations. Data from 48 NIF experiments are used to conditionally generate missing outputs and inputs (transparent boxes and images) and subsequently fine-tuned on.

to consistently capture across a larger ensemble of experiments (10; 11). Moreover, the availability of data for each experiment will vary as diagnostics are under repair and new tools come online, yet it is desirable to use all available data when fine-tuning a model.

In this work, we curate a dataset of 48 National Ignition Facility (NIF) experiments from the “Hybrid-E” campaign, which was the first platform to achieve a burning plasma and ignition (1; 12; 13; 14). Each experimental “shot” consists of partial scalar and image diagnostic information relative to the simulation-derived pre-training data, and these shots are used to fine-tune a multi-modal generative model. First, we use the pre-trained model to impute missing experimental modalities, similar to generative approaches for missing data imputation (15; 16). Although training loss is applied only to observed experimental quantities, these imputations act as a prior that helps prevent the model from drifting out of distribution during fine-tuning. Second, we incorporate simulation replay (17; 18) by randomly sampling from the original training data to maintain supervision over unobserved scalar and image modalities. Finally, we supervise noisy input predictions based on their ability to reproduce observed outputs.

We evaluate our method using cross-validation splits of 8 hold-out NIF experiments and show that fine-tuning improves prediction of both scalar quantities and neutron images across multiple conditional generation tasks. Additionally, we demonstrate improved consistency under round-trip inference tests, where inferred inputs are passed back through the forward surrogate model. While this work remains preliminary, the proposed framework enables adaptation of simulation-trained generative models to sparse and incomplete experimental observations without requiring complete correspondence between simulated and measured quantities.

2 Data and model overview

3D Rocket-Piston Simulator: The 3D rocket-piston (RP) model captures the impact of drive asymmetries on ICF implosions by discretizing a spherical implosion into 200 segments (19). Individual rocket-pistons are coupled through the hotspot pressure, with power-balance equations governing their collective dynamics based on Springer et al. (20), and a Callahan hohlraum model is used to derive the radiation drive (21). A suite of synthetic diagnostics at stagnation and peak burn produce scalar and image observables that can be compared to experiment. The model has previously been used to construct large-scale data sets for machine learning-based reconstruction tasks (6; 22; 23). The ensemble used for pre-training contains 443,610 simulations, each consisting of 28 scalar inputs, 12 scalar outputs, and 6 images which are a combination of 3 line-of-sight views and 2 neutron energy channels (primary and down-scattered).

NIF Experiments: Data was aggregated from 48 deuterium-tritium experiments performed between 2019 and 2025 as part of the Hybrid-E campaign at the National Ignition Facility (NIF) (12; 1; 13; 14). These data include 6 scalar quantities with corresponding uncertainties: log neutron yield, ion temperature, down-scattered ratio (DSR), and hot spot velocities in the x, y, and z directions. Out of 6 view/channel image combinations obtained from simulations, between 2-5 were available for each experiment. Image availability varies based on detector function, diagnostic overlap, and post-shot

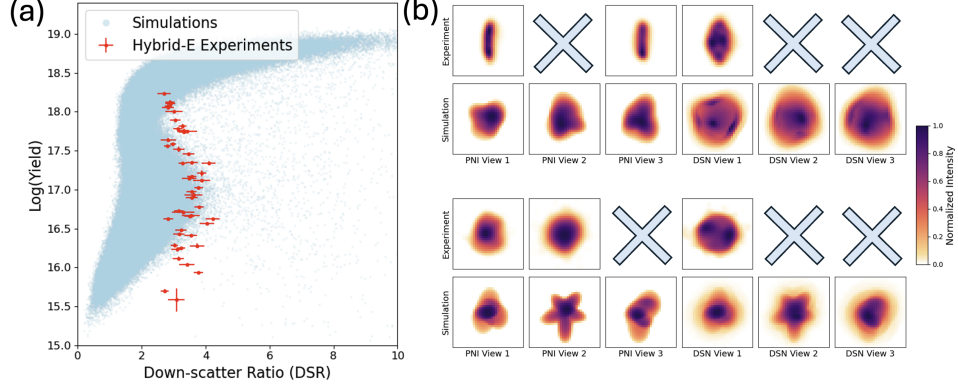


Figure 2: (a) Simulated (blue) and experimental (red) yield vs. down-scattered ratio (DSR) distributions. (b) Example image data for two shots showing all available views and channels. For both shots, the first row shows experimental data, where blue Xs indicate a missing channel. The second row shows images from the nearest neighbor training simulation in scalar space.

analyses. Numerous sources of variability stemming from laser delivery, capsule quality, etc. are not modeled by the RP simulation model but can have a significant impact on experiments. In Figure 2a we show the distribution shift between the neutron yield and down-scattered ratio (DSR) for experiments compared to simulations. Due to this shift, simulations with similar scalar values to experiments often have distinct image features (as shown in Figure 2b) and vice versa.

Joint diffusion model: We use a JointDiff model (6), pre-trained on 443,610 RP simulations, to learn the joint distribution $q(i, s, x)$ of simulation inputs, scalar outputs, and image outputs. Denoting the combined multi-modal state as $z = (i, s, x)$, the diffusion process constructs noisy states

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (1)$$

and trains a denoising network ϵ_θ to predict added noise using the diffusion objective (24; 25),

$$L_{\text{diff}} = \mathbb{E}_{t, \epsilon} [\|\epsilon - \epsilon_\theta(z_t, t, M)\|^2]. \quad (2)$$

A binary mask M specifies arbitrary observed quantities, which are held fixed while diffusion is applied to the remaining quantities. Randomizing M during training allows a single model to sample conditional distributions $q(z_M | z_M)$ enabling forward, inverse, imputation, and mixed-modality inference from arbitrary subsets of inputs, scalar outputs, and image channels. Images are processed using a U-Net architecture (26), while scalar inputs and outputs use fully connected networks.

Fine-tuning: Experimental data $q_e(s_e, x_e)$ contain partial output scalars and images, while direct capsule inputs are unavailable. To augment the limited experimental data, we independently sample the measured scalar uncertainties,

$$\tilde{s}_e = s_e + \epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0, \sigma_{s,e}^2).$$

We perform 6-fold cross-validation on 48 NIF experiments, training on 40 shots per fold. For each training shot, we generate 100 noised realizations, yielding 4,000 experimental samples per fold. For each realization, missing scalars, images, and capsule inputs are imputed using the pre-trained model f_θ conditioned on the experiment-specific mask M_e , while measured quantities remain fixed:

$$(\hat{i}_e, \hat{s}_e, \hat{x}_e) \leftarrow f_\theta(\tilde{s}_e, x_e, M_e).$$

During fine-tuning, experimental batches are sampled with 50% probability and mixed with simulated batches of equal size. For experimental batches, the diffusion loss is evaluated only on experimentally observed quantities, excluding imputed values from direct supervision. We additionally introduce a round-trip consistency loss to improve input inference without access to ground-truth experimental inputs. From the noisy input state i_t and predicted input noise $\hat{\epsilon}_i$, we estimate the corresponding clean inputs as

$$\hat{i}_0 = \frac{i_t - \sqrt{1 - \alpha_t} \hat{\epsilon}_i}{\sqrt{\alpha_t}}.$$

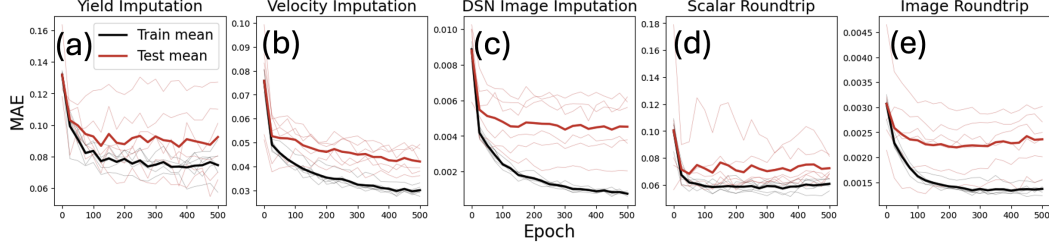


Figure 3: Train (black) and test (red) loss for 6-fold cross-validation performed on all 48 NIF experiments. For the first three tasks the ground truth values for yield (a), hot spot velocities (b), and view 1 DSN images (c) are separately withheld and inferred based on all other available scalars and images for that experiment. For the next two tests, round-trip inference is performed given all available scalars and images, and an MAE is computed for the reconstructed scalars (d) and images (e). Thin lines show results for each fold and thick lines show means across all folds.

These predicted inputs are then conditioned upon to reconstruct the experimentally observed outputs,

$$(\hat{x}_{\text{RT}}, \hat{s}_{\text{RT}}) \leftarrow f_{\theta}(x_t, \hat{i}_0, s_t, t),$$

and the resulting round-trip loss is

$$L_{\text{RT}} = \bar{\alpha}_t [\|x_e - \hat{x}_{\text{RT}}\|_{M_e}^2 + \|s_e - \hat{s}_{\text{RT}}\|_{M_e}^2].$$

The factor $\bar{\alpha}_t$ downweights round-trip predictions constructed from highly noised input states, and λ_{RT} weights the overall round-trip contribution to the loss. The total experimental loss is therefore

$$L = L_{\text{diff}} + \lambda_{\text{RT}} L_{\text{RT}}.$$

We observe similar performance for experimental sampling ratios between 25–75%, while higher ratios reduce generation stability for less represented image channels. Fine-tuning is performed for 500 epochs with improvement across three imputation and two round-trip tasks shown in Figure 3. Additional training details and procedure are provided in Algorithm 1 in the Appendix. We also experimented with freezing different portions of the architecture and more efficient weight updates with low-rank adaptation (27) (Figures A.1-A.2) to reduce overfitting and observed comparable performance under some conditions, however for this work we elect to update all weights.

3 Model Performance

Imputation: We first compare the fine-tuned model’s ability to predict missing scalars when given all other available experimental information. In this test, the missing scalar is measured experimentally

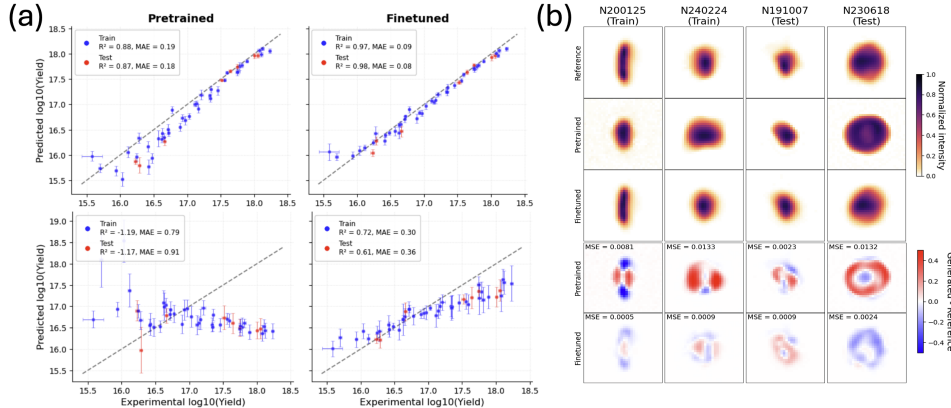


Figure 4: (a) Yield imputation across all experiments, with held-out test shots shown in red, conditioned on scalars and images (top left) or images only (bottom left). Errors report standard deviation across 10 generations. (b) Prediction of View 1 imputed images given all other experimental data for two training shots and two test shots. Bottom rows show MSE for pre-trained and fine-tuned predictions with respect to ground-truth image.

but withheld during conditional generation. In Figure 4, we show imputation performance on yield, which the pre-trained model predicts reasonably well but can struggle at lower values. After fine-tuning yield predictions are highly accurate with $R^2 > 0.97$ for both the train and test sets, and we no longer observe systematic under-prediction at low yield. Next, we pose the more challenging conditional prediction scenario in which the model only has access to experimental images and no scalars. For this task, the pre-trained model is likely too far out of distribution and shows no correlation with the experimental values. After fine-tuning, R^2 values improve to 0.72 for the train data and 0.61 for test data, indicating strong transferability despite significant variations in experimental image features and channel availability. Imputations for DSR and hot-spot velocities are included in Figure A.5, and also show significant improvements from fine-tuning.

In addition to improving scalar predictions, we also want to ensure that fine-tuning enhances experimental image prediction. In Figure 4b, we evaluate image imputation by removing the View 1 primary channel from a subset of shots that contain this ground-truth image and using all available scalar and other image channels to predict the missing channel. We compare generated and ground-truth images for each shot using both the pre-trained model (second row) and the fine-tuned model (third row), and observe improvement in each case. Interestingly, for the N230618 test shot, the pre-trained model substantially over-predicts intensity near the boundary while under-predicting at the center, whereas this behavior is reversed but more subtle for the fine-tuned model. Although predictions do not improve for every shot, we observe average improvement across each image view and channel. Improvement over training time for DSN images is shown in Figure 3c.

Round-trip: Next, we evaluate self-consistency between the forward and inverse models by performing a “round-trip” test in which we infer inputs from partial experimental data and then pass those inputs back through the forward model. Previously, it was shown that the pre-trained model could achieve reasonably strong round-trip reconstruction without fine-tuning (6), however in the top row of Figure 5 the round-trip over-predicts DSR for about half of samples and produces several velocity and yield outliers. After fine-tuning (bottom row), we observe no significant outliers and a reduction in DSR error with respect to experimental observation. We also measure round-trip reconstructions for image data and find an average decrease in pixel MSE relative to the pre-trained model. An example reconstruction for test shot N230618 is shown in Figure 5b, where generated images show improved fidelity to available ground-truth images. Improvement across scalars and images in each cross-validation split is shown in Figure 3d-e.

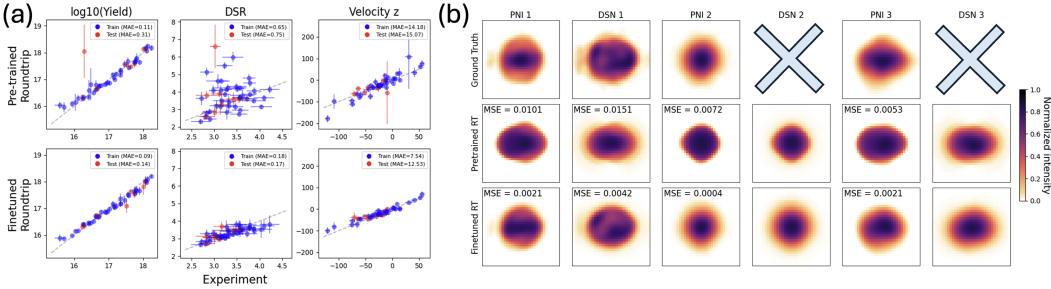


Figure 5: (a) Round-trip analysis evaluates model self-consistency. Partial experimental outputs are used to infer inputs, which are then passed back through the forward model to predict outputs. Means and standard deviations are computed over 10 generated samples for each experiment. (b) Round-trip image predictions for test shots N230618 given the pre-trained and fine-tuned models, with MSE computed against available ground-truth images. Blue Xs indicate missing experimental data.

4 Discussion

We present a framework for fine-tuning a multi-modal ICF surrogate using sparse NIF data. Partial imputation, simulation replay, and round-trip consistency improve missing-output predictions while preserving stability across unsupervised modalities. Limitations from simulation–experiment discrepancies and variable image features may be addressed by incorporating additional designs and diagnostics. Future applications include calibrating simulations through inferred inputs, imputing missing measurements, and designing experiments to improve implosion yield and robustness.

5 Acknowledgement

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program under Project No. 26-FS-008.

References

- [1] Kritcher, A., A. Zylstra, C. Weber, et al. Design of the first fusion experiment to achieve target energy gain $g > 1$. *Physical Review E*, 109(2):025204, 2024.
- [2] Hohenberger, M., D. S. Clark, A. E. Pak, et al. Demonstration of mix mitigation and record performance using continuous dopant targets in icf implosions on the nif. In *DPP 2025*. APS, 2025.
- [3] Hennick, C. Funding the fusion revolution. *MIT Energy Initiative*, 2025. Accessed: 2026-04-09.
- [4] Moses, E. I. The national ignition facility (nif): A path to fusion energy. *Energy Conversion and Management*, 49(7):1795–1802, 2008.
- [5] Rosen, M. D. The long road to ignition: an eyewitness account. *Physics of Plasmas*, 31(9):090501, 2024.
- [6] Jones, M., J. Kunimune, D. Casey, et al. Joint diffusion approach to multimodal inference in inertial confinement fusion. *PRX Intelligence*, 1(1):013012, 2026.
- [7] Kustowski, B., J. A. Gaffney, B. K. Spears, et al. Transfer learning as a tool for reducing simulation bias: application to inertial confinement fusion. *IEEE Transactions on Plasma Science*, 48(1):46–53, 2019.
- [8] Kustowski, B., J. A. Gaffney, B. K. Spears, et al. Suppressing simulation bias in multi-modal data using transfer learning. *Machine Learning: Science and Technology*, 3(1):015035, 2022.
- [9] Jones, M., B. Kustowski. Towards a multi-modal foundation model for inertial confinement fusion: Combining structured data and diagnostic images. In *1st ICML Workshop on Foundation Models for Structured Data*. 2025.
- [10] Bell, P. M., G. P. Grim. Target diagnostics physics and engineering for inertial confinement fusion iii. In *Proc. of SPIE Vol.*, vol. 9211, pages 921101–1. 2014.
- [11] Frenje, J. A. Nuclear diagnostics for inertial confinement fusion (icf) plasmas. *Plasma Physics and Controlled Fusion*, 62(2):023001, 2020.
- [12] Zylstra, A., O. Hurricane, D. Callahan, et al. Burning plasma achieved in inertial fusion. *Nature*, 601(7894):542–548, 2022.
- [13] Abu-Shawareb, H., R. Acree, P. Adams, et al. Achievement of target gain larger than unity in an inertial fusion experiment. *Physical review letters*, 132(6):065102, 2024.
- [14] Hohenberger, M., D. Clark, K. Humbird, et al. First demonstration of improved fusion yield with increased compression through reduced adiabat in inertial confinement fusion experiments at the national ignition facility. *Physical Review Letters*, 135(3):035101, 2025.
- [15] Zheng, S., N. Charoenphakdee. Diffusion models for missing value imputation in tabular data. *arXiv preprint arXiv:2210.17128*, 2022.
- [16] Wu, R., H. Wang, H.-T. Chen, et al. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.
- [17] Riemer, M., I. Cases, R. Ajemian, et al. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.
- [18] Rolnick, D., A. Ahuja, J. Schwarz, et al. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

- [19] Casey, D., J. Kunimune, O. Hurricane, et al. A multi-rocket piston model to study three-dimensional asymmetries in implosions at the national ignition facility. *High Energy Density Physics*, 54:101172, 2025.
- [20] Springer, P., O. Hurricane, J. Hammer, et al. A 3d dynamic model to assess the impacts of low-mode asymmetry, aneurysms and mix-induced radiative loss on capsule performance across inertial confinement fusion platforms. *Nuclear Fusion*, 59(3):032009, 2018.
- [21] Callahan, D., O. Hurricane, A. Kritcher, et al. A simple model to scope out parameter space for indirect drive designs on nif. *Physics of Plasmas*, 27(7), 2020.
- [22] Kunimune, J. H., D. T. Casey, B. Kustowski, et al. 3d reconstruction of an inertial-confinement fusion implosion with neural networks using multiple heterogeneous data sources. *Review of Scientific Instruments*, 95(7), 2024.
- [23] Kunimune, J., D. Casey, B. Kustowski, et al. Reconstruction of residual kinetic energy in inertial-confinement fusion implosions at the national ignition facility using multiple heterogeneous data sources and neural networks. *Physical Review E*, 113(6):065213, 2026.
- [24] Ho, J., A. Jain, P. Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.*, 33:6840–6851, 2020.
- [25] Song, J., C. Meng, S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [26] Ronneberger, O., P. Fischer, T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [27] Hu, E. J., Y. Shen, P. Wallis, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

A Fine-tuning Algorithm

Algorithm 1 Experimental fine-tuning with simulation replay and round-trip consistency.

```

1: Input: simulation data  $q(i, s, x)$ , exp. data  $q_e(s_e, x_e, M_e)$ , exp. uncertainty  $\sigma_{s,e}$ , pre-trained
   model  $f_\theta$ , diffusion steps  $T$ , noise schedule  $\bar{\alpha}_t$ , replay ratio  $\lambda_e$ , round-trip weight  $\lambda_{RT}$ 
2: for each experimental sample  $(s_e, x_e)$  do
3:    $\tilde{s}_e \leftarrow s_e + \epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0, \sigma_{s,e}^2)$  // augment with exp. noise
4:    $(\hat{i}_e, \hat{s}_e, \hat{x}_e) \leftarrow f_\theta(\tilde{s}_e, x_e, M_e)$  // impute missing quantities
5: end for
6: for batch = 1 to  $N$  do
7:   Sample  $t \sim \mathcal{U}(\{1, \dots, T\})$ ,  $\epsilon \sim \mathcal{N}(0, I)$ 
8:   Sample  $M \sim \text{Bernoulli}(0.5)$ 
9:   Sample batch type  $b_e \sim \text{Bernoulli}(\lambda_e)$ 
10:  if  $b_e = 0$  then // simulation replay
11:    Sample  $X_0 \sim q(i, s, x)$ 
12:  else // experimental batch
13:    Sample  $X_0, M_e \sim \hat{q}_e(i, s, x, M_e)$  // experimental batch
14:     $M[M_e] \leftarrow 0$  // exclude imputed targets
15:  end if
16:   $X_t \leftarrow \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ 
17:   $\tilde{X} \leftarrow M \odot X_0 + (1 - M) \odot X_t$  // condition on observed values
18:   $\hat{\epsilon} \leftarrow f_\theta(\tilde{X}, M, t)$ 
19:   $L_\theta \leftarrow \|\epsilon[M] - \hat{\epsilon}[M]\|_2^2$  // diffusion loss
20:  if  $b_e = 1$  then
21:     $\hat{i}_0 \leftarrow \frac{i_t - \sqrt{1 - \bar{\alpha}_t} \hat{\epsilon}_i}{\sqrt{\bar{\alpha}_t}}$  // predict clean inputs
22:     $(\hat{x}_{RT}, \hat{s}_{RT}) \leftarrow f_\theta(x_t, \hat{i}_0, s_t, t)$  // condition on predicted inputs
23:     $L_{RT} \leftarrow \alpha_t [\|x_e - \hat{x}_{RT}\|_{M_e}^2 + \|s_e - \hat{s}_{RT}\|_{M_e}^2]$  // regress on observed outputs only
24:     $L_\theta \leftarrow L_\theta + \lambda_{RT} L_{RT}$ 
25:  end if
26:   $\theta \leftarrow \theta - \eta \nabla_\theta L_\theta$ 
27: end for
28: return  $f_\theta$ 

```

B Training details

All base hyperparameters are consistent with pre-training (6) unless otherwise noted. The pre-trained model was trained on simulation data for 300 epochs using the Adam optimizer, a batch size of 128, and a learning rate of 3×10^{-4} , requiring approximately 20 hours on a single H100 GPU. Fine-tuning was performed on a mixture of experimental and simulation data for 500 epochs, with the batch size reduced to 64 and the learning rate to 5×10^{-5} , requiring approximately 3 hours on a single H100 GPU. Inference used 1,000 diffusion steps and required approximately 20 seconds for a batch of 100 samples on an H100 GPU.

C Ablations

Ablations were performed to probe the impact the experiment/simulation training data ratio, round-trip loss weighting, freezing different components of the neural network, and fine-tuning with Low Rank Adaptation. Results below reflect mean train and test performance across the 6-fold cross-validation set.

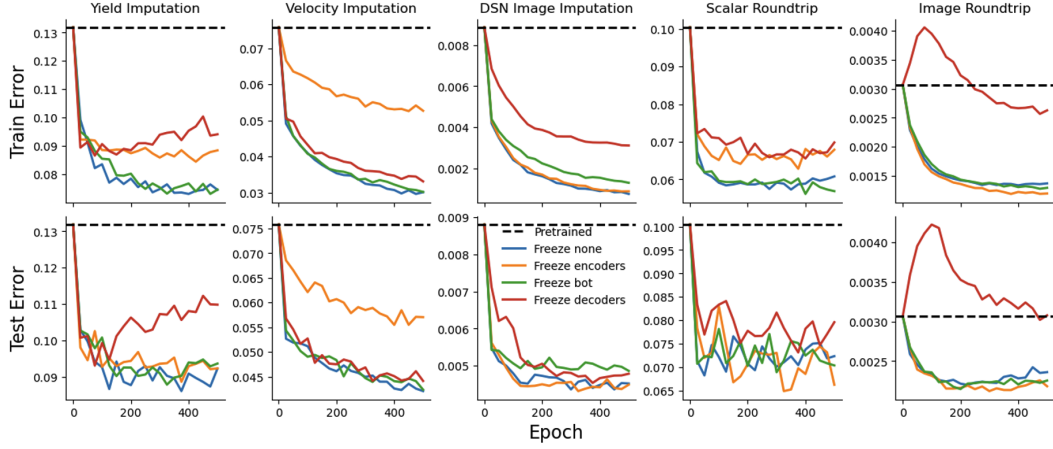


Figure A.1: Ablation over the set of model components updated during experimental fine-tuning. Training (top) and held-out test (bottom) errors are shown for yield imputation, velocity imputation, DSN image imputation, scalar round-trip prediction, and image round-trip prediction. We compare fine-tuning the full model with progressively freezing the encoder, decoder, or both. Dashed horizontal lines indicate the corresponding error of the pre-trained model.

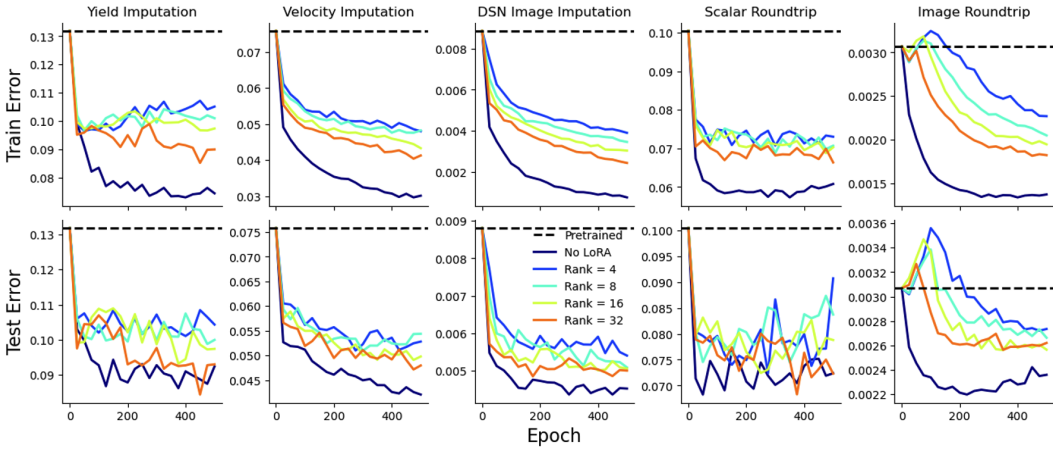


Figure A.2: Comparison of full fine-tuning with low-rank adaptation (LoRA). Training (top) and held-out test (bottom) errors are shown across the five evaluation tasks as a function of fine-tuning epoch. We compare standard fine-tuning without LoRA against LoRA with several ranks. Dashed horizontal lines indicate the corresponding pre-trained model error.

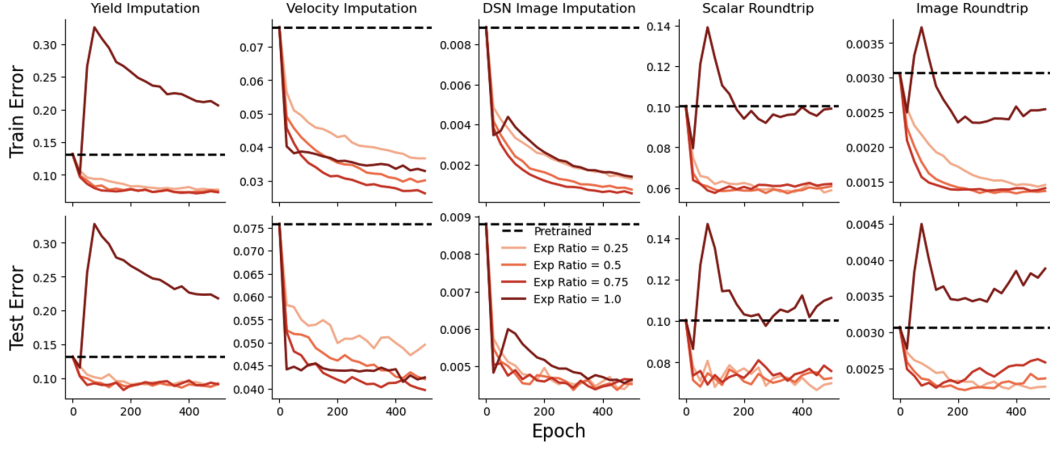


Figure A.3: Effect of the experimental-to-simulation data ratio during fine-tuning. Training (top) and held-out test (bottom) errors are shown for yield and velocity imputation, DSN image imputation, and scalar and image round-trip prediction. Curves correspond to different fractions of experimental samples relative to replayed simulation samples in each training batch. Dashed horizontal lines indicate the corresponding pre-trained model error.

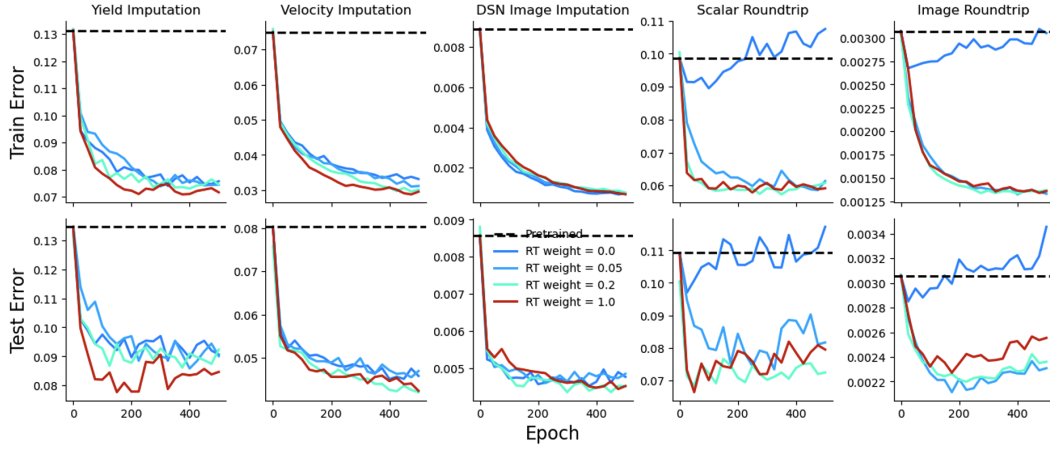


Figure A.4: Effect of the round-trip loss weight during experimental fine-tuning. Training (top) and held-out test (bottom) errors are shown for the five evaluation tasks for several round-trip loss weights. Increasing the round-trip weight improves scalar and image round-trip consistency, while excessively large weights can degrade performance on other objectives. Dashed horizontal lines indicate the corresponding pre-trained model error.

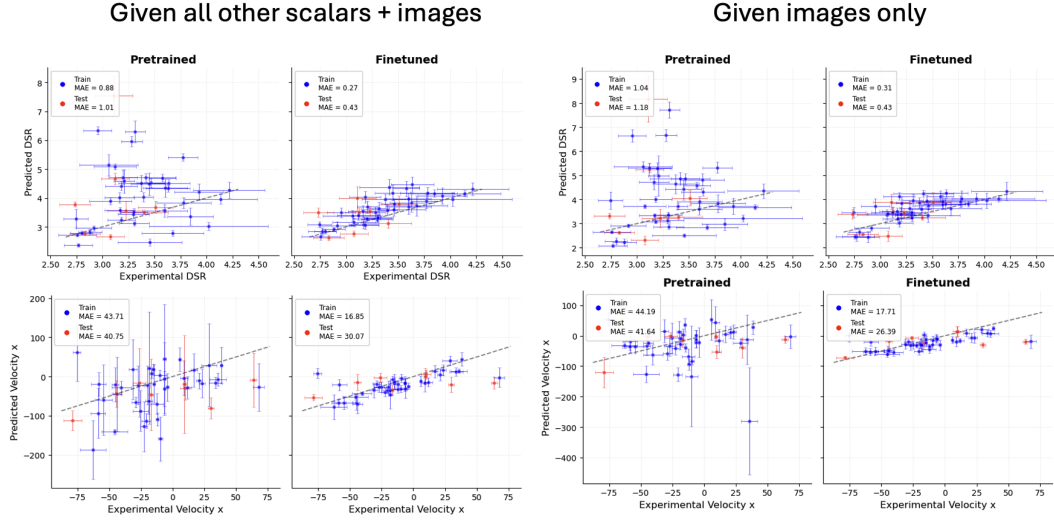


Figure A.5: Additional scalar imputation results before and after fine-tuning. Predictions are shown for DSR (top) and x-dimension of hot spot velocity (bottom), with experimental measurements on the horizontal axis and model predictions on the vertical axis. The left pair of columns conditions on all other available scalar and image diagnostics, while the right pair conditions only on images. Blue and red points denote training and held-out test experiments, respectively, with error bars indicating uncertainty. Fine-tuning substantially reduces prediction error in both conditioning settings and brings predictions closer to the dashed identity line.

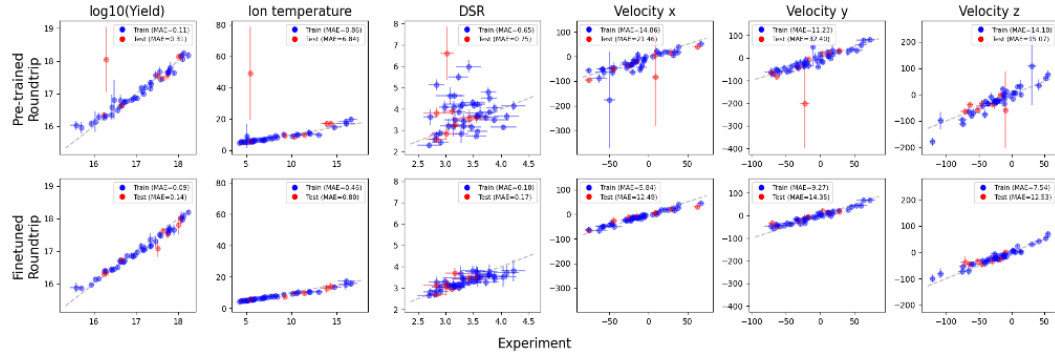


Figure A.6: Round-trip predictions for individual scalar experimental quantities before and after fine-tuning. The pre-trained model (top) and fine-tuned model (bottom) are evaluated by predicting experimental inputs through the model's round-trip procedure. Columns show representative quantities including yield, ion temperature, DSR, and three velocity components. Blue and red points denote training and held-out test experiments, respectively, and dashed lines indicate perfect agreement. Fine-tuning improves round-trip consistency across the scalar quantities, with particularly large reductions in error for DSR and velocity.