Towards a Multi-Modal Foundation Model for Inertial Confinement Fusion: Combining Structured Data and Diagnostic Images

Anonymous Authors[1]

Abstract

Inertial confinement fusion (ICF) offers a pathway to sustainable energy production, but achieving controlled fusion requires precise modeling of complex structured and image data. Recent breakthroughs underscore the need for scalable methods to analyze multi-modal diagnostic data and simulations, which include scalar inputs, scalar outputs, and image outputs. In this work, we present a diffusion-based generative framework designed to model the joint and conditional distributions of these structured and image data. By leveraging simulation data for pretraining, our approach addresses the challenge of experimental data scarcity and enables robust conditional modeling tasks. This work represents a prototype towards an ICF foundation model, and its architecture is transferable to diverse multi-modal scientific problems.

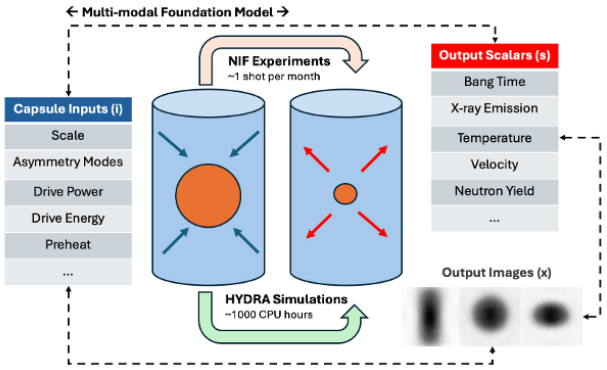


Figure 1. Structured capsule inputs (blue table) drive experimental and simulated capsule implosions that produce multimodal outputs – scalars (red table) and diagnostic images. A multi-model foundation model approach seeks to predict distribution of scalar outputs, images, or inputs conditioned on any other combination of information.

1. Introduction

Inertial confinement fusion (ICF) has emerged as a potentially viable path towards sustainable energy production via the nuclear fusion of hydrogen atoms into helium (Moses, 2008). However, achieving controlled fusion requires extremely high temperatures and pressures, along with precise control over energy inputs and symmetry. A fusion gain (ratio of output to input energy) > 1 was first achieved at the National Ignition Facility (NIF) in 2022, marking the first instance of "ignition" in a laboratory setting (Abu-Shawareb et al., 2024). Since then, NIF has broken its own record multiple times, but much higher yields must be reliably produced to power next-generation energy facilities (of Sciences Engineering & Medicine, 2021).

Experiments at NIF require extensive preparation and resources, and are limited to tens of high-yield tests per year; therefore, researchers rely heavily on radiation-hydrodynamics codes such as HYDRA (Marinak et al., 2001) to explore various design options (Figure 1). However, these capsule simulations can take thousands of CPU hours and are parameterized from high-dimensional structured inputs (i) that cannot be exhaustively explored. Experimental diagnostics produce sets of structured scalar (s) and image data (x) for each NIF experiment, and analogous outputs are computed from simulations. Capsule inputs are challenging to measure experimentally, and are often inferred from simulated inputs that correspond to similar outputs (Gaffney et al., 2019). In both experiment and simulation, there are numerous sources of uncertainty that must be mitigated in order to obtain robust results, and it is critical to understand the joint and conditional distributions of i, s, and x.

In this work, we prototype a foundation model for structured and image data in ICF, learning a joint distribution $p(i, s, x)$ of capsule inputs, scalar outputs, and X-ray images. Departing from previous models that rely on contrastive learning for multi-modal embeddings (Radford et al., 2021), we

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

propose a unified diffusion-based framework with multi-modal prediction heads and random masking for robust training across diverse scenarios. We demonstrate forward $p(s, x|i)$, inverse $p(i|s, x)$, and other bespoke conditional modeling tasks, providing a scalable and transferable generative framework for multi-modal data. Prior surrogate models predicted scalars and images deterministically (Anirudh et al., 2020), while MCMC methods predicted inputs from outputs without incorporating images (Gaffney et al., 2019). Although no direct comparison exists in prior ICF work, our framework unifies and generalizes these objectives. We also propose a fine-tuning method that iteratively refines input predictions under experimental uncertainties, enabling adaptation to limited data. Our key contributions are:

1. *We train a single generative model to learn the joint probability of structured and image data.*
2. *We introduce the first multi-modal model for ICF input prediction, transferable to other scientific tasks.*
3. *We demonstrate iterative fine-tuning on experimental data without access to ground-truth inputs.*

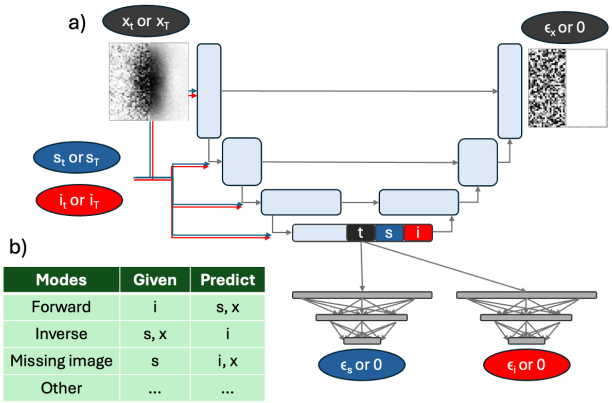
2. Methods

2.1. Training and testing data

Simulation inputs $i \in \mathbb{R}^9$ were sampled by latin-hypercube sampling; ∼90 thousand outputs $s \in \mathbb{R}^{10}$ and images $x \in \mathbb{R}^{48\times48}$ were generated by HYDRA simulation code (Marinak et al., 2001) as previously reported (Nora et al., 2017). A random 80/10/10 train/validation/test split was performed. All inputs and scalar outputs were normalized $[0, 1]$ followed by an inverse sigmoid transform $y = \log(x)/(1 - x)$ and images were normalized $[-1, 1]$. Experimental ICF data consists of 10 Deuterium-Tritium shots performed at NIF during the BigFoot campaign (Casey et al., 2018). Fine-tuning was performed and tested using a 7/3 split as specified by (Kustowski et al., 2022).

2.2. Model architecture and training

We employ a U-net (Ronneberger et al., 2015) architecture to jointly predict noise over inputs (i), output images (x), and output scalars (s) (Figure 2). Image noise is predicted via a series of encoding and decoding convolution layers, and structured noise is predicted via fully connected networks from the U-net bottleneck layer. For each training batch $X_T = [i_T, s_T, x_T]$, a set of masks $M = [M_i, M_s, M_x]$ with matching dimensions are sampled from a Bernoulli distribution. Diffusion times $t \sim \mathcal{U}(0, T)$ are sampled and X_t is computed from a linear schedule α_t with cumulative noise $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ (Ho et al., 2020). If a given modality is masked, then its true (unnoised) value X_T is passed to the model; otherwise the noised version X_t is passed. For each unmasked modality the corresponding output head is regressed against the true noise, whereas masked modalities are regressed against a zero vector. Further details on diffusion and U-net hyperparameters are in the Appendix.



Modes	Given	Predict
Forward	i	s, x
Inverse	s, x	i
Missing image	s	i, x
Other	...	...

Figure 2. a) Architecture of the multi-modal diffusion model. Noised input and output scalars condition U-net encoding/decoding layer. Each data modality is concatenated in the bottleneck layer, from which additional heads predict input and output noise. b) Table showing inference modes based conditioning (masked) modalities and predicted ones.

Algorithm 1 Training Loop

Input: Simulation data containing joint distribution $q(i, s, x)$ of all modalities, diffusion steps T, and untrained model f_θ, cumulative noise schedule $\bar{\alpha}$
for batch$= 1$ **to** N **do**
  $X_T \sim q(i, s, x)$
  $t \sim \mathcal{U}(0, T)$
  $\epsilon \sim \mathcal{N}(0, 1)$
  $M \sim$ Bernoulli(0.5)
  $X_t = \sqrt{\bar{\alpha}_t} X_T + \sqrt{1 - \bar{\alpha}_t}\epsilon$
  $X[j] = \begin{cases} X_t[j] & \textbf{if } M[j] = 0, \\ X_T[j] & \textbf{if } M[j] = 1. \end{cases}$
  $\hat{\epsilon} = f_\theta(X, M, t)$
  $L_\theta = \|\epsilon - \hat{\epsilon}\|$
  Take gradient step $\nabla_\theta L_\theta$
end for
return f_θ

2.3. Inference

During the inference stage we specify a prediction mode (forward, inverse, etc.) that determines which data modalities are fixed and which are denoised from an initially Gaussian distribution. As in a traditional diffusion model, we make noise prediction for T steps and iteratively re-

move noise from each unmasked modality. A conditional distribution, for example $p(s, x|i)$ for the forward mode, is determined by generating N samples each with varied initial noise profiles.

Algorithm 2 Inference Loop

1: **Input:** Conditioning modalities $X = [i, s, x]$, masks $M = [M_i, M_s, M_x]$, diffusion timesteps T, and noise schedule parameters $\alpha, \bar{\alpha}$

2: $X_0[j] = \begin{cases} \mathcal{N}(0,1) & \textbf{if } M[j] = 0, \\ X[j] & \textbf{if } M[j] = 1. \end{cases}$

3: **for** $t = 1$ **to** T **do**

4: $\quad \epsilon \sim \mathcal{N}(0,1)$

5: $\quad \hat{\epsilon} = f_\theta(X, M, t)$

6: $\quad X_t = \frac{1}{\sqrt{\alpha}}\left(X_{t-1} - \frac{1-\alpha}{\sqrt{1-\bar{\alpha}}}\hat{\epsilon}\right) + \sqrt{(1-\alpha)}\epsilon$

7: **end for**

8: **return** X_T

2.4. Fine-tuning on NIF experiments

Unlike simulations, NIF experiments yield uncertainties $\sigma_{s,\text{exp}}$ associated with each output scalar mean $\mu_{s,\text{exp}}$. We leverage these uncertainties to augment fine-tuning data by sampling 100 experimental scalars from each of our 7 train samples $s_{\text{exp}} \sim \mathcal{N}(\mu_{s,\text{exp}}, \sigma_{s,\text{exp}})$. Because there are no ground-truth experimental inputs, we use our pre-trained model to generate input predictions conditioned on the experimental distribution $\hat{i}_{\text{exp}} = f_\theta(s_{\text{exp}}, x_{\text{exp}})$. After one epoch of fine-tuning on these inputs and outputs $(\hat{i}_{\text{exp}}, s_{\text{exp}}, x_{\text{exp}})$, we repeat the procedure until we see convergence in the KL divergence in the experimental and roundtrip output distributions.

3. Results

3.1. Reducing Error with Multi-modal Conditioning

To evaluate our model in the forward mode $p(s, x|i)$, we generate an ensemble of 10 outputs for each test sample. We compare the mean of each distribution to true outputs and plot the mean error in Figure 3a (blue). Since outputs are well constrained by inputs, the error is consistently low. We then repeat the predictions using two alternative conditioning scenarios: input with output image $p(s|i, x)$ and output image only $p(s|x)$. The former (green) yields slightly improved or similar predictions compared to input-only conditioning, while the latter performs worse overall. As expected, this indicates inputs constrain simulation outputs better than images do, except when predicting hot spot radius, which depends more on image geometry.

A more challenging test is the inverse mode $p(i|s, x)$. Very different simulation inputs can produce similar outputs, often making this task unconstrained. We compare input predictions given scalars $p(i|s)$ (blue) or images $p(i|x)$ (red) only, and find that error varies significantly depending on the conditioning data. For example, asymmetry modes 1 and 2 are much better predicted from image data, while the preheat and scale of the implosion are better determined from scalars. Encouragingly, we find that combined conditioning (green) always outperforms single-modality predictions. For some inputs – such as drive trough adjustment and dopant fraction – there is significant improvement over single modality conditioning, indicating synergetic constraints from the scalars and images.

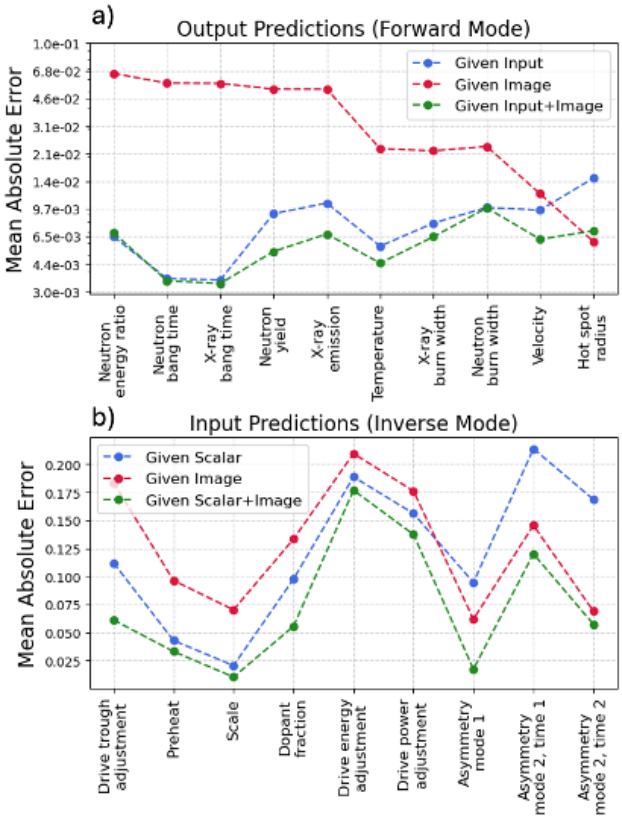


Figure 3. a) Mean output prediction accuracy on test set given conditioning on inputs and/or images b) Mean input prediction accuracy on test set conditioned on output scalars and/or images.

3.2. Inverse Modeling Constraints and Self-Consistency

Next, we select a random simulation test sample and generate a distribution of 500 inputs using the three conditioning scenarios described above. In Figure 4a, we show these distributions for five inputs, each normalized to $[0, 1]$. We note that certain inputs, such as scale and dopant fraction, are more tightly constrained by scalars alone (blue), while asymmetry is again more constrained by image data (red).

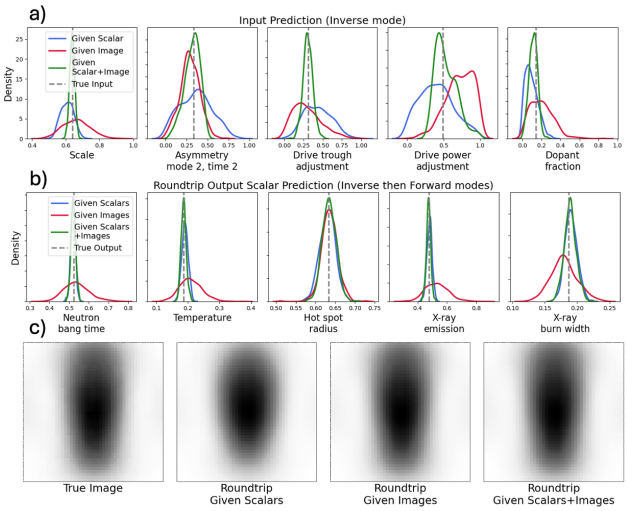


Figure 4. a) Input distributions predicted from a randomly sampled test output scalar and/or image. True inputs shown by dashed gray line b) Roundtrip output prediction given the input distributions shown in(a). True output shown by dashed gray line. c) True image and mean of sampled images from each roundtrip model in (b).

Interestingly, for both drive trough adjustment and power adjustment, the means of $p(i|s)$ and $p(i|x)$ are shifted in opposite directions relative to the ground truth, while the $p(i|s,x)$ mean (green) is much more accurate. This reveals that missing data modalities can systematically shift the predicted input distribution. For each input, the entire $p(i|s,x)$ distribution is contained within both the $p(i|s)$ and $p(i|x)$ distributions, indicating the model is successfully capturing conditional dependencies. Additional test inputs and outputs are included Appendix Figures 6-7.

To evaluate the quality of our input distributions and the self-consistency of the model, we use the forward mode $p(s,x|i)$ to project our 500 predicted inputs back into outputs. As expected, the combined conditioning performs best on this "round-trip" test, with distributions sharply peaked near the ground truth outputs. The scalar-only inputs also perform well but with increased uncertainty reflecting the loss in image constraints. Round-trip distributions for $p(i|x)$ inputs show that images alone provide insufficient conditioning, and only certain outputs, such as hot spot radius, are correctly predicted by the forward model.

While we focus our analysis on structure data, we also perform round-trip image generation and find strong reconstruction when images were included as conditional information for the inputs. Images generated from $p(i|s)$ have similar symmetries as the ground truth but are noticeably worse reconstructions than image-conditioned inputs $p(i|x)$ and $p(i|x,s)$. Image generation is generally robust and will be the subject of future analysis.

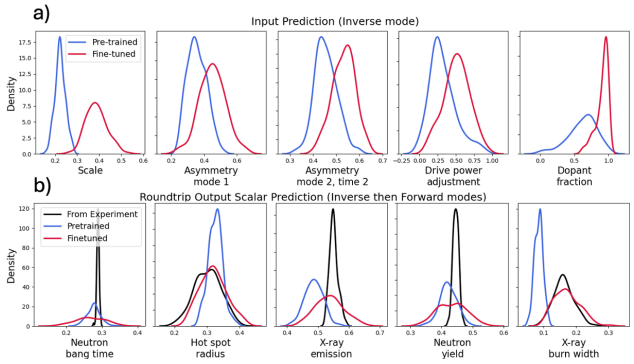


Figure 5. a) Input distributions predicted from experimental outputs via pretrained and finetuned model b) Roundtrip output prediction of pretrained and finetuned input distributions compared to ground-truth experimental distributions (black)

3.3. Fine-tuning improvement experimental predictions

Lastly, we evaluate our fine-tuned model on our experimental test samples. In Figure 5a, we sample 500 sets of outputs based on the mean and uncertainty a single experimental shot (predictions for all three test shots shown in Appendix Figures 9-10) and compute $p(i|s,x)$ for the same five inputs shown in Figure 4a. These distributions are computed using both our pre-trained model (blue) and the model fine-tuned for 10 epochs on experimental data. We observe a shift in the input distributions for each case, with minimal overlap for certain inputs, such as scale and dopant fraction, indicating that the joint distributions for these values varies significantly between simulation and experiment.

Since there are no ground-truth inputs to compare against, we evaluated these inputs by performing a roundtrip analysis and comparing the resulting output distributions to the experimental ground-truth distributions. For most outputs, fine-tuning significantly improved agreement with experimental data. For example, the KL divergence for X-ray burn width decreased from 10.4 to 0.4 after fine-tuning. However, certain outputs, such as neutron bang time, showed worse agreement, with a slight increase in KL divergence from 1.1 to 2.4. These results suggest that while fine-tuning improves self-consistency, seven experimental samples may be insufficient to fully update the joint distribution and certain trade-offs must be made to fit the data.

4. Conclusions

We present a generative foundation model for ICF that unifies structured and image data and enables robust predictions of capsule inputs, scalar outputs, and diagnostic images. Our approach learns joint distributions across all modalities, produces arbitrary conditional distributions, and supports iterative fine-tuning on experimental data.

References

Abu-Shawareb, H., Acree, R., Adams, P., Adams, J., Addis, B., Aden, R., Adrian, P., Afeyan, B., Aggleton, M., Aghaian, L., et al. Achievement of target gain larger than unity in an inertial fusion experiment. *Physical review letters*, 132(6):065102, 2024.

Anirudh, R., Thiagarajan, J. J., Bremer, P.-T., and Spears, B. K. Improved surrogates in inertial confinement fusion with manifold and cycle consistencies. *Proceedings of the National Academy of Sciences*, 117(18):9741–9746, 2020.

Casey, D., Thomas, C., Baker, K., Spears, B., Hohenberger, M., Khan, S., Nora, R., Weber, C., Woods, D., Hurricane, O., et al. The high velocity, high adiabat,"bigfoot" campaign and tests of indirect-drive implosion scaling. *Physics of Plasmas*, 25(5), 2018.

Dome272. Diffusion-models-pytorch. https://github.com/dome272/Diffusion-Models-pytorch, 2023. Accessed: 2025-05-15.

Gaffney, J. A., Brandon, S. T., Humbird, K. D., Kruse, M. K. G., Nora, R. C., Peterson, J. L., and Spears, B. K. Making inertial confinement fusion models more predictive. *Physics of Plasmas*, 26(8):082704, 2019.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.*, 33:6840–6851, 2020.

Kustowski, B., Gaffney, J. A., Spears, B. K., Anderson, G. J., Anirudh, R., Bremer, P.-T., Thiagarajan, J. J., Kruse, M. K., and Nora, R. C. Suppressing simulation bias in multi-modal data using transfer learning. *Machine Learning: Science and Technology*, 3(1):015035, 2022.

Marinak, M. M., Kerbel, G. D., Gentile, N. A., Jones, O. S., Munro, D. H., Dittrich, T. R., and Haan, S. W. Three-dimensional hydra simulations of national ignition facility targets. *Physics of Plasmas*, 8:2275–2280, 2001.

Moses, E. I. The national ignition facility (nif): A path to fusion energy. *Energy Conversion and Management*, 49 (7):1795–1802, 2008.

Nora, R., Peterson, J. L., Spears, B. K., Field, J. E., and Brandon, S. Ensemble simulations of inertial confinement fusion implosions. *Stat. Anal. Data Min.*, 10(4):230–237, 2017.

of Sciences Engineering, N. A. and Medicine. *Bringing Fusion to the U.S. Grid*. The National Academies Press, 2021. ISBN 978-0-309-68538-2.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

A. Appendix

A.1. Diffusion Model

The diffusion architecture is adapted from (Dome272, 2023), which is based on the DDPM framework of (Ho et al., 2020). The model employs 12 convolutional channels and consists of 3 encoding and 3 decoding layers. A total of 1000 diffusion steps are used with a linear noise schedule ranging from $\beta = [0.0001, 0.02]$.

Diffusion time is projected into a 48-dimensional space via positional encoding. Input and output scalars are embedded into the same dimension and concatenated with the time embedding and three mask tokens to indicate noised modalities. The combined embedding is reshaped to match the output dimensions of each encoding and decoding layer and added to the layer outputs.

Noise predictions for input and output are handled by separate fully connected heads derived from the flattened bottleneck layer. Each head includes a single hidden layer with 100 dimensions, followed by output layers of 9 dimensions for input noise and 10 dimensions for output noise. Sigmoid Linear Units (SiLU) are used as activation functions throughout the architecture.

A.2. Training Parameters

The model is trained using a learning rate of 0.0003 and a batch size of 64. An exponential moving average (EMA) is applied to the model weights with a smoothing factor of 0.995. Training is conducted for 200 epochs, requiring approximately 12 hours on a single GPU. Prediction error for forward and inverse modeling tasks converged at the end of training.

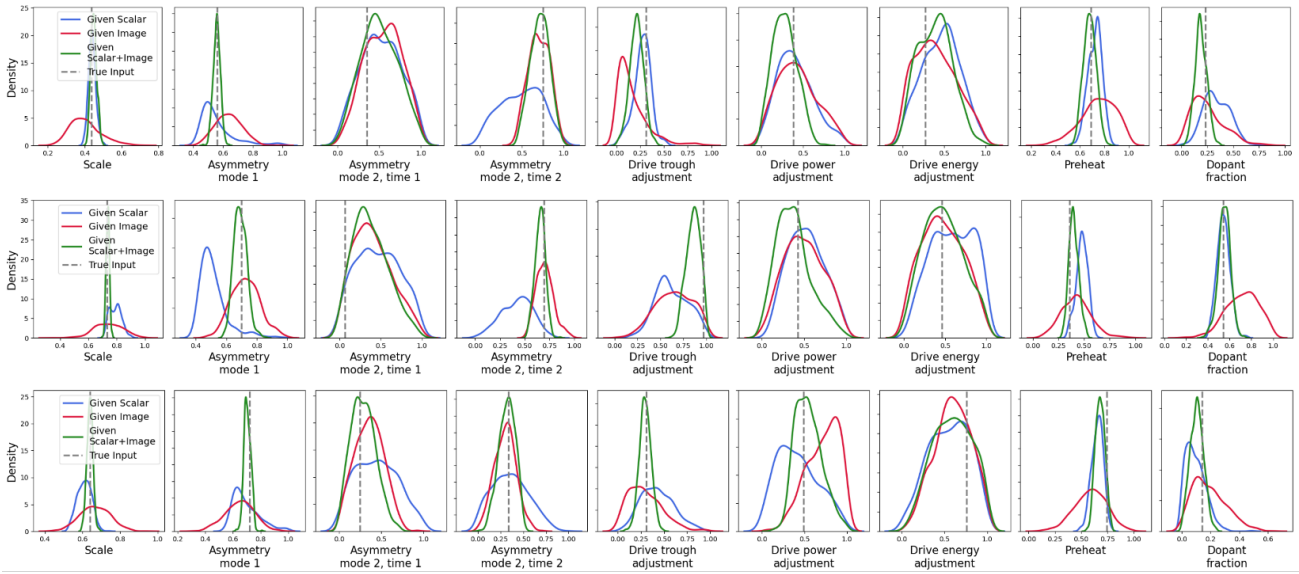


Figure 6. a) All input distributions predicted from three randomly sampled test output scalar and/or image.

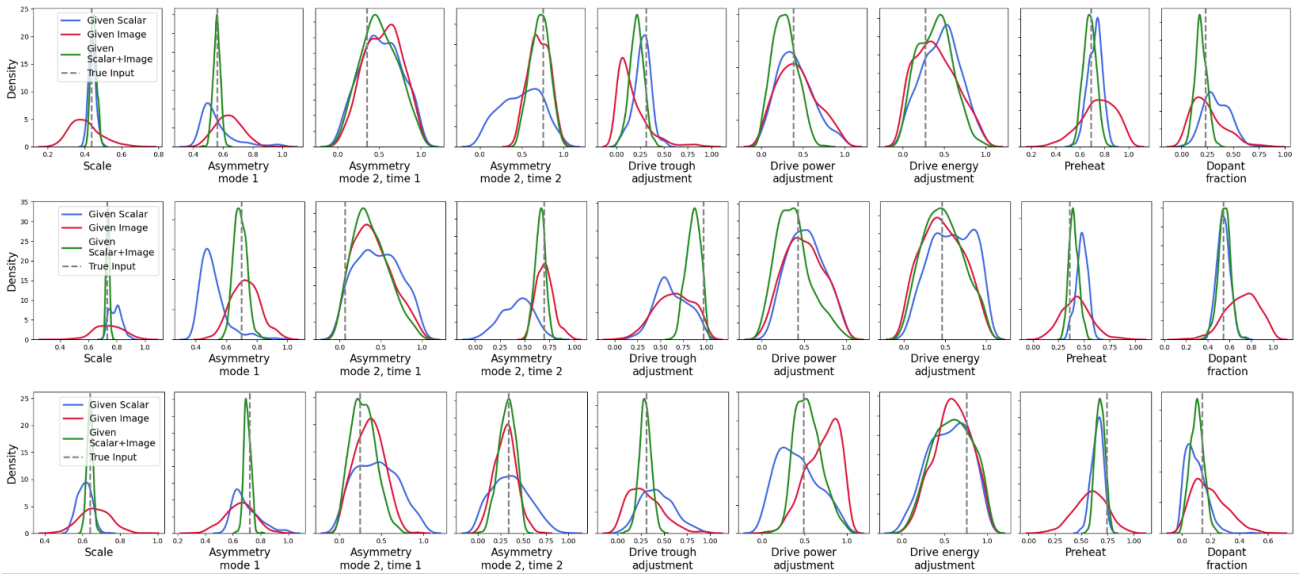


Figure 7. b) Roundtrip output prediction given the three sets of input distributions shown above

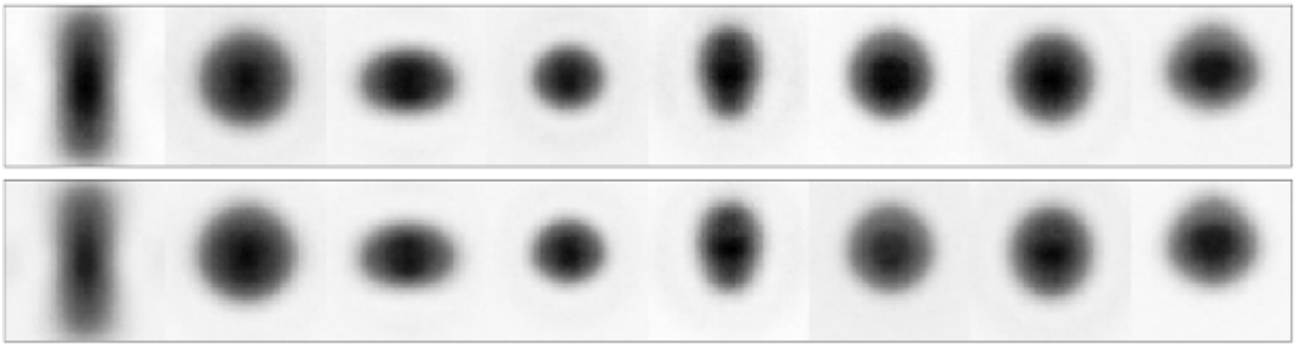


Figure 8. Ground truth images (top row) and generated images (given input conditioning only) for eight samples in the simulation test set. Agreement is consistently strong across varied image geometries and structured inputs.

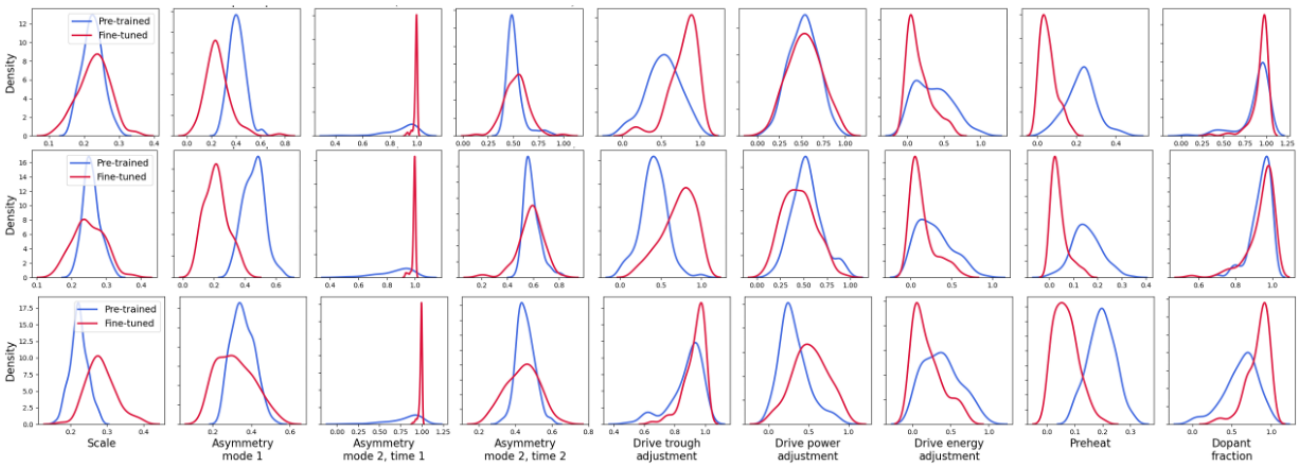


Figure 9. All input distributions predicted from experimental outputs via pretrained and finetuned model for all three test shots, where is row is a different shot.

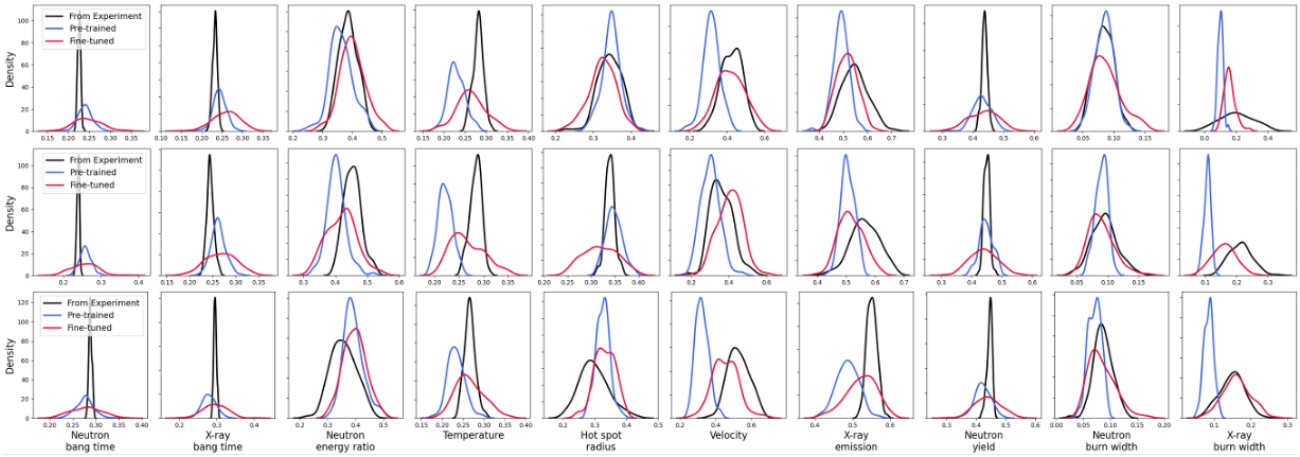


Figure 10. Roundtrip output prediction of pretrained (blue) and finetuned (red) input distributions shown above compared to ground-truth experimental distributions (black).

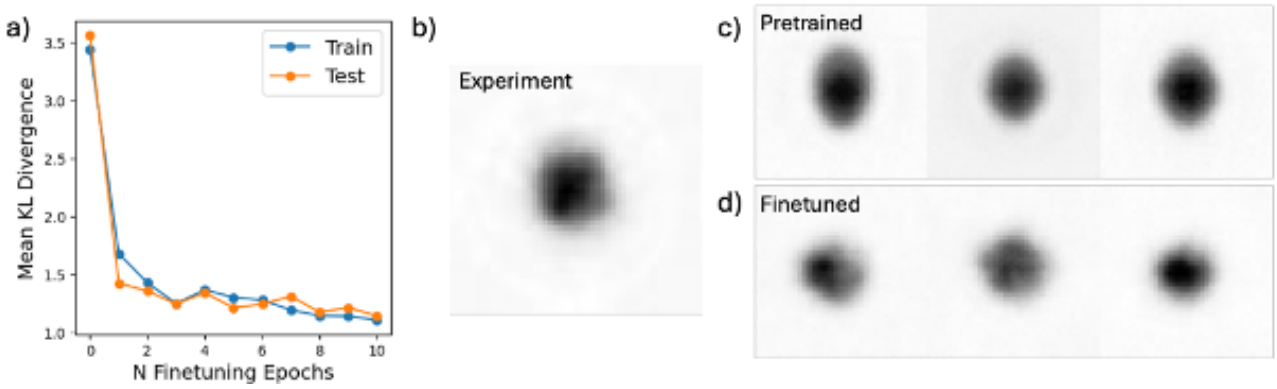


Figure 11. a) Mean KL divergence of roundtrip and ground-truth outputs for train and test sets during fine-tuning. b) Reference images from NIF test set. c) Generated roundtrip samples from pre-trained model showing consistent symmetries d) Generated roundtrip samples with finetuned model showing breaking of symmetries more consistent with experiment