Joint Diffusion Approach to Multimodal Inference in Inertial Confinement Fusion

Michael Jones [1,*] Justin Kunimune [2] Daniel Casey [1] Bogdan Kustowski [1] Eugene Kur [1] and Kelli Humbird [1]
[1] *Lawrence Livermore National Laboratory, Livermore, California 94550, USA*
[2] *Plasma Science and Fusion Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

(Received 20 February 2026; accepted 17 July 2026; published 18 August 2026)

A combination of physics-based simulation and experiments has been critical to achieving ignition in inertial confinement fusion (ICF). Simulation and experiment both produce a mixture of scalar and image outputs; however, only a subset of simulated data are available experimentally. We introduce a generative framework, called JointDiff, which enables predictions of conditional simulation input and output distributions from partial, multimodal observations. The model leverages joint diffusion to unify forward surrogate modeling, inverse inference, and output imputation into one architecture. We train our model on a large ensemble of three-dimensional multirocket piston simulations and demonstrate high accuracy, statistical robustness, and promising transfer to experiments performed at the National Ignition Facility, with further improvements achieved through experimental fine-tuning. This work establishes JointDiff as a flexible generative surrogate for multimodal scientific tasks, with implications for understanding diagnostic constraints, aligning simulation to experiment, and accelerating ICF design.

DOI: 10.1103/dldf-7tfm

I. INTRODUCTION

Inertial confinement fusion (ICF) has emerged as a leading approach for controlled nuclear fusion [1,2], providing fundamental plasma physics insights and a potential pathway toward alternative energy. Scientific progress in ICF relies on a close interplay between simulation and experiment, in which each is iteratively refined based on data from the other. Machine learning models can enhance this interplay by acting as fast surrogates for simulations and by identifying input conditions that reproduce observed experimental outputs [3–6]. In practice, both simulations and experiments often provide only partial views of the system: experiments yield sparse and heterogeneous diagnostics, while simulations, though richer and more uniform, represent lower fidelity outputs. Moreover, these outputs are inherently multimodal—a combination of scalars and images—making it challenging to learn consistent conditional relationships across subsets of observed quantities [7]. This motivates models that can represent joint distributions over all observables and support inference under arbitrary patterns of missing data.

Multimodal generative models have become powerful tools across scientific domains, enabling advances in biomolecular structure prediction, medical diagnostics, and materials design [8–10]. These models capture complex relationships across heterogeneous data types and often employ generative decoders such as denoising diffusion probabilistic models (DDPMs) [11,12] to generate high-quality samples. Multimodal prediction is typically achieved either by combining data into a shared latent space prior to diffusion [13,14] or by aligning preexisting latent representations using contrastive learning techniques [15]. While contrastive approaches allow the reuse of pretrained models without exhaustive retraining, they may fail to capture subtle intermodal dependencies by not explicitly modeling the full joint distribution [16].

An alternative strategy is to train a generative model directly on the joint distribution of multimodal data. This approach has gained traction in domains such as generative biology, where diffusion models jointly update continuous atom positions and discrete atom types [17–19]. It has further been shown that joint diffusion is feasible and theoretically well founded in arbitrary state spaces [20,21]; however, the application of joint diffusion models as physics-based surrogates has been underexplored.

We propose JointDiff, a model architecture and training scheme that leverages joint diffusion as a generative surrogate model for ICF. We train our model on a large ensemble of three-dimensional (3D) multi-rocket piston (RP) simulations [22] and demonstrate accuracy in forward and inverse modeling tasks as well as imputation of missing outputs. We show that JointDiff produces meaningful conditional distributions from partial multimodal observations, enabling tunable uncertainty quantification and investigation of input sensitivity to missing diagnostics. We demonstrate reasonable transferability to experiments at the National Ignition Facility (NIF), and we show further improvement to predictive performance after short experimental fine-tuning.

Deep learning models have previously been applied as multimodal ICF surrogates [3,4] and techniques such as Markov chain Monte Marlo (MCMC) have been used to reconstruct simulation inputs from experimental outputs [5–7]; however, implementing MCMC with multimodal data is

*Contact author: jones313@llnl.gov

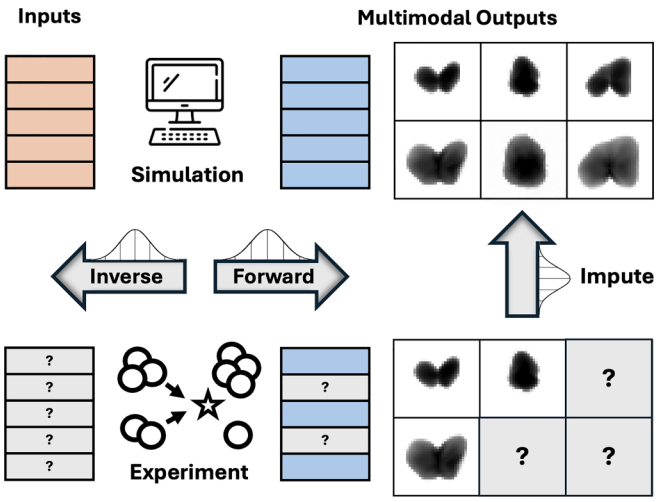

FIG. 1. JointDiff is generative surrogate for forward, inverse, and imputation tasks. Simulations (upper row) produce a complete set of multimodal outputs given a complete set of inputs. Surrogate models are trained to replace simulations and perform the forward or inverse prediction tasks. Experiments (bottom row) produce incomplete sets of outputs, and inputs to the capsule are unknown. In addition to forward and inverse surrogate capabilities, JointDiff predicts conditional distribution of inputs and outputs given partial multimodal observations.

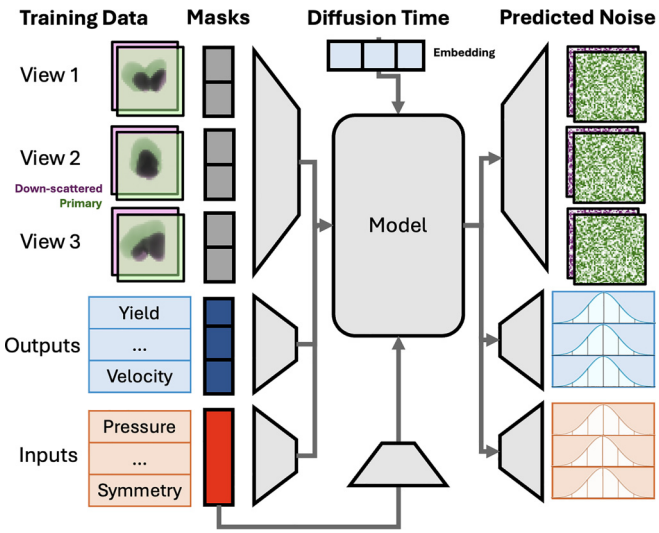

FIG. 2. Architecture of the JointDiff model. Output images contain primary (green) and down-scattered (pink) neutron intensities, which are encoded as separate image channels. Trapezoids represent data encoders and decoders, which are convolutional networks for images and fully connected networks for scalar data. Gray and blue boxes correspond to discrete masks that inform the model whether true or noisy data are provided for each modality and are embedded by a fully connected network. Diffusion time is encoded by a sin/cos positional embedding. The time, mask, and input/output embeddings are concatenated to the image embeddings at each convolution step. Individual decoders predict noise to be removed at diffusion time t for each image and set of input and output scalars.

sensitive to choice of priors, computationally intensive, and may fail to converge. DDPMs have been applied as surrogates for particle-in-cell simulations [23], which are crucial to understand fundamental plasma interactions during ICF implosions. A preliminary multimodal application showed that DDPMs can model two-dimensional (2D) radiation hydrodynamics HYDRA simulations, capturing the joint relationship between scalar and single-channel image outputs and reproducing one-to-many inverse mappings to inputs [24]. In this work, we generalize and expand this framework to an ensemble of 440 000 simulations of 3D ICF implosions using a multimodal U-Net-based architecture amenable to arbitrary image views, image channels, and scalar inputs and outputs (Fig. 2). This introduces significantly greater complexity, as the model must learn conditional distributions across a combinatorially larger set of modality configurations.

We evaluate JointDiff on three core prediction tasks: (1) surrogate modeling (simulation inputs → outputs), (2) inverse modeling (outputs → inputs), and (3) imputation of missing outputs (partial outputs → complete outputs and inputs). Each task (shown in Fig. 1) addresses a critical component of the ICF design process—accelerating simulations for new input conditions, inferring inputs for previous experiments, and enriching available outputs—leading to the development of higher-yield and more robust experiments. We perform round-trip consistency tests to demonstrate that the learned distributions remain stable and self-consistent even when a large fraction of simulation data is masked. Finally, we evaluate the transferability of our approach using recent ICF experiments performed at the NIF, which contain only partial scalar and image diagnostics relative to simulations. Our model's round-trip predictions reconstruct most experimental observables, with discrepancies in specific scalar and image features lending insight into limitations of the underlying RP physics model. Furthermore, we leverage experimental fine-tuning to significantly enhance the model's ability to impute missing experimental scalars and images under various conditional scenarios. Together, these results show that JointDiff is a robust and flexible framework for predicting conditional distributions in ICF, and, more broadly, in scientific domains characterized by partial multimodal observations.

II. RESULTS

A. Joint diffusion enables multimodal conditional prediction

To predict conditional distributions given multimodal (scalar and image) data, we use a joint diffusion objective [15,20]. During training, noise is gradually added to both scalars and images and a single architecture learns to "denoise" both modalities at the same time. We apply random masking over inputs and outputs such that the model learns the distribution of outputs given inputs (forward model) as well as inputs given outputs (inverse model). Output masks are varied individually over scalars and images, enabling prediction of inputs and missing outputs conditioned on partial observations. When making predictions, the model is guided by whatever data are available, and the prediction targets are masked. The model architecture is shown in Fig. 2 and additional details are provided in Sec. IV.

We train our model on over 440 000 simulations generated by an RP physics model. Although the RP model is

lower fidelity than full radiation-hydrodynamics codes such as HYDRA [25], it has been shown to reasonably reproduce key observables, including asymmetries, hot spot velocity, bang time, and neutron images generated by HYDRA [22]. The model builds on prior work [26,27], which demonstrates that the implosion can be decomposed into discrete segments to capture the effects of 3D asymmetries while preserving the essential dynamics of stagnation. Crucially, the RP model enables the generation of large-scale 3D simulation datasets spanning a high-dimensional input space, which remains computationally infeasible with high-resolution HYDRA simulations. This 3D ensemble provides access to synthetic neutron images that are aligned to the same lines of sight used at NIF. Additional details on the RP model are provided in Sec. IV and in Refs. [7,22]. Each simulation training example contains a set of 28 scalar inputs, 12 output scalars, and 3 image line of sight each with 2 energy channels. The input scalars represent parameters such as initial pressure, adiabat, and drive symmetry modes that define each simulation. Scalar outputs include quantities that can be directly compared to experiment, such as the total neutrons generated by the implosion (yield) and the time of peak neutron production (bang time), as well as inferred quantities that cannot be measured experimentally such as the areal density (ρR) and residual kinetic energy (RKE), which are measures of hot spot confinement and energy not coupled to the hot spot, respectively. The images correspond to three different views of the implosion, with each image containing two channels: one for higher-energy (primary) neutrons and one for lower-energy (down-scattered) neutrons. A complete list of inputs and outputs can be found in Tables S1 and S2 of the Supplemental Material [28].

B. JointDiff is an accurate surrogate for rocket-piston simulations

We first evaluate JointDiff by predicting simulation outputs given scalar inputs for 986 test samples excluded from training. In Fig. 3(a), we show mean and standard deviations given ten output predictions for each test sample (additional velocities and inferred outputs shown in Fig. S1 of the Supplemental Material [28]). JointDiff achieves excellent prediction accuracy, with R^2 values ranging from 0.935 to 0.999 across scalars. On average, 92.8% of true samples fall within two standard deviations of the predicted distribution, varying from 84.7% to 97.4% across scalars. This is close to the 95.5% expected for a Gaussian distribution; however, we emphasize that predicted distributions are not Gaussian and calibration varies by scalar. Still, we observe a significant majority of ground-truth outputs fall within predicted distribution. Additionally, JointDiff can help identify outlier samples. Notably, the four samples with the largest yield prediction errors also correspond to those with the highest yield uncertainty. This demonstrates how predicted distributions and large error bars can flag less trustworthy predictions, in contrast to deterministic models that cannot directly quantify uncertainty. Additionally, targeted sampling and training in these regions can further reduce error.

We next evaluate the model's ability to generate accurate neutron images conditioned on simulation inputs. In Fig. 3(b), we compare ground-truth primary neutron images (View 1) to the mean of ten images generated for five test set samples. Each test image is randomly sampled from a different quintile of total neutron yield, which spans 3 orders of magnitude and results in distinct intensity profiles. Despite the diversity in yield and 3D geometry, the model consistently generates detailed and visually similar images across the sampled range. When analyzing the mean absolute error (MAE) for each generated image compared to the ground truth, we observe that the largest pixelwise errors occur at the boundaries of the intensity profiles. Since these intergeneration discrepancies are not apparent in the mean images, they likely correspond to regions of higher variance that can be used to identify regions of lower confidence in the predicted images.

TABLE I. Imputing missing scalar outputs given all other scalars and image outputs. Standard deviations are computed across ten generations for each sample in the test set.

Output	R^2	Within 2σ
log(yield)	1.000	95.5
Ion temp	1.000	97.9
Bang time	0.932	87.8
Burn width	1.000	92.5
DSR	0.999	95.1

C. JointDiff predicts inputs and missing outputs given multimodal conditioning

We test the capabilities of JointDiff when guided by multimodal output scalars and image information, focusing on two tasks: imputing missing scalars and predicting input distributions. Imputation is particularly relevant to scientific applications where simulation outputs may be unobservable in experiments or intermittently unavailable due to measurement challenges. This is a common occurrence in ICF experiments, where particular diagnostics might fail, are blocked by ride-along experiments, or are physically incapable of resolving simulation outputs. In Table I, we evaluate predicted distributions for five outputs that were left out (masked) during inference. In addition to partially conditioned generation, we find that the JointDiff model can produce fully unconditioned samples that are physically meaningful and consistent with the training distribution (Fig. S3 of the Supplemental Material [28]). Similar to the outputs analysis above, we obtain standard deviations over ten diffusion model generations for each test sample, and we observe similar or slightly improved accuracy as well as similar calibration to forward model predictions shown in Fig. 3. Improved performance compared to the forward model is likely due to stronger correlation between outputs that serve to constrain the missing values. We note that bang time has the lowest accuracy and highest uncertainty in both cases, indicating more noise in the simulated diagnostic or lower correlation with other outputs.

For the same test set, we mask inputs and guide conditioning with all output scalars and images. In Table II, we show predicted distributions for five input scalars (all scalars shown in Fig. S2 of the Supplemental Material [28]) compared to their ground-truth values. Again, we observe strong accuracy across scalars, with R^2 ranging from 0.967 to 0.999, and find

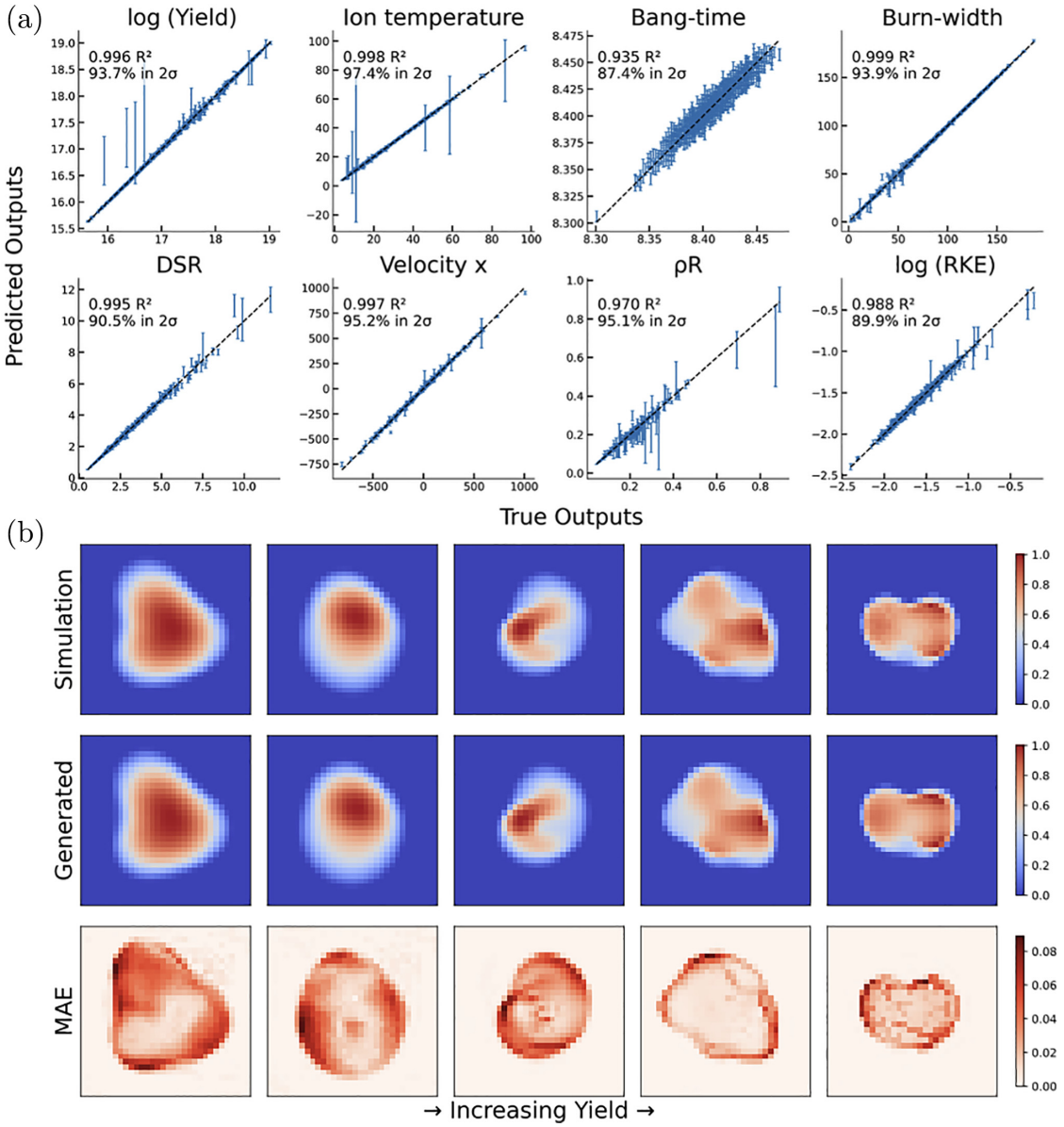

FIG. 3. JointDiff predicts accurate distributions of RP simulation outputs. (a) For each set of test inputs, ten output predictions are made. Error bars show the mean and standard deviation across these ten predictions. R^2 metrics are computed between the predicted means and the simulated ground truth. The percent of ground-truth samples within two standard deviations of the predicted distribution is also shown for each scalar. A subset of eight outputs is shown with all inputs in Fig. S1 of the Supplemental Material [28]. (b) Primary neutron images (View 1) sampled from each quintile of yield in the test data. The first row shows the ground-truth simulated images, the second row shows the mean model prediction across generated samples, and the third row shows MAE between each generated sample and the ground truth.

that a majority of predicted standard deviations include the ground-truth scalar, with a mean of 93.7% within two standard deviations. For both the inverse and forward modeling tasks, we perform ablations with task-specific and deterministic variants of JointDiff and show comparable performance (Table S3 of the Supplemental Material [28]).

In Fig. 4, we test the inverse model's robustness and sensitivity by removing partial image information. For six inputs (all inputs shown in Fig. S4 of the Supplemental Material [28]), we show the prediction MAE when removing each individual image view, as well as all primary or all down-scattered neutron images. In most cases, there is minimal change to the MAE when partial information is removed; however, we find that particular inputs are sensitive to the removal of particular images. For example, removing View 3 increases the $l = 2$, $m = 1$ symmetry error by $10\times$, while having minimal impact on $l = 2$, $m = -1$. Removing View 1 has a similar impact on $l = 2$, $m = -1$ while preserving the accuracy of $l = 2$, $m = 1$. These inputs represent diagonal drive symmetry features, and this analysis reveals that they cannot both be resolved without access to both equatorial views (1 and 3). Furthermore, we find that excluding down-scattered images has the greatest effect on adiabat and P2 swing predictions. This is likely because down-scattered images offer the most insight into the confining shell's profile—the adiabat sets the shell's density, and the P2 swing strongly changes the shell's density distri-

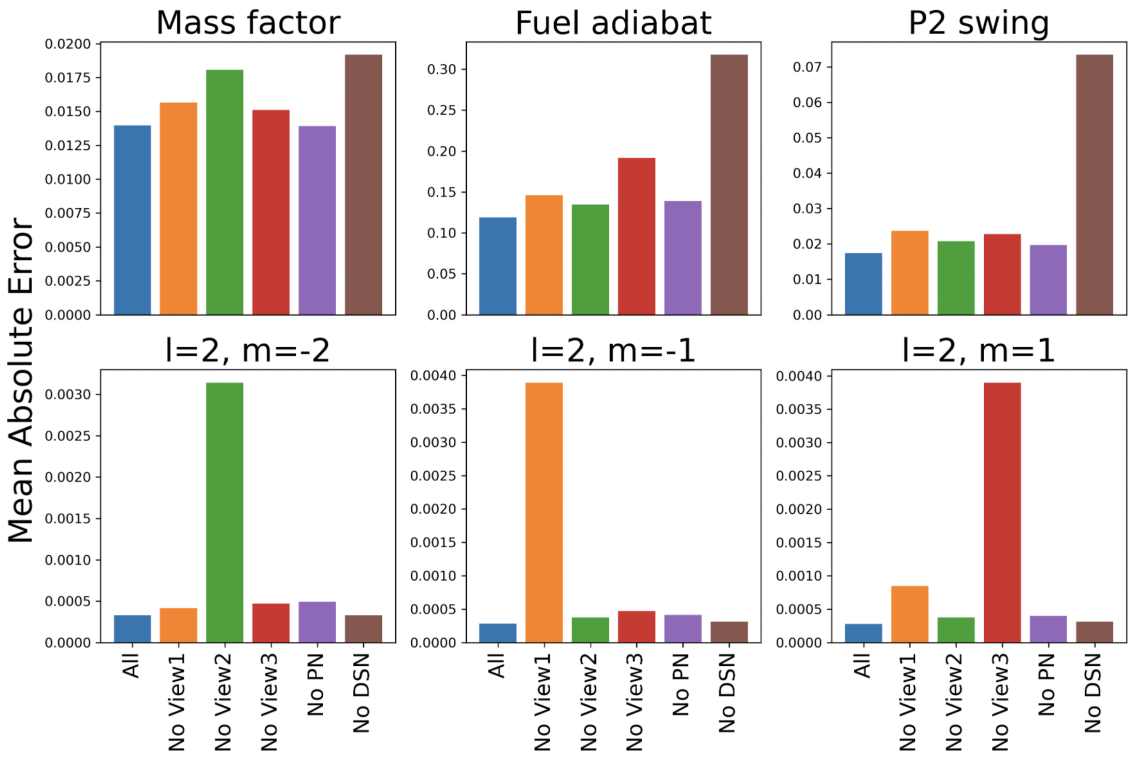

FIG. 4. Sensitivity of input predictions to removal of output image views and channels. Mean average error for six inputs when providing all outputs (blue) and when removing partial image information (each view and each channel). All inputs are shown in Fig. S4 of the Supplemental Material [28].

bution. These correlations are important to consider when (1) gauging uncertainty for experiments with missing diagnostics and (2) designing new diagnostic capabilities at the NIF and future facilities.

D. Conditional input distributions remain self-consistent in the absence of multiple outputs

Next, we test the inverse model's ability to adapt and predict inputs when a larger fraction of output data are unavailable. In this test, we specifically remove outputs that are regularly unavailable in NIF shots. For example, in NIF experiments, View 3 is often blocked by other diagnostics and the View 2 (polar) detector cannot capture down-scattered neutrons as the camera's line of sight is not long enough to separate neutron energies in time [29]. Accordingly, we mask both channels of View 3 as well as the down-scatter channel of View 2. Furthermore, 4/12 scalars in our simulation dataset, measuring ρR and RKE quantities, are inferred outputs that have no direct experimental equivalent. Burn-width measurements can also be challenging at high yield where burn duration is comparable to detector resolution, and the RP bang time tends to be systematically lower than experiments [22]. Therefore, we remove half (6/12) of scalar outputs and half (3/6) of image channels, and we test the model's ability to predict meaningful capsule input distributions. We emphasize that this model was not retrained or fine-tuned for this specific mask configuration and it therefore retains the forward and inverse modeling capabilities showcased above.

TABLE II. Predicting simulation inputs from scalar and image outputs (inverse model). Drive symmetry is averaged across all 23 symmetry modes. Standard deviations are computed across ten generations for each sample in the test set.

Input	R^2	Within 2σ
Mass factor	0.969	96.7
Alpha factor	0.984	96.2
Fuel adiabat	0.987	97.9
P2 swing	0.987	98.7
Symmetry	0.991	95.8

In Fig. 5(a), we show a distribution of inputs given 500 generations, all conditioned on a single set of simulation outputs. We highlight all $l = 2$ symmetry modes as these are main contributors to performance degradation at NIF. We compare distributions generated from full knowledge of scalar and image outputs (blue) to those generated from partial knowledge (orange). While there exists a clear mapping for most outputs to inputs when all information is provided, many inputs may satisfy the subset of partially observed outputs. We test if the diffusion model can capture the broader input distributions in these scenarios, while still constraining inputs that are fully described by output subset. For inputs such as the P2 swing and most symmetry modes, these distributions look very similar, indicating that adding the missing scalars and image views does not enhance confidence in input predictions. Other quantities, such as the initial pressure and alpha factor, have wider distributions with slightly shifted means when partial information is given, indicating

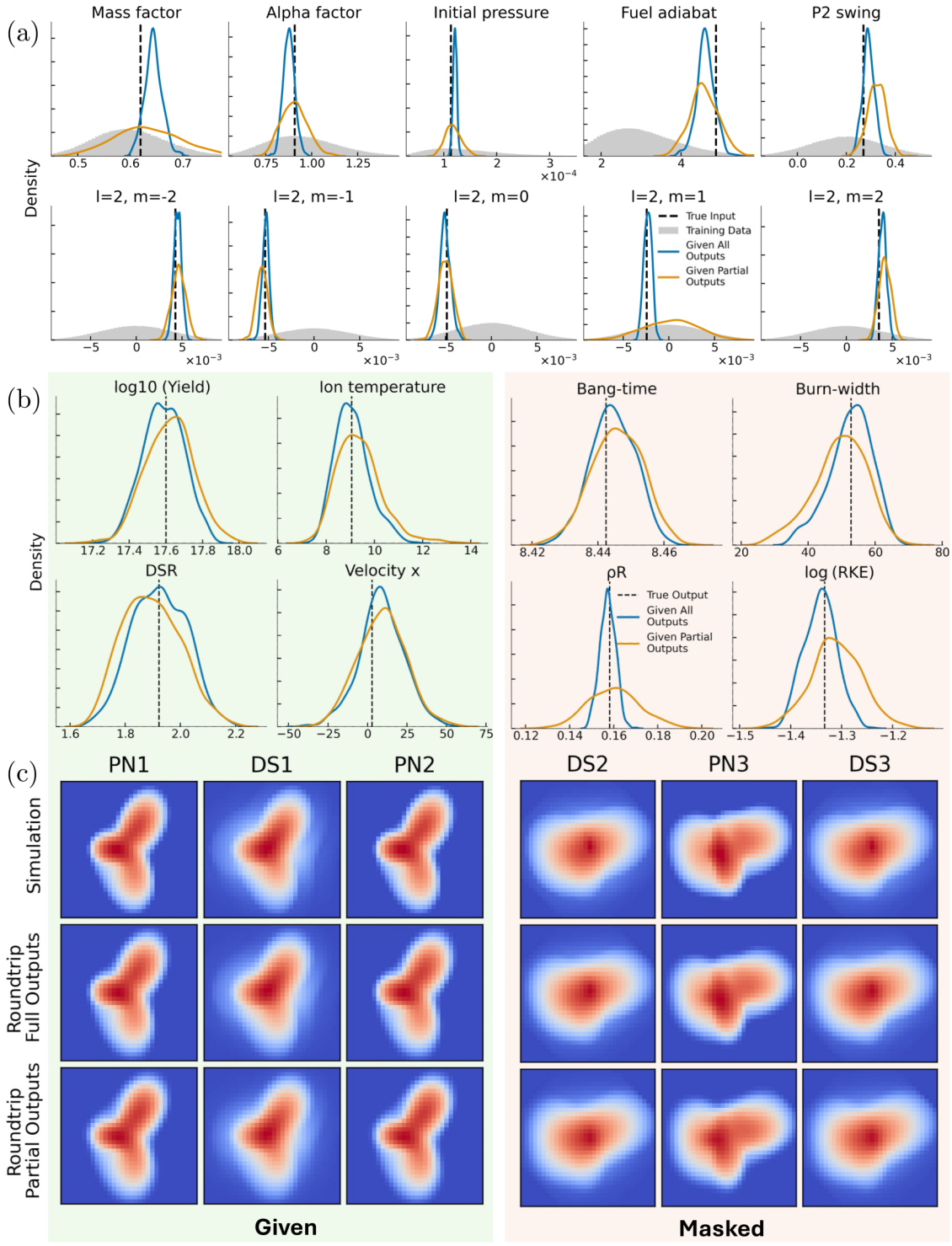

FIG. 5. Input and round-trip distributions for single test sample. (a) Distribution of inputs for 500 generations when providing model with all outputs (blue) or half of outputs (orange). Ground-truth inputs are shown by black dashed line. Distribution of all training data is shown in gray. (b) Round-trip output distributions produced by running inputs from panel (a) through the forward model. Both round-trip distributions of given outputs (green shading) are similar and centered on the original outputs, despite originating from different input distributions. Round-trip distributions for masked outputs (orange shading) are less consistent but retain significant overlap. (c) Round-trip images produced by running inputs from panel (a) through the forward model. Simulated images (first row) are compared to mean images of full outputs (second row) and partial outputs (third row).

that adding the missing information is helpful in constraining these distributions beyond what is commonly observed in experiments. Interestingly, we note that the $l = 2$, $m = 1$ mode collapses entirely onto the training data distribution when given partial outputs, whereas it is tightly constrained around the ground-truth input when given full outputs. This result makes sense in the context of Fig. 4, where we observe a much larger MAE for the $l = 2$, $m = 1$ mode when View 3 is removed. We show additional examples in Fig. S5 of the Supplemental Material [28], which show similar trends to Fig. 5 with some expected variation given differences in image profile and scalar magnitudes. This analysis provides guidance for both the relative utility of various ICF diagnostics and the confidence in input predictions when a given diagnostic fails.

Given the high dimensionality of the input space, it is challenging to verify that the predicted input distribution appropriately describes the outputs they are conditioned on. Indeed, we observe that the ground-truth inputs (vertical black lines) tend to fall within both the full and partial conditional distributions, but these represent only a single solution that might describe the observed output. In order to evaluate the complete distribution, we enlist the forward version of the model (which was shown to be a highly accurate surrogate in Fig. 3) to project samples back into output space. This "round-trip" analysis, visualized in Fig. 5(b) (additional examples shown in Fig. S6 of the Supplemental Material [28]), shows that both the partial outputs and full output distributions are self-consistent with the original conditioning values, as demonstrated by distributions centered on the ground-truth black dashed lines. For the scalar outputs that are provided in both cases (green shading), the round-trip distributions are very similar in both mean and variance. For outputs that were missing in the partial distribution (orange shading), distributions are similar for bang time and burn width, but the partial distributions are wider for ρR and RKE, indicating that the wider inferred distributions do not adequately constrain these values. Overall, strong round-trip consistency across variably constrained inputs indicates that the model can recover nonunique output-to-input mappings under partial observations.

We perform the same round-trip analysis on images. Image channels provided to both models are shown in the green box in Fig. 5(c), and masked images are shown in the red box. Again, we observe strong round-trip reconstructions when comparing the mean round-trip prediction to the original images, although there are some features slightly blurred in the partial round trips. Taken together, these results verify that the predicted input distributions correctly correspond to their conditioning values and that the forward and inverse modes are self-consistent, even when given partial information and without strict self-consistency enforced as in Ref. [3].

E. JointDiff is transferable to NIF experiments

After verifying that the model satisfies round-trip tests for partial simulations observables, we perform the same analysis on a set of NIF experiments. Each experimental shot contains data for the six scalars and three image channels described above, and can therefore be used to produce an analogous input and round-trip output distributions. We emphasize that for this test no experimental data were used to train or fine-tune the JointDiff model; therefore, we do not expect the model to reproduce various experimental features that are not captured by the relatively simple RP model. Indeed, if experimental data were significantly out-of-distribution compared to RP simulations and the JointDiff model were not robust, we would expect the round-trip distributions to be entirely uncorrelated with the observed experimental outputs they were conditioned on. However, in Fig. 6(a), we observe reasonably strong correlations for most scalars, indicating the model is transferable to experiment.

In Fig. 6(b), we examine the round-trip distribution of scalars and images predicted for the recent N230729 and N240907 shots. We observe that predicted output distributions are wider than simulations, indicating higher uncertainty, but that ground-truth outputs are captured within the distributions. Across all shots, DSR is the hardest scalar to match, potentially indicating inconsistency in the underlying RP model. In Fig. 6(c), we show round-trip predictions of View 1 primary neutron images for all experiments. Overall image shape is well captured in reconstructions; however, more subtle features are blurred or lost, especially in N221204 and N240210. These features may not be present in the RP training data, and fine-tuning JointDiff on experiments or higher fidelity simulations [4,30] would likely improve round-trip self-consistency.

This analysis provides a means of assessing both JointDiff robustness to out-of-distribution data as well as the physical similarity of RP training data to experiments. Importantly, we show that the model can produce an input distribution that approximately reproduces observed outputs. These analyses may guide future development of the RP model—for example, to align DSR predictions—and sampling strategies to approach experimental conditions near the ignition cliff.

F. Imputation is improved with experimental fine-tuning

Although we observe reasonable round-trip performance, the distribution shift between experiments and simulations introduces challenges for direct output imputation. To address this, we demonstrate how JointDiff can be fine-tuned on experimental data to improve performance on unseen shots. Fine-tuning under sparse and partially observed conditions is inherently difficult; therefore, we (1) augment experimental data using measured uncertainties, (2) impute missing values with the pretrained model to provide a prior, and (3) replay a small fraction of simulation data to preserve supervision over unobserved modalities (see Sec. IV for details). We randomly select two shots for testing and fine-tune on the remaining six experiments.

In Fig. 7(a), we impute yield (blue points) given all other available scalars and images, and in Fig. 7(b), we impute yield given experimental images only. In the first case, there is strong correlation but systematic underprediction compared to experiment, while in the second case experimental images are minimally predictive without scalar support. After fine-tuning (green points), predictions are significantly improved in both cases, as demonstrated by lower MAE for both the train data and test data (shown by green diamonds) and higher rate of ground-truth values within error bars. We note that there are substantial variations between experimental images, and we

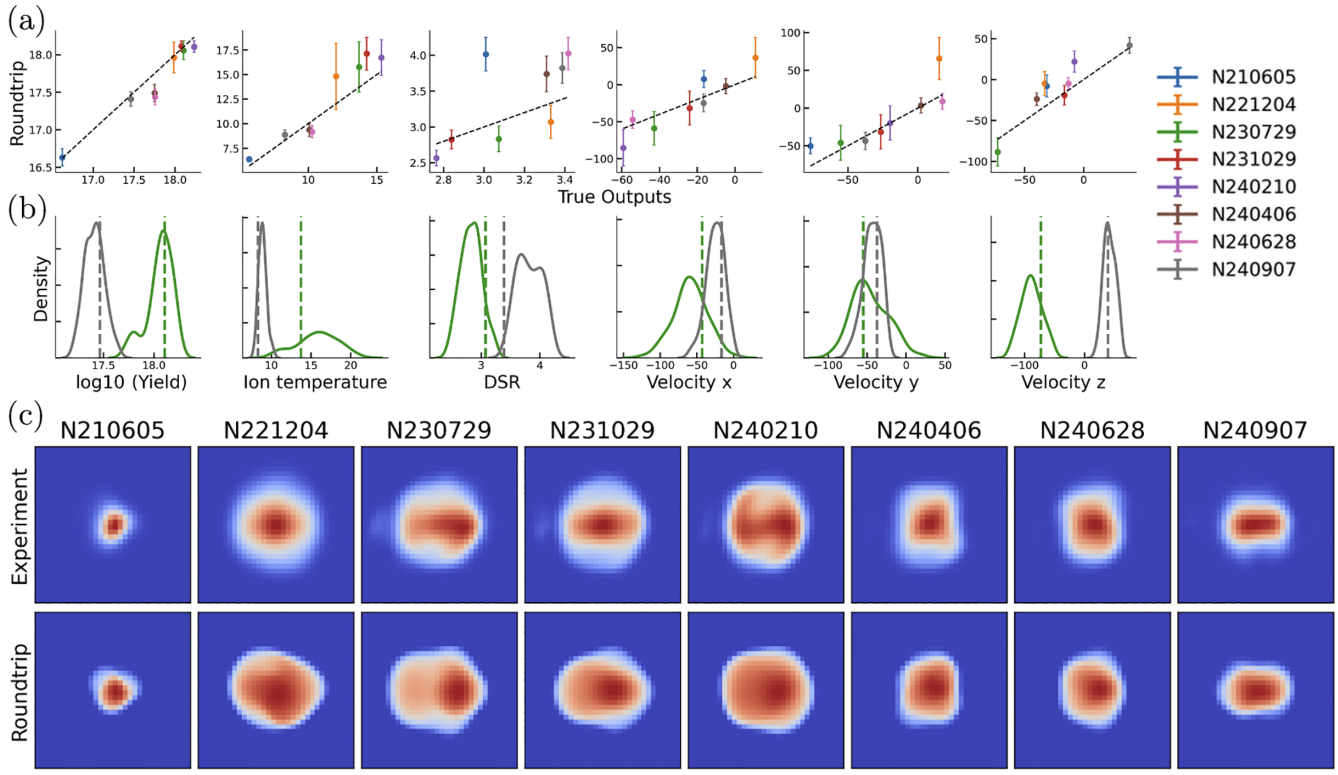

FIG. 6. Round-trip predictions given partial experimental outputs. (a) Round-trip mean and standard deviations for six output scalars measured in eight NIF experiments. (b) Round-trip output distributions over 100 generations for N230729 (green) and N240907 (gray) experiments. Experimental output values are shown as dashed lines in green and gray, respectively. (c) Experimental (top row) and round-trip generated images (bottom row) of View 1 primary neutrons for all eight shots.

are encouraged that the test predictions improve in Fig. 7(b) despite the model never seeing test images during fine-tuning. Interestingly, we find that the error after fine-tuning in scenario B is about equal to that of the pretrained model in scenario A, highlighting a trade-off between the addition of more conditional information and exposure to experiments during training.

As a second test, we compare the pretrained and fine-tuned models in their ability to reproduce a missing experimental image. In particular, we consider the View 1 down-scattered image, for which the pretrained model tends to miss symmetry features [Fig. 7(c), second row] when conditioned on other experimental scalars and images. After fine-tuning, we observe that generated training images [Fig. 7(c), first two columns] are nearly indistinguishable from the ground truth, with pixelwise MAE < 0.01. Improvements for the test images vary, but in both cases performance improves over the pretrained model. Notably, the fine-tuned model captures two lobes of higher intensity that were absent in the pretrained prediction, with MAE decreasing by an average of 44%. These predictions remain limited by the quantity and similarity of available experimental images, but are expected to improve as this analysis is scaled to larger NIF datasets.

III. CONCLUSION AND OUTLOOK

We presented JointDiff, a probabilistic framework for multimodal conditional generation that simultaneously performs forward prediction, inverse inference, and modality imputation using arbitrary subsets of scalar and image diagnostics. Applied to rocket-piston simulations, the model accurately reproduces distributions of observables, infers latent simulation inputs, and maintains self-consistency even when substantial fractions of scalar and image information are masked. Furthermore, the model and underlying RP training data support predictions on NIF experiments, and output imputation can be significantly improved through fine-tuning.

Beyond predictive performance, JointDiff provides a framework for assessing the diagnostic information content of experimental measurements. We showed that removing particular lines of sight increases uncertainty in inferred drive asymmetry modes, while down-scattered neutron measurements strongly constrain quantities such as adiabat and P2 swing. The systematic discrepancies observed for DSR predictions suggest limitations in the underlying RP model, motivating ongoing efforts to improve the treatment of neutron transport and shock timing. Looking forward, JointDiff may help inform optimal diagnostic configurations at the NIF and future inertial fusion facilities by identifying which measurements most strongly constrain particular physical quantities and most effectively reduce predictive uncertainty.

Several directions may further improve model performance and physical fidelity. Future work will incorporate higher-fidelity simulations such as HYDRA [25], larger experimental datasets, and additional diagnostic modalities. Methodologically, the joint diffusion objective could be

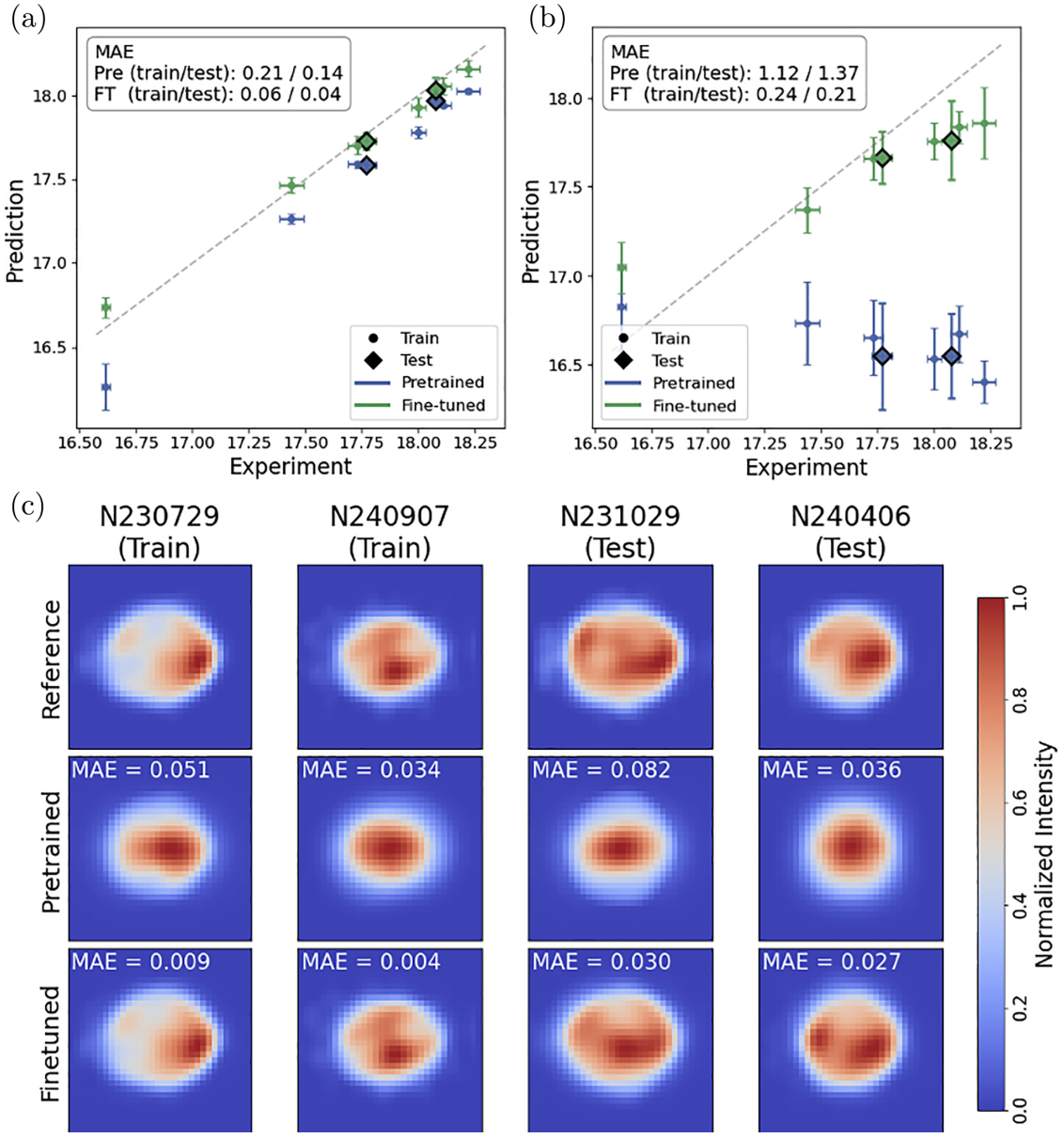

FIG. 7. Scalar and image predictions with pretrained and fine-tuned model. (a) and(b) Mean and standard deviation of yield predictions across eight NIF experiments, conditioned on (a) all available scalar and image data and (b) experimental images only. Pretrained predictions are shown in blue and fine-tuned predictions in green; diamond markers denote test shots. MAEs, computed over ten generations, are reported on each plot. (c) View 1 down-scattered neutron image predictions for four shots (two training and two test), imputed from all other available experimental data using the pretrained model (second row) and the fine-tuned model (third row). Pixelwise MAEs are computed over ten generations.

integrated with architectures such as diffusion transformers [21,31], curriculum learning strategies [32], or latent-space diffusion approaches [33]. Finally, incorporating experimental uncertainties during inference may further constrain predicted distributions and improve agreement with traditional Bayesian inference approaches.

Overall, these results demonstrate that JointDiff provides a flexible framework for probabilistic inference in ICF, while more broadly illustrating how multimodal generative models can perform forward prediction, inverse inference, and data imputation within partially observed physical systems.

IV. METHODS

A. Multirocket piston model

Simulated implosions are modeled using a radiation drive derived from the Callahan hohlraum model [34] and solve the rocket equations for 200 coupled rocket pistons that discretize the capsule shell in 3D. After the initial acceleration phase, the pistons are linked through hotspot pressure using power-balance equations, following the framework of Springer *et al.* [35] but with an explicit hohlraum model applied from the onset of implosion. Synthetic diagnostics at stagnation and peak burn are generated through postprocessing. Simulations

and diagnostic computation are performed in 15 s on a single CPU. Full methodological details are provided in Ref. [22].

B. Training data and preprocessing

An ensemble of 443 610 RP simulations was used with inputs sampled near the predicted experimental ignition cliff [22]. A random split of 986 samples (0.2%) was held out for testing, and 100 additional samples were used for validation during training. Images were normalized to [0, 1] by dividing by max pixel intensity for each view and channel. Scalar inputs and outputs were normalized to [0, 1] given mean and then passed through an inverse sigmoid $\hat{x} = \ln x/(1-x)$ to more closely approximate zero-centered Gaussians and to ensure predicted values cannot substantially exceed physical constraints of the simulation ensemble. All evaluation metrics reported in Tables I and II and Fig. 3 are performed after removing normalization.

C. ICF experiments

Deuterium-Tritium ICF experiments were performed at the NIF between June 2021 and September 2024. Additional information on specific shots can be found in Refs. [2,36–40].

D. Denoising diffusion probabilistic model

DDPMs generate high-resolution data distributions by first noising training data to gradually approach a Gaussian distribution, then iteratively denoising samples drawn from a Gaussian prior [11,12]. During training, samples X_t can be drawn at arbitrary diffusion times t according to $X_t = \sqrt{\bar{\alpha}_t}X_T + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $t \sim [0, 1]$, $\epsilon \sim \mathcal{N}(0, 1)$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_t$ corresponding to the noise schedule. We use a linear noise schedule that decreases α_t from 0.9999 to 0.98 over 1000 steps.

During inference, initial samples are drawn $X_0[j] \sim \mathcal{N}(0, 1)$ if $M[j] = 0$, where M corresponds to a binary conditioning mask (described below). All other $X_0[j]$ are set to their true (unmasked) values determined by simulation or experiment. A neural network model (described below) iteratively predicts how to remove noise from the sample, given the mask and diffusion time $\hat{\epsilon} = f_\theta(X_t, M, t)$. The sample is updated according to $X_{t+1} = \frac{1}{\sqrt{\alpha}}(X_t - \frac{1-\alpha}{\sqrt{1-\bar{\alpha}}}\hat{\epsilon}) + \sqrt{(1-\alpha)}\epsilon$ and noise is continuously removed over 1000 steps. Both image and scalar quantities are treated with identical (de)noising procedures, the only difference being the dimensions of data modalities.

E. Neural network architecture

A multimodal neural network architecture is used to encode and predict noise in both scalar and image pixel space. We adapt a U-Net architecture [41] developed by Ref. [42] to ingest and predict multimodal data. Each image view is embedded by independent convolution networks containing 48 channels and 3 up-/down-sampling layers. Both input scalars, output scalars, and mask tokens are embedded by fully connected neural networks containing one 64-dimensional hidden layer. Diffusion time is embedded by a sin/cos positional encoding. Scalar, mask, and time embeddings are concatenated and added to U-Net embedding at each up- and down-sampling step. Image embedding for each view receive the same time, scalar, and mask conditioning but are independent from each other until the bottleneck layer (lowest dimension of the U-Net). At the bottleneck layer, outputs from each image view are pooled along the channel dimension, flattened, and concatenated along with scalar and mask embedding. Output and input scalar decoders take the flattened bottleneck layer as input and predict noise for each scalar. Image decoders pool information from the bottleneck layer and upsample three times to recover the original pixel and channel dimensions.

Models are constructed in PyTorch [43], and trained using the Adam optimizer [44]. We use a batch size of 256 and learning rate of 0.0003. An exponential moving average with constant 0.995 is applied to weights. Training was performed on a single NVIDIA H100 GPU and converged in approximately 8 h. Inference is carried out with batch sizes of up to 8000, with an average latency of about 0.2 s per sample. Additional architecture details and hyperparameters are included in Tables S1 and S2 of the Supplemental Material [28].

F. Masking

Binary masks inform the model if a conditioning variable (scalar or image) is provided for conditioning, or if it should be predicted during the denoising (inference) process. Certain masks are linked to others if those quantities can never be known without the other. Specifically, all input masks are linked since they must all be specified to perform a simulation and cannot be partially observed in experiment. The four inferred output scalars (ρR weighted harmonic mean, ρR mean, RKE at bang time, and RKE at minimum volume) are linked for the same reason [27]. The three velocity scalars are linked because they are x, y, and z components of the same vector. All other output scalars have individual masks along with each image view and channel. During training, all independent masks are sampled according to a Bernoulli distribution with constant 0.5. Ablations were performed with varied constant values and with curriculum learning strategies to gradually lower the constant as function of training time, but a fixed constant of 0.5 provided optimal performance across forward, inverse, and imputation tasks.

G. Fine-tuning

Experimental data were randomly split into six training shots and two test shots. For each training shot, noise was sampled from experimentally measured scalar uncertainties, $\tilde{s}e \leftarrow s + \epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0, \sigma s^2)$, generating 1000 augmented samples per shot. Images do not have corresponding experimental uncertainties and were not modified. All missing scalar and image quantities were then imputed using the pretrained model. These imputed values are not regressed upon during training, but instead serve as priors to ensure the model observes consistent, partially noised samples. During fine-tuning, simulation data are sampled 10% of the time using the same batch size and learning rate as experimental data, providing supervision on missing modalities. All base hyperparameters are consistent with pretraining, except for the learning rate, which is reduced from 3×10^{-4} to 1×10^{-4},

and the number of epochs, which is reduced from 300 pretraining epochs to 100 fine-tuning epochs. Fine-tuning is completed in approximately 20 min on a single NVIDIA H100 GPU.

ACKNOWLEDGMENTS

The authors thank the international Inertial Confinement Fusion collaboration and the LLNL ICF program. This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344.

K.H., E.K., B.K., and M.J. designed the research. M.J performed research. D.C. and J.K. generated simulation training data. M.J. wrote the paper. K.H., D.C, B.K., and J.K. advised the paper.

There are no competing interests to declare.

DATA AVAILABILITY

The data that support the findings of this article are pending institutional review and not publicly available at the time of acceptance. The data are available from the authors upon reasonable request.

[1] A. B. Zylstra, O. A. Hurricane, D. A. Callahan, A. L. Kritcher, J. E. Ralph, H. F. Robey, J. S. Ross, C. V. Young, K. L. Baker, D. T. Casey, *et al.*, Burning plasma achieved in inertial fusion, Nature (London) **601**, 542 (2022).

[2] H. Abu-Shawareb, R. Acree, P. Adams, J. Adams, B. Addis, R. Aden, P. Adrian, B. Afeyan, M. Aggleton, L. Aghaian, *et al.*, Achievement of target gain larger than unity in an inertial fusion experiment, Phys. Rev. Lett. **132**, 065102 (2024).

[3] R. Anirudh, J. J. Thiagarajan, P.-T. Bremer, and B. K. Spears, Improved surrogates in inertial confinement fusion with manifold and cycle consistencies, Proc. Natl. Acad. Sci. USA **117**, 9741 (2020).

[4] B. Kustowski, J. A. Gaffney, B. K. Spears, G. J. Anderson, R. Anirudh, P.-T. Bremer, J. J. Thiagarajan, M. K. Kruse, and R. C. Nora, Suppressing simulation bias in multi-modal data using transfer learning, Mach. Learn.: Sci. Technol. **3**, 015035 (2022).

[5] J. A. Gaffney, K. Humbird, A. Kritcher, M. Kruse, E. Kur, B. Kustowski, R. Nora, and B. Spears, Data-driven prediction of scaling and ignition of inertial confinement fusion experiments, Phys. Plasmas **31**, 092702 (2024).

[6] B. K. Spears, S. Brandon, D. T. Casey, J. E. Field, J. A. Gaffney, K. D. Humbird, A. L. Kritcher, M. K. Kruse, E. Kur, B. Kustowski, *et al.*, Predicting fusion ignition at the National Ignition Facility with physics-informed deep learning, Science **389**, 727 (2025).

[7] J. H. Kunimune, D. T. Casey, B. Kustowski, V. Geppert-Kleinrath, L. Divol, D. N. Fittinghoff, P. L. Volegov, M. K. Kruse, J. A. Gaffney, R. C. Nora, *et al.*, 3D reconstruction of an inertial-confinement fusion implosion with neural networks using multiple heterogeneous data sources, Rev. Sci. Instrum. **95**, 073506 (2024).

[8] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3, Nature (London) **630**, 493 (2024).

[9] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, *et al.*, BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, arXiv:2303.00915.

[10] Y. Wu, M. Ding, H. He, Q. Wu, S. Jiang, P. Zhang, and J. Ji, A versatile multimodal learning framework bridging multiscale knowledge for material design, npj Comput. Mater. **11**, 276 (2025).

[11] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. **33**, 6840 (2020).

[12] J. Song, C. Meng, and S. Ermon, Denoising diffusion implicit models, in *International Conference on Learning Representations* (ICLR, 2021).

[13] F. Bao, S. Nie, K. Xue, C. Li, S. Pu, Y. Wang, G. Yue, Y. Cao, H. Su, and J. Zhu, One transformer fits all distributions in multi-modal diffusion at scale, in *International Conference on Machine Learning* (PMLR, 2023), pp. 1692–1717.

[14] S. Li, K. Kallidromitis, A. Gokul, Z. Liao, Y. Kato, K. Kozuka, and A. Grover, Any-to-any generation with multi-modal rectified flows, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2025), pp. 13178–13188.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, Learning transferable visual models from natural language supervision, in *International Conference on Machine Learning* (PMLR, 2021), pp. 8748–8763.

[16] H. Chen, X. Wang, Y. Zhou, B. Huang, Y. Zhang, W. Feng, H. Chen, Z. Zhang, S. Tang, and W. Zhu, Multi-modal generative AI: Multi-modal LLM, diffusion and beyond, IEEE Trans. Circuits Syst. Video Technol. **36**, 5621 (2024).

[17] A. E. Chu, J. Kim, L. Cheng, G. El Nesr, M. Xu, R. W. Shuai, and P.-S. Huang, An all-atom protein generative model, Proc. Natl. Acad. Sci. USA **121**, e2311500121 (2024).

[18] W. Qu, J. Guan, R. Ma, K. Zhai, W. Wu, and H. Wang, *P*(all-atom) is unlocking new path for protein design, in *International Conference on Machine Learning* (PMLR, 2025).

[19] A. Campbell, J. Yim, R. Barzilay, T. Rainforth, and T. Jaakkola, Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design, in *International Conference on Machine Learning* (PMLR, 2024), pp. 7240–7260.

[20] P. Holderrieth, M. Havasi, J. Yim, N. Shaul, I. Gat, T. Jaakkola, B. Karrer, R. T. Chen, and Y. Lipman, Generator matching: Generative modeling with arbitrary Markov processes, arXiv:2410.20587.

[21] K. Rojas, Y. Zhu, S. Zhu, F. X.-F. Ye, and M. Tao, Diffuse everything: Multimodal diffusion models on arbitrary state spaces, in *International Conference on Machine Learning* (PMLR, 2025).

[22] D. T. Casey, J. Kunimune, O. A. Hurricane, O. L. Landen, P. Springer, R. M. Bionta, C. V. Young, R. C. Nora, B. J. Macgowan, J. A. Gaffney, *et al.*, A multi-rocket piston model to

study three-dimensional asymmetries in implosions at the National Ignition Facility, High Energy Density Phys. **54**, 101172 (2025).

[23] C. Liu, C. Wu, S. Cao, M. Chen, J. C. Liang, A. Li, M. Huang, C. Ren, D. Liu, Y. N. Wu, *et al.*, Diff-PIC: Revolutionizing particle-in-cell nuclear fusion simulation with diffusion models, in *International Conference on Learning Representations* (ICLR, 2025).

[24] M. Jones and B. Kustowski, Towards a multi-modal foundation model for inertial confinement fusion: Combining structured data and diagnostic images, *presented at the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, BC, Canada* (2025).

[25] M. M. Marinak, G. D. Kerbel, N. A. Gentile, O. S. Jones, D. H. Munro, T. R. Dittrich, and S. W. Haan, Three-dimensional Hydra simulations of National Ignition Facility targets, Phys. Plasmas **8**, 2275 (2001).

[26] O. A. Hurricane, D. T. Casey, O. Landen, A. L. Kritcher, R. Nora, P. K. Patel, J. A. Gaffney, K. D. Humbird, J. E. Field, M. K. G. Kruse, *et al.*, An analytic asymmetric-piston model for the impact of mode-1 shell asymmetry on ICF implosions, Phys. Plasmas **27**, 062704 (2020).

[27] O. A. Hurricane, D. T. Casey, O. Landen, D. A. Callahan, R. Bionta, S. Haan, A. L. Kritcher, R. Nora, P. K. Patel, P. T. Springer, *et al.*, Extensions of a classical mechanics "piston-model" for understanding the impact of asymmetry on ICF implosions: The cases of mode 2, mode 2/1 coupling, time-dependent asymmetry, and the relationship to coast-time, Phys. Plasmas **29**, 012703 (2022).

[28] See Supplemental Material at http://link.aps.org/supplemental/10.1103/dldf-7tfm for additional figures, extended prediction results, and supplementary analyses.

[29] V. E. Fatherley, S. H. Batha, C. R. Danly, L. A. Goodwin, H. W. Herrmann, H. J. Jorgenson, J. I. Martinez, F. E. Merrill, J. A. Oertel, D. W. Schmidt, *et al.*, System design of the NIF neutron imaging system north pole, in *Target Diagnostics Physics and Engineering for Inertial Confinement Fusion VI* (SPIE, Bellingham, WA, 2017), Vol. 10390, pp. 46–57.

[30] K. D. Humbird, J. L. Peterson, J. Salmonson, and B. K. Spears, Cognitive simulation models for inertial confinement fusion: Combining simulation and experimental data, Phys. Plasmas **28**, 042709 (2021).

[31] W. Peebles and S. Xie, Scalable diffusion models with transformers, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, Piscataway, NJ, 2023), pp. 4195–4205.

[32] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, Curriculum learning, in *Proceedings of the 26th Annual International Conference on Machine Learning* (ACM, New York, NY, 2009), pp. 41–48.

[33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2022), pp. 10684–10695.

[34] D. A. Callahan, O. A. Hurricane, A. L. Kritcher, D. T. Casey, D. E. Hinkel, Y. P. Opachich, H. F. Robey, M. D. Rosen, J. S. Ross, M. S. Rubery, *et al.*, A simple model to scope out parameter space for indirect drive designs on NIF, Phys. Plasmas **27**, 072704 (2020).

[35] P. T. Springer, O. A. Hurricane, J. H. Hammer, R. Betti, D. A. Callahan, E. M. Campbell, D. T. Casey, C. J. Cerjan, D. Cao, E. Dewald, *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, Nucl. Fusion **59**, 032009 (2019).

[36] H. Abu-Shawareb, R. Acree, P. Adams, J. Adams, B. Addis, R. Aden, P. Adrian, B. B. Afeyan, M. Aggleton, L. Aghaian, *et al.*, Lawson criterion for ignition exceeded in an inertial fusion experiment, Phys. Rev. Lett. **129**, 075001 (2022).

[37] O. A. Hurricane, Inertial confinement fusion: Status and challenges, Annu. Rev. Nucl. Part. Sci. **75**, 153 (2025).

[38] A. B. Zylstra, A. L. Kritcher, O. A. Hurricane, D. A. Callahan, J. E. Ralph, D. T. Casey, A. Pak, O. L. Landen, B. Bachmann, K. L. Baker, *et al.*, Experimental achievement and signatures of ignition at the National Ignition Facility, Phys. Rev. E **106**, 025202 (2022).

[39] A. L. Kritcher, A. B. Zylstra, C. R. Weber, O. A. Hurricane, D. A. Callahan, D. S. Clark, L. Divol, D. E. Hinkel, K. Humbird, O. Jones, *et al.*, Design of the first fusion experiment to achieve target energy gain $G > 1$, Phys. Rev. E **109**, 025204 (2024).

[40] O. A. Hurricane, A. Allen, B. L. Bachmann, K. L. Baker, S. Baxamusa, S. D. Bhandarkar, J. Biener, S. R. M. Bionta, T. Braun, T. Briggs, *et al.*, Present understanding of ignition and gain using indirect-drive inertial confinement fusion target designs on the US National Ignition Facility, Plasma Phys. Controlled Fusion **67**, 015019 (2025).

[41] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, Cham, Switzerland, 2015), pp. 234–241.

[42] Dome272, Diffusion-models-PyTorch, https://github.com/dome272/Diffusion-Models-pytorch, accessed 15 May 2025.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, PyTorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. **32**, 8024 (2019).

[44] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *International Conference on Learning Representations* (ICLR, 2015).