

Decomposition-Guided Diffusion Language Models for Inertial Confinement Fusion Prediction

Xiang Zhang¹, Varchas Gopalaswamy², Rahman Ejaz², Riccardo Betti², Dongfang Liu^{1†}

¹Purdue University, ²University of Rochester

[†]Corresponding author

Abstract

Inertial confinement fusion (ICF) is a leading pathway toward clean energy, but each shot at the National Ignition Facility costs on the order of \$1M, making accurate AI surrogates a high-value target. We study *exogenous-driven* ICF waveform prediction, where a 512-step neutron-rate diagnostic must be inferred directly from a laser pulse and target design parameters, with no historical response observed. The regime stresses standard time-series predictors with temporal sparsity (picosecond peak in a nanosecond window), input-output scale mismatch (<300 real shots), and peak sensitivity (picosecond timing). We propose **ICF-DLM**, to our knowledge the first LM-based ICF predictor, combining (i) a physics-typed decomposition into yield Y_{DT} , peak timing t_{peak} , and local waveform $\mathbf{w}_{\text{local}}$; (ii) bidirectional denoising that defers commitment to peak location; and (iii) a physics-driven PPO reward re-injecting metric structure across numeric tokens. On **ICFBench** (50K simulations + 232 experimental shots), ICF-DLM cuts peak-timing error from 11.6 to 9.2 steps over a matched autoregressive LLaMA-3-8B and outperforms classical sequence models and LLM-based time-series predictors. Beyond ICF, the recipe shows potential to address science domains with low data and sparse events.

1 Introduction

Inertial confinement fusion (ICF) is a leading pathway toward clean energy (Nuckolls et al., 1972; Lindl, 1995; Betti and Hurricane, 2016; Craxton et al., 2015; Abu-Shawareb et al., 2024), but physical experiments are extraordinarily expensive. A single shot at the National Ignition Facility (Radha et al., 2016; Humbird et al., 2021) costs on the order of \$1 million, with only a few feasible per year. High-fidelity radiation-hydrodynamics simulations are cheaper but remain computationally costly and suffer from sim-to-real domain shift

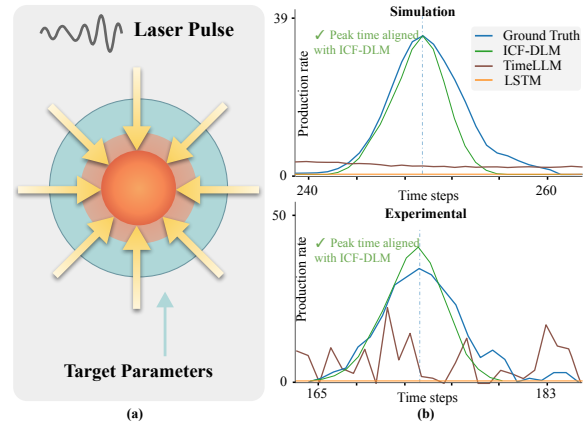


Figure 1: (a) Exogenous-driven ICF waveform prediction maps a laser pulse and target parameters to a 512-step neutron-rate diagnostic. (b) ICF-DLM outputs on simulation (top) and experimental (bottom) data.

(Humbird et al., 2019). This motivates a question for AI for science (Carleo et al., 2019; Song et al., 2025). Can pretrained language models serve as fast and accurate ICF surrogates?

We frame ICF prediction as exogenous-driven waveform generation. Given a time-resolved laser pulse and a few target design parameters, we predict the 512-step neutron-rate diagnostic with no historical observation of the response (Fig. 1(a)). The regime exposes three properties that break standard predictors. **(P1) Temporal sparsity**: neutron emission occupies a picosecond window inside a nanosecond sequence, so pointwise objectives reward modeling the baseline rather than the event. **(P2) Input-output scale mismatch**: a few hundred inputs determine the 512-step output, with only ~ 300 real shots to anchor the mapping. **(P3) Peak sensitivity**: small timing shifts cascade into large L_2 errors, so committing early to the wrong peak dominates the loss.

Naively casting this task as next-token prediction with an autoregressive LM or as continuous regression with a Transformer leaves the structure of (P1)–(P3) unexploited. Autoregressive decoding

commits to peak placement before the surrounding waveform is generated; deterministic regressors collapse to conditional means and obscure the heavy-tailed yield distribution (Salinas et al., 2020); classical sequence models trained on $\mathcal{O}(10^2)$ shots overfit and fail to transfer from simulation (Craxton et al., 2015). Diffusion language models (Ho et al., 2020; Austin et al., 2021; Li et al., 2022; Nie et al., 2026) instead offer bidirectional refinement and a generative formulation amenable to reinforcement learning over numeric tokens, two ingredients aligned with (P1)–(P3).

We present ICF-DLM, to our knowledge the first LM-based predictor for ICF. It addresses the above challenges through coordinated design choices. **(i) Physics-typed decomposition:** Following how physicists read radiation-hydrodynamics outputs, we factorize the waveform into total yield Y_{DT} , peak timing t_{peak} , and a short local waveform $\mathbf{w}_{\text{local}}$ around the peak. Predicting these descriptors jointly, rather than emitting 512 unstructured steps, concentrates supervision on the picosecond emission window and shrinks the effective output dimensionality to a few dozen tokens, both important under sparse events and a small experimental set. **(ii) Bidirectional denoising:** An LLaDA-8B (Nie et al., 2026) backbone refines all three descriptors in parallel under global bidirectional attention. The peak location is never committed before the surrounding waveform is generated, so small early-step mistakes can be corrected later, which is critical when timing shifts dominate the loss. **(iii) Physics-driven reward:** A composite PPO (Schulman et al., 2017) reward scores magnitude accuracy, peak alignment, and waveform fidelity at the descriptor level. This re-injects the metric structure discarded by token-level cross-entropy, which treats adjacent numeric tokens as fully distinct classes, and supplies a continuous signal at the peak where small errors propagate to large L_2 deviations.

We further curate **ICFBench**, pairing 50,000 simulation samples with 232 quality-controlled experimental shots, the largest ICF prediction benchmark to date. ICF-DLM outperforms classical sequence models, LLM-based time-series predictors, and a matched autoregressive LLaMA-3-8B trained with the same SFT+PPO recipe (Fig. 1(b); §4.2), with ablations isolating each design choice (§4.3). Beyond ICF, the recipe (bidirectional generation, physics-typed factorization, metric-aware reward) may transfer to LM-based scientific prediction in

low-data sparse-event regimes.

2 Related Work

2.1 AI for Inertial Confinement Fusion

Data-driven surrogates have been widely pursued where experiments and simulations are costly (Reichstein et al., 2019; Carleo et al., 2019; LeCun et al., 2015), including in ICF, where prior work maps laser pulse and target parameters to yield, hot-spot quantities, and implosion dynamics (Gopalaswamy et al., 2019; Ejaz et al., 2024b,a; Gopalaswamy et al., 2025; Lees et al., 2021; Cao et al., 2022; Liu et al., 2025), mostly through deterministic point-estimate models (Safavian and Landgrebe, 1991; Breiman, 2001; Hearst et al., 1998; Ahmed et al., 2010; Hatfield et al., 2019). Stronger inductive biases come from physics-informed and multi-fidelity learning (Raissi et al., 2019; Karniadakis et al., 2021; Li et al., 2020; Sanchez-Gonzalez et al., 2020; Han et al., 2018; Meng and Karniadakis, 2020), yet point estimates underrepresent the intrinsic variability of ICF, where small perturbations and diagnostic noise produce large yield deviations (Goncharov et al., 2008; Radha et al., 2016; Gopalaswamy et al., 2019; Gaffney et al., 2024). Transfer learning narrows the sim-to-real gap (Pan and Yang, 2009; Humbird et al., 2019; Humbird and Peterson, 2022), yet to our knowledge no prior work applies *pretrained language models* to ICF. ICF-DLM fills this gap, framing exogenous-driven waveform prediction as conditional sequence generation under (P1)–(P3) and releasing a benchmark pairing 50K simulations with 232 experimental shots.

2.2 Language Models for Time-Series and Numeric Prediction

Pretrained language models adapt to numeric time-series in two lines. *Reprogramming-based* approaches (TimeLLM (Jin et al., 2024), LPI-LLM (Chen et al., 2024)) project continuous patches into the LM embedding space (Zhou et al., 2023). *Tokenization-based* approaches discretize values into the LM vocabulary so next-token prediction performs regression (Wang et al., 2025; Golkar et al., 2023). Both rely on autoregressive forecasting with modality-matched inputs and outputs, an assumption ICF violates under (P2). We therefore replace AR decoding with bidirectional diffusion (P3) and pointwise cross-entropy with a physics-driven reward; the matched-AR ablation (§4.3) iso-

lates the gain.

Beyond LM-based predictors, dedicated architectures formulate physical evolution as recurrent (Hochreiter and Schmidhuber, 1997; Cho et al., 2014; Ranzato et al., 2015), Neural-ODE (Chen et al., 2018; Ruthotto and Haber, 2020; Greydanus et al., 2019; Cranmer et al., 2020), Transformer-variant (Vaswani et al., 2017; Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Nie et al., 2022; Liu et al., 2024; Wang et al., 2024), or state-space (Gu et al., 2021; Gu and Dao, 2023) prediction; their pointwise objectives (Lim et al., 2021; Sapankevych and Sankar, 2009; Le Guen and Thome, 2019) and strict autoregressive ordering (Ranzato et al., 2015; Laban et al., 2026) compound with (P3) to produce poor peak placement (Craxton et al., 2015). Diffusion-based predictors (CSDI (Tashiro et al., 2021), TimeDiff (Shen and Kwok, 2023), Diffusion-TS (Yuan and Qiao, 2024)) sidestep the AR-ordering issue through iterative denoising, but use regression backbones on modality-matched horizons. ICF-DLM differs from these diffusion predictors in four respects, namely an exogenous-driven setting in which input and output share no modality, a pretrained diffusion LM with discrete numeric tokens rather than a regression backbone, a physically factorized output rather than a flat waveform, and PPO with a composite physics reward rather than denoising-only losses.

2.3 Diffusion Language Models

Discrete diffusion language models (Austin et al., 2021; Li et al., 2022; Savinov et al., 2021; Lou et al., 2023; Sahoo et al., 2024; Nie et al., 2026; Jing et al., 2026) iteratively denoise masked tokens with bidirectional context, in contrast to the left-to-right factorization of autoregressive LMs (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023; Grattafiori et al., 2024; Chowdhery et al., 2023). LLaDA (Nie et al., 2026) matches strong autoregressive baselines on language tasks while offering parallel generation, global conditioning, and an iterative refinement schedule (Ho et al., 2020) that can revise earlier commitments as later tokens crystallize. We exploit this revision capability against (P3), committing to peak location only after the surrounding waveform has been refined. We are the first, to our knowledge, to combine a pretrained diffusion LM (Nie et al., 2026), numeric tokenization (Wang et al., 2025; Golkar et al., 2023), and physics-driven RL reward shaping for scientific time-series prediction.

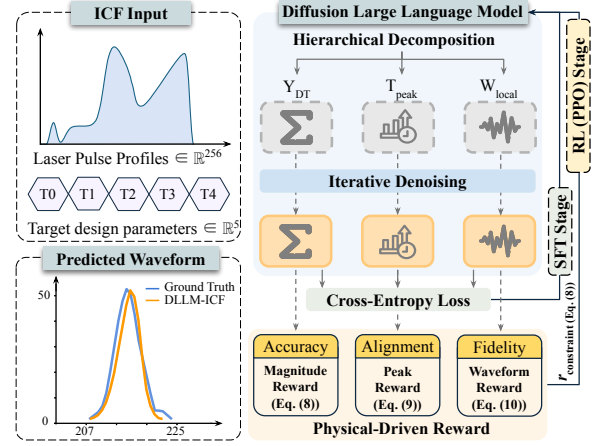


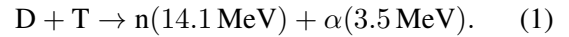
Figure 2: ICF-DLM pipeline. The three physical descriptors (Y_{DT} , t_{peak} , w_{local}) are denoised jointly under physics-driven PPO rewards for magnitude, peak timing, and waveform fidelity.

3 Methods

ICF-DLM (Fig. 2) maps a laser pulse and target parameters to a 512-step neutron-rate waveform through three stages, namely input encoding, joint denoising of the physical descriptors (Y_{DT} , t_{peak} , w_{local}) under physics-driven PPO rewards, and waveform reconstruction.

3.1 Problem Setup

Physical context. In ICF, high-power lasers compress a spherical deuterium-tritium (DT) target. As the fuel reaches extreme temperatures and densities, it initiates the fusion reaction:



The experimental conditions are encoded by a composite input $\mathbf{x} \in \mathbb{R}^{261}$ formed by five **target parameters** $\vec{T} \in \mathbb{R}^5$ (e.g., outer radius, shell/ice thickness) and a **pulse shape** $\vec{P} \in \mathbb{R}^{256}$, the time-resolved laser power that drives the implosion.

Output waveform. The primary implosion diagnostic is the neutron rate $\dot{n}(t)$. We define the output $\mathbf{y} = [\dot{n}(t_1), \dots, \dot{n}(t_T)] \in \mathbb{R}^T$ with $T=512$. The waveform has a sharp emission peak at the *bang time* t_{peak} and is negligible elsewhere, reflecting (P1). The total yield is the integral:

$$Y_{DT} = \int_0^T \dot{n}(t) dt. \quad (2)$$

Problem formulation. We learn $f: \mathbf{x} \mapsto \mathbf{y}$ predicting the full neutron-rate waveform via a hierarchical physics-typed factorization that decomposes $\mathbf{y}$ into the yield (Y_{DT}), peak timing (t_{peak}), and local shape (w_{local}).

3.2 ICF-DLM

3.2.1 Diffusion Language Modeling

The descriptors $(Y_{DT}, t_{\text{peak}}, \mathbf{w}_{\text{local}})$ are emitted as discrete tokens. Continuous values in $[-0.5, 0.5]$ map to a vocabulary $\mathcal{V}_{\text{num}}$ of $V=10,000$ uniformly-spaced bins, and the structured prompt $\mathbf{x}=\text{tok}(\vec{P}, \vec{T})$ and the masked output tokens form the model input.

LLaDA (Nie et al., 2026) couples a forward masking process with a reverse unmasking process. Given the clean descriptor tokens $\mathbf{z}_0$ (the tokenized $(Y_{DT}, t_{\text{peak}}, \mathbf{w}_{\text{local}})$) and a mask ratio $\beta \in [0, 1]$, the forward process replaces a fraction β of positions with a $[M]$ token. The reverse process parameterizes the masked positions through a bidirectional Transformer:

$$p_{\theta}(\mathbf{z}_0 \mid \mathbf{z}_{\beta}, \mathbf{x}) = \prod_{i \in \mathcal{M}} p_{\theta}(z_0^{(i)} \mid \mathbf{z}_{\beta}, \mathbf{x}), \quad (3)$$

trained by the masked cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{\beta \sim U(0,1)} \sum_{i \in \mathcal{M}} \log p_{\theta}(z_0^{(i)} \mid \mathbf{z}_{\beta}, \mathbf{x}). \quad (4)$$

At inference we denoise over $S=128$ steps with confidence-based remasking and block size 128 (ablated in §4.3).

3.2.2 Physics-Typed Hierarchical Decomposition

Targeting (P1) and (P2), we decompose the prediction into three physical descriptors inspired by radiation-hydrodynamics (rad-hydro) workflows (Delettrez et al., 1987; Radha et al., 2005; Igumenchev et al., 2016). Physicists routinely characterize ICF responses using scalar energy descriptors (e.g., yield Y_{DT}), temporal landmarks (e.g., bang time t_{peak}), and local waveform structure ($\mathbf{w}_{\text{local}}$); see Appendix S7. Rather than predicting $\mathbf{y} \in \mathbb{R}^T$ directly, we predict:

$$(Y_{DT}, t_{\text{peak}}, \mathbf{w}_{\text{local}}) = f(\mathbf{x}; \theta), \quad (5)$$

where Y_{DT} is the integrated total yield, t_{peak} is the peak timing, and $\mathbf{w}_{\text{local}} \in \mathbb{R}^{T_{\text{local}}}$ is the local waveform around the peak. The DLM denoises all three jointly at each step, with each descriptor receiving an independent reward (§3.2.3). The full waveform is reconstructed at inference by embedding $\hat{\mathbf{w}}_{\text{local}}$ at $\hat{t}_{\text{peak}}$, with the remaining steps filled by a near-zero baseline (Appendix S2.1). The factorization concentrates supervision on the peak

window and reduces effective output dimensionality from $T=512$ to ~ 20 tokens, jointly addressing the sparsity and low-data conditions of (P1)–(P2). We impose the physical identity $\hat{Y}_{DT} = \sum_t \hat{y}_t$ at reconstruction. Specifically, the predicted local waveform is rescaled by $\hat{Y}_{DT} / \sum_t \hat{w}_{\text{local},t}$ before being embedded at $\hat{t}_{\text{peak}}$. In this manner, the yield descriptor and the waveform descriptor remain consistent by construction.

3.2.3 Physics-Driven Reward Modeling

Optimization is coarse-to-fine (Ouyang et al., 2022). We first apply supervised fine-tuning with $\mathcal{L}_{\text{SFT}}$ (§3.2.1) (Chung et al., 2024; Wei et al., 2021) to initialize the hierarchical output format $(Y_{DT} \rightarrow t_{\text{peak}} \rightarrow \mathbf{w}_{\text{local}})$, and then continue with PPO (Schulman et al., 2017), which optimizes a clipped surrogate of the reward gain:

$$\mathcal{L}_{\text{PPO}}(\theta) = -\mathbb{E}[\min(\rho_{\theta} \hat{A}, \text{clip}_{\epsilon}(\rho_{\theta}) \hat{A})] + \beta_{\text{KL}} \text{KL}(\pi_{\theta} \parallel \pi_{\text{ref}}), \quad (6)$$

where ρ_{θ} is the per-token importance ratio, $\hat{A}$ the advantage from the composite reward r , π_{ref} the SFT-initialized policy, and $\text{clip}_{\epsilon}(\cdot) = \text{clip}(\cdot, 1-\epsilon, 1+\epsilon)$. Token-level cross-entropy is insufficient. The three descriptors carry very different numerical scales and error sensitivities, and cross-entropy treats adjacent numeric tokens as fully distinct classes, inducing a metric mismatch under discrete tokenization. We therefore define the composite reward:

$$r = r_{\text{mag}} + r_{\text{peak}} + r_{\text{wave}} + r_{\text{constraint}}, \quad (7)$$

which re-injects metric structure across numeric tokens, targeting (P1) and (P3).

Magnitude accuracy reward. Yield Y_{DT} is a scalar spanning several orders of magnitude across shots. To produce a bounded, scale-invariant signal, we use a reciprocal of the relative yield error.

$$r_{\text{mag}}(\hat{Y}_{DT}, Y_{DT}^*) = \frac{1}{1 + \left| \frac{\hat{Y}_{DT} - Y_{DT}^*}{Y_{DT}^*} \right|}. \quad (8)$$

A perfect prediction maps to 1.0 and large errors approach 0 asymptotically.

Peak alignment reward. Peak timing exhibits sharp local sensitivity, since small temporal offsets cause large waveform deviations. To provide gradient near the peak while preserving signal for large timing errors, we use an exponential decay.

$$r_{\text{peak}}(\hat{t}, t^*) = \exp(-|\hat{t} - t^*|). \quad (9)$$

This addresses the peak-sensitivity challenge identified in the introduction.

Waveform fidelity reward. The squared waveform error grows rapidly with per-step deviation and can dominate the combined reward. We apply a logarithmic compression to keep its contribution balanced.

$$r_{\text{wave}}(\hat{\mathbf{w}}, \mathbf{w}^*) = -\log(1 + \|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2). \quad (10)$$

This balances peak timing, magnitude, and waveform fidelity in the combined reward.

Physical constraint reward. We penalize physically implausible predictions, namely negative yields or peaks outside $[0, T]$:

$$r_{\text{constraint}} = \begin{cases} -\lambda_c & \text{if } \hat{Y}_{DT} < 0 \text{ or } \hat{t} \notin [0, T] \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

The total reward is clipped to a bounded range to prevent gradient explosion while preserving relative ranking (Schulman et al., 2017).

4 Experiments

4.1 Experimental Setup

4.1.1 Dataset and Benchmark

We introduce the **ICFBench** dataset, the first large-scale benchmark for ICF neutron yield time-series prediction. The dataset comprises (1) **Simulation data**: 50,000 samples generated from radiation-hydrodynamics codes, split into 40,000 / 5,000 / 5,000 for train / validation / test; and (2) **Experimental data**: 232 quality-controlled physical shots split into 182 training shots and a 50-shot held-out test set used for all physical-experimental results in this paper. Each sample consists of 5 scalar target parameters ($\vec{T}$) and a 256-dimensional pulse shape ($\vec{P}$), with target neutron rate ($\dot{n}$) time-series of 512 steps. All models predict the neutron rate from target design parameters. Full dataset statistics, the quality-control protocol, and the per-table evaluation ledger are in Appendix S1.

4.1.2 Training Setup

We build on LLaDA-8B (Nie et al., 2026) with LoRA adapters for parameter-efficient tuning (Hu et al., 2021; Wang et al., 2023; Han et al., 2024a). Training proceeds in two stages: supervised fine-tuning followed by PPO with the physics-driven rewards in §3.2.3. All experiments use 8 NVIDIA H100 GPUs. More details are in Appendix S2.

4.1.3 Baselines

We compare ICF-DLM against seven baselines covering classical deep time-series models (RNN (Ranzato et al., 2015), LSTM (Hochreiter and Schmidhuber, 1997), GRU (Cho et al., 2014), Transformer (Vaswani et al., 2017), FEDformer (Zhou et al., 2022)), diffusion-based time-series models (Diffusion-TS (Yuan and Qiao, 2024)), and LLM-based time-series models (TimeLLM (Jin et al., 2024)).

4.1.4 Evaluation Metrics

We report four metrics, all lower-the-better. **MSE / MAE** are squared and absolute errors on the local waveform $\mathbf{w}_{\text{local}}$. **PTE** is the Peak Timing Error, $|\hat{t}_{\text{peak}} - t_{\text{peak}}^*|$. **YE** is the Yield Error, the relative error of $\hat{Y}_{DT}$ in percent.

4.2 Benchmark Results

4.2.1 Simulation Main Results

We evaluate on the 5,000-shot simulation test set of ICFBench. As shown in Table 1, ICF-DLM achieves the best MSE, MAE, PTE, and YE across all baselines, with MSE 0.122 and YE 18.6%. Compared to classical deep time-series models, ICF-DLM reduces MSE by 60.0% over the strongest baseline Transformer (0.305) and YE by 39.8% over the strongest baseline GRU (30.9%), likely because the explicit decomposition of yield, peak timing, and local waveform captures the nonlinearity of ICF dynamics that generic deep sequence models do not target. Compared to Diffusion-TS, ICF-DLM reduces MSE by 14.1% (0.142 $\rightarrow$ 0.122) and PTE by 73.9% (30.7 $\rightarrow$ 8.0 steps), with physics-driven rewards and hierarchical decomposition providing an additional inductive bias beyond a generic diffusion backbone. Compared to TimeLLM, ICF-DLM reduces MSE by 69.1% (0.395 $\rightarrow$ 0.122) and PTE by 38.5% (13.0 $\rightarrow$ 8.0 steps), consistent with the picosecond-scale burst window benefiting from bidirectional diffusion refinement rather than left-to-right decoding.

4.2.2 Physical-Experimental Main Results

All models are pre-trained on simulation data and fine-tuned on the experimental training set.

As shown in Table 1, ICF-DLM achieves the best MSE, MAE, PTE, and YE on the 50-shot held-out test set. Compared to classical deep time-series models, ICF-DLM reduces MSE by 72.3% over the strongest baseline FEDformer (0.231) and

Table 1: Quantitative results on simulation and physical-experimental datasets for **exogenous-driven ICF waveform prediction**. Method categories: † Classical Deep Time-Series Models; ‡ Diffusion-based Time-Series Models; ‡ LLM-based Time-Series Models; ★ Ours.

Method	Simulation				Physical Experimental			
	MSE↓	MAE↓	PTE↓	YE (%)↓	MSE↓	MAE↓	PTE↓	YE (%)↓
RNN†	0.401 $\pm$ 0.002	0.550 $\pm$ 0.003	13.3 $\pm$ 0.58	35.5 $\pm$ 0.4	0.304 $\pm$ 0.014	0.472 $\pm$ 0.018	15.2 $\pm$ 0.76	303.8 $\pm$ 35
LSTM†	0.362 $\pm$ 0.018	0.513 $\pm$ 0.022	10.7 $\pm$ 2.52	34.8 $\pm$ 4.4	0.304 $\pm$ 0.019	0.472 $\pm$ 0.016	20.7 $\pm$ 6.83	404.5 $\pm$ 165
GRU†	0.338 $\pm$ 0.006	0.480 $\pm$ 0.010	9.0 $\pm$ 1.00	30.9 $\pm$ 3.2	0.236 $\pm$ 0.022	0.394 $\pm$ 0.006	25.7 $\pm$ 8.75	412.6 $\pm$ 120
Transformer†	0.305 $\pm$ 0.001	0.444 $\pm$ 0.001	8.3 $\pm$ 0.58	32.5 $\pm$ 1.6	0.238 $\pm$ 0.006	0.390 $\pm$ 0.012	25.0 $\pm$ 1.32	356.3 $\pm$ 35
FEDformer†	0.315 $\pm$ 0.004	0.452 $\pm$ 0.005	9.0 $\pm$ 0.00	33.4 $\pm$ 0.6	0.231 $\pm$ 0.023	0.400 $\pm$ 0.015	30.7 $\pm$ 1.15	252.0 $\pm$ 18
Diffusion-TS‡	0.142 $\pm$ 0.007	0.251 $\pm$ 0.011	30.7 $\pm$ 2.08	95.6 $\pm$ 1.5	0.144 $\pm$ 0.019	0.231 $\pm$ 0.026	25.0 $\pm$ 4.00	141.0 $\pm$ 55
TimeLLM‡	0.395 $\pm$ 0.003	0.539 $\pm$ 0.001	13.0 $\pm$ 0.00	35.1 $\pm$ 0.6	0.347 $\pm$ 0.004	0.452 $\pm$ 0.006	15.8 $\pm$ 0.58	148.5 $\pm$ 10
ICF-DLM (Ours)★	0.122$\pm$0.003	0.178$\pm$0.004	8.0$\pm$0.15	18.6$\pm$0.5	0.064$\pm$0.002	0.181$\pm$0.003	9.2$\pm$0.17	35.2$\pm$0.8

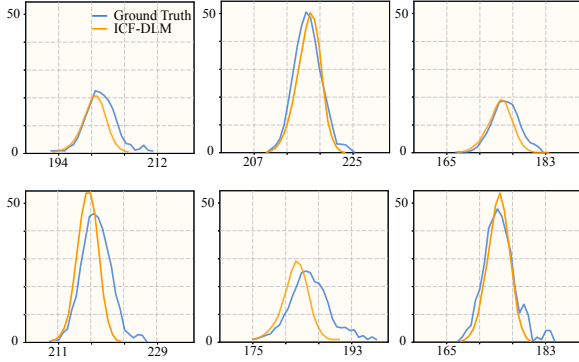


Figure 3: Comparison of ICF-DLM predictions (orange solid) versus physical experimental results (blue solid) for six expert-selected input configurations.

PTE by 39.5% over the strongest baseline RNN (15.2 steps). Classical recurrent baselines (LSTM, GRU) show notable PTE inflation on experimental data (20.7 and 25.7 steps) compared to their simulation performance (10.7 and 9.0 steps), while ICF-DLM maintains PTE within 9.2 steps across both distributions, likely due to the explicit decomposition of peak timing as a separate prediction target. Compared to TimeLLM, ICF-DLM reduces MSE by 81.6% (0.347 $\rightarrow$ 0.064) and PTE by 41.8% (15.8 $\rightarrow$ 9.2 steps), with bidirectional diffusion refinement and physics-driven rewards transferring cleanly across the sim-to-real gap. The high YE of all baselines on experimental data (above 140%) reflects the difficulty of sim-to-real transfer (Appendix S1). ICF-DLM, in contrast, transfers cleanly across the gap.

4.2.3 Real-world ICF Experiments Guided by ICF-DLM

To validate the practical utility of ICF-DLM for guiding real ICF experiments, we collaborate with domain experts to conduct a prospective validation study. Domain experts propose six input configurations of scientific interest for exploring new im-

plosion regimes, ICF-DLM predicts the expected neutron-rate waveforms before any shot is taken, the corresponding physical experiments are subsequently conducted, and predictions are evaluated against the measured outcomes.

Figure 3 compares ICF-DLM predictions and physical-experimental results. The close agreement between predicted and measured waveforms demonstrates that ICF-DLM can reliably guide experimental design by providing accurate forecasts before costly physical shots are conducted. The agreement on both peak timing and waveform amplitude suggests that the hierarchical decomposition transfers to expert-proposed configurations beyond the random held-out split, offering useful predictive information for experimental planning.

4.3 Ablation Studies

Effect of Backbone and Training Stage. We ablate two design axes. (i) The *backbone*. We replace the masked-language diffusion model (LLaDA-8B) backbone of ICF-DLM with an autoregressive model (LLaMA-3-8B) under identical numeric tokenization, prompt, and LoRA configuration. (ii) The *training stage*. We remove PPO with the physics-driven rewards and keep only SFT. As shown in Table 2, replacing AR with a DLM under the same SFT + PPO pipeline lowers MSE from 0.129 to 0.064 and PTE from 11.6 to 9.2 steps. Removing simulation pre-training raises MSE from 0.064 to 0.150 and PTE from 9.2 to 33.1 steps, confirming the value of sim-to-real transfer. Adding PPO on the DLM further lowers MSE from 0.100 to 0.064 and halves YE from 68% to 35%. The DLM gain is consistent with bidirectional denoising revising yield, peak timing, and local waveform jointly in each step. The PPO gain reflects physical information about Y_{DT} and t_{peak} that the pointwise SFT loss does not provide.

Table 2: Backbone $\times$ training stage on the test set.

Variant	MSE	PTE	YE (%)
AR LLaMA-3-8B, SFT + PPO	0.129	11.6	39
DLM LLaDA-8B, no simulation pre-train	0.150	33.1	142
DLM LLaDA-8B, SFT only	0.100	9.6	68
DLM LLaDA-8B, SFT + PPO	0.064	9.2	35

Effect of Reward Components. We ablate three reward terms that each target a distinct physical quantity. (i) The waveform reward r_{wave} supervises the local waveform shape $\mathbf{w}_{\text{local}}$. (ii) The magnitude reward r_{mag} supervises the total neutron yield Y_{DT} . (iii) The peak reward r_{peak} supervises the bang time t_{peak} . Each reward is removed in turn while the other two remain. As shown in Table 3, removing r_{wave} hurts most on MSE (from 0.064 to 0.360) and YE (from 35% to 186%). Removing r_{mag} hurts most on PTE (from 9.2 to 17.8 steps) and YE (from 35% to 148%). Removing r_{peak} inflates PTE (from 9.2 to 11.9 steps). The three rewards are synergistic, each regularizing the others through shared physical structure.

Table 3: Reward component ablation on the test set.

Variant	MSE	PTE	YE (%)
Full reward	0.064	9.2	35
W/o r_{peak}	0.087	11.9	93
W/o r_{mag}	0.128	17.8	148
W/o r_{wave}	0.360	16.4	186

Effect of Remasking Strategy. We ablate four remasking strategies that decide which tokens are re-sampled at each denoising step. (i) *Confidence* re-masks tokens with the lowest predicted probability. (ii) *Margin confidence* re-masks tokens with the smallest gap between top-1 and top-2 probabilities. (iii) *Random* re-masks uniformly at random. (iv) *Negative entropy* re-masks tokens with the highest predictive entropy. As shown in Table 4, confidence remasking gives the best numbers (MSE 0.064, PTE 9.2, YE 35%), outperforming margin confidence by 0.016 MSE, random by 0.032, and negative entropy by 0.070. Confidence remasking concentrates each refinement step on the tokens the model is least certain about, which matches the iterative denoising paradigm.

Table 4: Remasking strategy on the test set.

Variant	MSE	PTE	YE (%)
Confidence	0.064	9.2	35
Margin confidence	0.080	9.8	38
Random	0.096	10.4	43
Neg. entropy	0.134	11.7	45

Effect of Block Size. The block size controls how many tokens are denoised jointly before the next block, where larger blocks model longer-range dependency in one shot. As shown in Table 5, block size 128 gives the best numbers (MSE 0.064, PTE 9.2, YE 35%), outperforming 64 by 0.026 MSE and 32 by 0.018 MSE. Larger blocks improve global coherence across the joint generation of Y_{DT} , t_{peak} , and $\mathbf{w}_{\text{local}}$.

Table 5: Block size on the test set.

Variant	MSE	PTE	YE (%)
128	0.064	9.2	35
64	0.090	9.6	37
32	0.082	9.9	38

Effect of Denoise Steps. The number of denoise steps controls how many refinement iterations the diffusion decoder takes at inference, where more steps allow finer token correction at the cost of compute. As shown in Table 6, 128 steps gives the best numbers (MSE 0.064, PTE 9.2, YE 35%), outperforming 64 steps by 0.023 MSE and 32 steps by 0.032 MSE. More steps allow ICF-DLM to perform additional bidirectional refinement iterations over the masked numeric tokens, progressively improving prediction accuracy, so we use 128 steps as the default.

Table 6: Denoise steps on the test set.

Variant	MSE	PTE	YE (%)
Steps = 32	0.096	9.8	39
Steps = 64	0.087	9.4	37
Steps = 128	0.064	9.2	35

Effect of Architectural Choices. We further ablate three architectural choices of ICF-DLM by replacing each component while holding the rest of the pipeline fixed. (i) The *LM backbone*, replaced by a conditional MLP regressor that consumes the same (T, P) input and predicts $(Y_{DT}, t_{\text{peak}}, \mathbf{w}_{\text{full}})$ through dedicated regression heads. (ii) The *hierarchical decomposition*, replaced by a flat variant that predicts the full 512-step waveform directly without the $Y_{DT} \rightarrow t_{\text{peak}} \rightarrow \mathbf{w}_{\text{local}}$ factorization. (iii) The *numeric tokenization*, kept in place but read out with expected-value (“soft”) decoding $\sum_i p_i v_i$ at the $\langle l \rangle$ and $\langle t \rangle$ positions instead of argmax, which approximates a continuous regression head reading off the same hidden state. As shown in Table 7, removing the decomposition is the most damaging change, raising MSE from 0.064 to 0.251 (a $3.9\times$ increase) and YE to nearly

100% as the model collapses to a near-constant baseline output. Removing the LM backbone preserves peak timing (PTE 9.4 vs. ICF-DLM’s 9.2) but inflates YE from 35% to 55%, with the gap driven by the low-yield tail ($Y_{DT} < 1.5$) on which the MLP is unstable. Switching argmax to soft decoding leaves three metrics largely unchanged (PTE 13.0, YE 45%). Hierarchical decomposition contributes the largest gain among the three. The LM backbone narrows the right tail of yield error under simulation-to-experiment shift, while the discrete numeric tokenization is a representational choice rather than a performance constraint.

Table 7: Architectural choice ablations on the test set.

Variant	MSE	PTE	YE (%)
W/o decomposition	0.251	179.5	99
W/o numeric tokenization	0.064	13.0	45
W/o LM backbone	0.146	9.4	55

5 Discussion

ICF is a natural stress test for language-model-based numeric prediction. Three observations follow that should generalize beyond fusion.

Bidirectional decoding helps when outputs are globally coupled. Under matched conditions, the diffusion backbone improves over a matched-AR LLaMA-3-8B on waveform MSE with smaller gains on peak-timing and yield error (Table 2). Iterative denoising for time-series has been explored by CSDI (Tashiro et al., 2021) and Diffusion-TS (Yuan and Qiao, 2024), but those models operate on continuous values with regression backbones and assume input-output modality match; the structural advantage we observe here is specifically for LM-style discrete-token decoding under exogenous-driven prediction, where the model has no temporal scaffolding from past observations. The qualitative reason is that AR decoding commits to a peak location early, and the surrounding waveform must remain consistent with that commitment, whereas bidirectional denoising lets the peak adjust as the waveform crystallizes. We further verify this by tracing the denoising trajectory on the test set, where shots undergo bidirectional revision during decoding, reducing PTE by 2.48 steps (Appendix S4). We expect this advantage to recur for tasks whose outputs factorize into globally coupled physical quantities that resist a natural left-to-right ordering.

Numeric tokenization needs metric-aware supervision. Discrete numeric tokenization induces

a quantization error well below the typical prediction error, so quantization itself is not the bottleneck. The mismatch surfaces during RL. Cross-entropy treats adjacent numeric tokens as fully distinct classes, while the physics-driven rewards see them as nearly equivalent; the magnitude reward r_{mag} absorbs this mismatch by re-injecting metric structure across the vocabulary. Consistent with this view, removing r_{mag} degrades MSE and yield error more than removing r_{peak} (Table 3), even though r_{peak} is the only reward that directly supervises peak timing.

A recipe for LM-based scientific prediction in low-data regimes. Three design choices appear to matter for LM-based numeric prediction when scientific data are scarce and critical events are sparse. First, factorize the output along physically meaningful axes so the model is supervised on the descriptors physicists already use. Second, generate bidirectionally so commitment to globally coupled landmarks can be deferred. Third, pair token cross-entropy with at least one metric-aware reward to close the discretization gap. ICF stresses all three simultaneously; we expect the recipe to apply to other sparse-event scientific prediction tasks, an extension we leave to future work.

6 Conclusion

We presented ICF-DLM, to our knowledge the first language-model predictor for ICF neutron-rate waveforms, combining physics-typed output decomposition, bidirectional denoising that defers commitment to peak location, and a physics-driven PPO reward that restores metric structure across numeric tokens. On the held-out experimental test set, it outperforms classical sequence models, LLM-based time-series predictors, and a matched autoregressive LLaMA-3-8B, pointing to a transferable recipe for LM-based scientific prediction in low-data sparse-event regimes.

Limitations

Three caveats bound the scope of our claims. First, all experimental data come from a single ICF facility and a single diagnostic, the neutron-rate waveform; generalization to other diagnostics (X-ray imaging, neutron time-of-flight) and to facilities with different driver geometries is left to future work. Second, although the recipe conceptually applies to other scientific time-series with similar structural properties, we have not empirically validated it on domains such as tokamak disruption pre-

diction, supernova light curves, or shock-induced phase transitions. Third, diffusion generation requires multiple forward passes per prediction, so per-sample latency exceeds single-pass predictors. This is acceptable for ICF planning, where each physical shot is far more expensive than inference, but would need distillation (Han et al., 2024b) or few-step variants for real-time settings.

Ethical Considerations

This work targets scientific discovery in fusion energy research, a domain with significant potential societal benefit for clean energy generation. The ICFBench data are derived from publicly funded research at national laboratories. Our methods produce fast surrogate predictions of neutron-rate waveforms; they are not a replacement for facility-grade simulation or experimental design tools, and any practical use should retain expert review and standard quality control on the underlying physical predictions.

Inertial confinement fusion overlaps with restricted nuclear-weapons-relevant physics, and predictive surrogates of neutron yield could in principle reduce the experimental cost of designing high-yield implosions for non-peaceful applications. Our contribution is intentionally narrow, namely a tokenizer-level diffusion-LM recipe for structured numeric prediction in low-data regimes, evaluated on simulation-derived data and a small set of facility-released experimental shots. Inference latency, evaluation cost, and the small experimental shot count are explicitly low enough that the artifact does not measurably accelerate weapons-relevant design optimization beyond what existing surrogate physics tools already provide.

Acknowledgements

This research was supported by the DOE Fusion Energy Science Program under grant DE-SC0024381.

References

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

Shyamal Anadkat, et al. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.

Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar, and Hisham El-Shishiny. 2010. [An empirical comparison of machine learning models for time series forecasting](#). *Econometric reviews*, 29(5-6):594–621.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. [Structured denoising diffusion models in discrete state-spaces](#). In *NeurIPS*.

R Betti and OA Hurricane. 2016. [Inertial-confinement fusion with lasers](#). *Nature Physics*, 12(5):435–448.

Leo Breiman. 2001. [Random forests](#). *Machine learning*, 45(1):5–32.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. [Language models are few-shot learners](#). In *NeurIPS*.

SH Cao, Dhruvir Patel, Aarne Lees, C Stoeckl, MJ Rosenberg, V Gopalaswamy, Han Wen, Hu Huang, Alexander Shvydky, Riccardo Betti, et al. 2022. [Predicting hot electron generation in inertial confinement fusion with particle-in-cell simulations](#). *Physical Review E*, 106(5):055214.

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. 2019. [Machine learning and the physical sciences](#). *Reviews of Modern Physics*, 91(4):045002.

Mingkai Chen, Taowen Wang, Shihui Cao, James Chenhao Liang, Chuan Liu, Chunshu Wu, Qifan Wang, Ying Nian Wu, Michael Huang, Chuang Ren, et al. 2024. [Inertial confinement fusion forecasting via large language models](#). *arXiv preprint arXiv:2407.11098*.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. [Neural ordinary differential equations](#). In *NeurIPS*.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using rnn encoder–decoder for statistical machine translation](#). In *EMNLP*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. [Palm: Scaling language modeling with pathways](#). *Journal of machine learning research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. [Scaling instruction-finetuned language models](#). *Journal of machine learning research*, 25(70):1–53.

- Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. 2020. [Lagrangian neural networks](#). *arXiv preprint arXiv:2003.04630*.
- RS Craxton, KS Anderson, TR Boehly, VN Goncharov, DR Harding, JP Knauer, RL McCrory, PW McKenty, DD Meyerhofer, JF Myatt, et al. 2015. [Direct-drive inertial confinement fusion: A review](#). *Physics of Plasmas*, 22(11).
- J Delettrez, R Epstein, MC Richardson, PA Jaanimagi, and BL Henke. 1987. [Effect of laser illumination nonuniformity on the analysis of time-resolved x-ray measurements in uv spherical transport experiments](#). *Physical Review A*, 36(8):3926.
- Rahman Ejaz, Varchas Gopalaswamy, Riccardo Betti, Aarne Lees, and Christopher Kanan. 2024a. [Can kans \(re\) discover predictive models for direct-drive laser fusion?](#) *arXiv preprint arXiv:2409.08832*.
- Rahman Ejaz, Varchas Gopalaswamy, A Lees, C Kanan, D Cao, and R Betti. 2024b. [Deep learning-based predictive models for laser direct drive at the omega laser facility](#). *Physics of Plasmas*, 31(5).
- Jim A Gaffney, Kelli Humbird, Andrea Kritcher, Michael Kruse, Eugene Kur, Bogdan Kustowski, Ryan Nora, and Brian Spears. 2024. [Data-driven prediction of scaling and ignition of inertial confinement fusion experiments](#). *Physics of Plasmas*, 31(9).
- Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanassee, Michael McCabe, Ruben Ohana, Liam Holden Parker, et al. 2023. [xval: A continuous number encoding for large language models](#). In *NeurIPS 2023 AI for Science Workshop*.
- Valeri N Goncharov, TC Sangster, PB Radha, R Betti, TR Boehly, TJB Collins, RS Craxton, JA Delettrez, R Epstein, V Yu Glebov, et al. 2008. [Performance of direct-drive cryogenic targets on omega](#). *Physics of Plasmas*, 15(5).
- V Gopalaswamy, A Lees, R Ejaz, CA Thomas, TJB Collins, KS Anderson, W Ebmeyer, and R Betti. 2025. [Automated and highly parallelized bayesian optimization scheme for direct drive fusion experiments on omega](#). *Physical Review Research*, 7(1):013009.
- Varchas Gopalaswamy, Riccardo Betti, JP Knauer, N Luciani, D Patel, KM Woo, A Bose, IV Igumenshchev, EM Campbell, KS Anderson, et al. 2019. [Tripled yield in direct-drive laser fusion through statistical modelling](#). *Nature*, 565(7741):581–586.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Samuel Greydanus, Misko Dzamba, and Jason Yosinski. 2019. [Hamiltonian neural networks](#). In *NeurIPS*.
- Albert Gu and Tri Dao. 2023. [Mamba: Linear-time sequence modeling with selective state spaces](#). *arXiv preprint arXiv:2312.00752*.
- Albert Gu, Karan Goel, and Christopher Ré. 2021. [Efficiently modeling long sequences with structured state spaces](#). *arXiv preprint arXiv:2111.00396*.
- Cheng Han, Qifan Wang, Yiming Cui, Wenguan Wang, Lifu Huang, Siyuan Qi, and Dongfang Liu. 2024a. [Facing the elephant in the room: Visual prompt tuning or full finetuning?](#) *arXiv preprint arXiv:2401.12902*.
- Cheng Han, Qifan Wang, Sohail A Dianat, Majid Rabani, Raghuveer M Rao, Yi Fang, Qiang Guan, Lifu Huang, and Dongfang Liu. 2024b. [Amd: Automatic multi-step distillation of large-scale vision models](#). In *ECCV*.
- Jiequn Han, Arnulf Jentzen, and Weinan E. 2018. [Solving high-dimensional partial differential equations using deep learning](#). *Proceedings of the National Academy of Sciences*, 115(34):8505–8510.
- Peter Hatfield, Steven Rose, Robbie Scott, Ibrahim Almosallam, Stephen Roberts, and Matt Jarvis. 2019. [Using sparse gaussian processes for predicting robust inertial confinement fusion implosion yields](#). *IEEE Transactions on Plasma Science*, 48(1):14–21.
- Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. [Support vector machines](#). *IEEE Intelligent Systems and their applications*, 13(4):18–28.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. [Denoising diffusion probabilistic models](#). In *NeurIPS*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural computation*, 9(8):1735–1780.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *arXiv preprint arXiv:2106.09685*.
- Kelli D Humbird, J Luc Peterson, J Salmonson, and Brian K Spears. 2021. [Cognitive simulation models for inertial confinement fusion: Combining simulation and experimental data](#). *Physics of Plasmas*, 28(4).
- Kelli D Humbird and Jayson Luc Peterson. 2022. [Transfer learning driven design optimization for inertial confinement fusion](#). *Physics of Plasmas*, 29(10).
- Kelli D Humbird, Jayson Luc Peterson, BK Spears, and Ryan G McClarren. 2019. [Transfer learning to model inertial confinement fusion experiments](#). *IEEE Transactions on Plasma Science*, 48(1):61–70.

- IV Igumenshchev, VN Goncharov, FJ Marshall, JP Knauer, EM Campbell, CJ Forrest, DH Froula, V Yu Glebov, RL McCrory, SP Regan, et al. 2016. [Three-dimensional modeling of direct-drive cryogenic implosions on omega](#). *Physics of Plasmas*, 23(5).
- Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2024. [Time-llm: Time series forecasting by reprogramming large language models](#). In *ICLR*.
- Shirou Jing, Chunshu Wu, Chuan Liu, Arghavan Bahadorinejad, Feitong Qiao, Dongfang Liu, and Tony Geng. 2026. [Infodllm: an information-adaptive framework for discrete diffusion language model pre-training](#). In *ICML*.
- George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. [Physics-informed machine learning](#). *Nature Reviews Physics*, 3(6):422–440.
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. [Llms get lost in multi-turn conversation](#). In *ICLR*.
- Vincent Le Guen and Nicolas Thome. 2019. [Shape and time distortion loss for training deep time series forecasting models](#). In *NeurIPS*.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. [Deep learning](#). *nature*, 521(7553):436–444.
- A Lees, R Betti, JP Knauer, V Gopalaswamy, D Patel, KM Woo, KS Anderson, EM Campbell, D Cao, J Carroll-Nellenback, et al. 2021. [Experimentally inferred fusion yield dependencies of omega inertial confinement fusion implosions](#). *Physical review letters*, 127(10):105001.
- Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. [Diffusion-lm improves controllable text generation](#). In *NeurIPS*.
- Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. 2020. [Fourier neural operator for parametric partial differential equations](#). *arXiv preprint arXiv:2010.08895*.
- Bryan Lim, Serkan Ö Arık, Nicolas Loeff, and Tomas Pfister. 2021. [Temporal fusion transformers for interpretable multi-horizon time series forecasting](#). *International journal of forecasting*, 37(4):1748–1764.
- John Lindl. 1995. [Development of the indirect-drive approach to inertial confinement fusion and the target physics basis for ignition and gain](#). *Physics of plasmas*, 2(11):3933–4024.
- Chuan Liu, Chunshu Wu, Ming kai Chen, James Liang, Ang Li, Michael Huang, Chuang Ren, Yingnian Wu, Dongfang Liu, Tong Geng, et al. 2025. [Diff-pic: Revolutionizing particle-in-cell nuclear fusion simulation with diffusion models](#). In *ICLR*.
- Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2024. [itrans-former: Inverted transformers are effective for time series forecasting](#). In *ICLR*.
- Aaron Lou, Chenlin Meng, and Stefano Ermon. 2023. [Discrete diffusion modeling by estimating the ratios of the data distribution](#). *arXiv preprint arXiv:2310.16834*.
- Xuhui Meng and George Em Karniadakis. 2020. [A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse pde problems](#). *Journal of Computational Physics*, 401:109020.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2026. [Large language diffusion models](#). In *NeurIPS*.
- Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. [A time series is worth 64 words: Long-term forecasting with transformers](#). *arXiv preprint arXiv:2211.14730*.
- John Nuckolls, Lowell Wood, Albert Thiessen, and George Zimmerman. 1972. [Laser compression of matter to super-high densities: Thermonuclear \(ctr\) applications](#). *Nature*, 239(5368):139–142.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). In *NeurIPS*.
- Sinno Jialin Pan and Qiang Yang. 2009. [A survey on transfer learning](#). *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- PB Radha, VN Goncharov, TJB Collins, JA Delettrez, Y Elbaz, V Yu Glebov, RL Keck, DE Keller, JP Knauer, JA Marozas, et al. 2005. [Two-dimensional simulations of plastic-shell, direct-drive implosions on omega](#). *Physics of Plasmas*, 12(3).
- PB Radha, M Hohenberger, DH Edgell, JA Marozas, FJ Marshall, DT Michel, MJ Rosenberg, W Seka, A Shvydky, TR Boehly, et al. 2016. [Direct drive: Simulations and results from the national ignition facility](#). *Physics of plasmas*, 23(5).
- Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. [Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations](#). *Journal of Computational physics*, 378:686–707.
- Marc’ Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. [Sequence level training with recurrent neural networks](#). *arXiv preprint arXiv:1511.06732*.

- Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and F Prabhat. 2019. [Deep learning and process understanding for data-driven earth system science](#). *Nature*, 566(7743):195–204.
- Lars Ruthotto and Eldad Haber. 2020. [Deep neural networks motivated by partial differential equations](#). *Journal of Mathematical Imaging and Vision*, 62(3):352–364.
- S Rasoul Safavian and David Landgrebe. 1991. [A survey of decision tree classifier methodology](#). *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674.
- Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. [Simple and effective masked diffusion language models](#). In *NeurIPS*.
- David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. [Deepar: Probabilistic forecasting with autoregressive recurrent networks](#). *International journal of forecasting*, 36(3):1181–1191.
- Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. 2020. [Learning to simulate complex physics with graph networks](#). In *ICML*.
- Nicholas I Sapankevych and Ravi Sankar. 2009. [Time series prediction using support vector machines: a survey](#). *IEEE computational intelligence magazine*, 4(2):24–38.
- Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. 2021. [Step-unrolled denoising autoencoders for text generation](#). *arXiv preprint arXiv:2112.06749*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *arXiv preprint arXiv:1707.06347*.
- Lifeng Shen and James Kwok. 2023. [Non-autoregressive conditional diffusion models for time series prediction](#). In *ICML*.
- Ruibing Song, Chuan Liu, Chunshu Wu, Ang Li, Dongfang Liu, Yingnian Wu, and Tong Geng. 2025. [Ds-llm: Leveraging dynamical systems to enhance both training and inference of large language models](#). In *ICLR*.
- Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. [Csdi: Conditional score-based diffusion models for probabilistic time series imputation](#). In *NeurIPS*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. [Llama: Open and efficient foundation language models](#). *arXiv preprint arXiv:2302.13971*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *NeurIPS*.
- Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2025. [Chattime: A unified multimodal time series foundation model bridging numerical and textual data](#). In *AAAI*.
- Qifan Wang, Yuning Mao, Jingang Wang, Hanchao Yu, Shaoliang Nie, Sinong Wang, Fuli Feng, Lifu Huang, Xiaojun Quan, Zenglin Xu, et al. 2023. [Aprompt: Attention prompt tuning for efficient adaptation of pre-trained language models](#). In *EMNLP*.
- Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. 2024. [Timemixer: Decomposable multiscale mixing for time series forecasting](#). In *ICLR*.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. [Finetuned language models are zero-shot learners](#). *arXiv preprint arXiv:2109.01652*.
- Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. [Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting](#). In *NeurIPS*.
- Xinyu Yuan and Yan Qiao. 2024. [Diffusion-ts: Interpretable diffusion for general time series generation](#). In *ICLR*.
- Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. [Informer: Beyond efficient transformer for long sequence time-series forecasting](#). In *AAAI*.
- Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. [Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting](#). In *ICML*.
- Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2023. [One fits all: Power general time series analysis by pretrained lm](#). In *NeurIPS*.

APPENDIX

This appendix contains additional details for the EMNLP 2026 submission, titled “*Decomposition-Guided Diffusion Language Models for Inertial Confinement Fusion Prediction*”. The appendix is organized as follows:

- §S1 provides detailed descriptions of the ICFBench dataset, including data collection, preprocessing, and statistics.
- §S2 extends the Methods formalism with the full training algorithm, hyperparameters, LoRA configuration, prompt template, and inference settings.
- §S3 describes baseline implementations.
- §S4 analyzes denoising trajectories to characterize how bidirectional revision contributes to peak placement.
- §S5 stratifies per-shot errors by ground-truth Y_{DT} magnitude across yield bins.
- §S6 presents qualitative case studies with visualization of predictions.
- §S7 discusses the physical interpretation of hierarchical decomposition.
- §S8 lists artifact licenses.

S1 Dataset Details

S1.1 ICFBench Dataset Overview

The ICFBench dataset is constructed from two complementary sources to enable both large-scale training and realistic evaluation:

Simulation Data We utilize 50,000 samples generated from radiation-hydrodynamics (rad-hydro) simulation codes. These simulations span diverse implosion regimes by systematically varying laser pulse shapes, target specifications, and experimental configurations. Each simulation provides:

- **Input:** 5 scalar target parameters ($\vec{T}$) governing target geometry and a 256-dimensional pulse shape ($\vec{P}$).
- **Output:** Full neutron rate ($\dot{n}$) time-series of 512 steps, total yield Y_{DT} , and peak timing t_{peak} .

Experimental Data The experimental subset consists of 232 physical shots from real ICF experiments, partitioned into 182 training shots and 50 held-out test shots.

S1.2 Target Parameter Ranges

Table S1 summarizes the ranges of target parameters ($\vec{T}$) in both simulation and experimental datasets.

Table S1: Target design parameter ranges for simulation and experimental data.

Param	Simulation		Experimental	
	Range	Mean	Range	Mean
T_0	[0.42, 0.76]	0.64	[0.51, 0.73]	0.62
T_1	[0.60, 2.97]	1.47	[0.88, 2.90]	1.52
T_2	[0.08, 0.90]	0.42	[0.34, 0.68]	0.56
T_3	[1.05, 1.47]	1.26	[1.04, 1.49]	1.28
T_4	[0.00, 3.90]	1.95	[0.25, 4.62]	0.57

S1.3 Yield Distribution

The neutron yield Y_{DT} (in units of 10^{19}) exhibits different distributions between simulation and experimental data:

- **Simulation:** Range $[7.2 \times 10^{-4}, 27.0]$, mean 1.75 ± 1.82 , median 1.13
- **Experimental:** Range $[0.16, 17.5]$, mean 4.19, reflecting selection bias toward higher-yield shots in physical experiments

S1.4 Dataset Statistics Summary

Table S2 summarizes key statistics of the ICFBench dataset.

Table S2: ICFBench dataset statistics.

Statistic	Simulation	Experimental
Total samples	50,000	232
Training set	40,000	182
Validation set	5,000	—
Test set	5,000	50
Pulse shape	256	256
Target parameters	5	5
Output length	512	512
Y_{DT} range	$[7.2 \times 10^{-4}, 27.0]$	$[0.16, 17.5]$
Y_{DT} mean	1.75	4.19

S2 Implementation Details

S2.1 Pipeline Components

ICF-DLM has three components (Fig. 2). The *input encoder* serializes $\vec{P}$ and $\vec{T}$ into a structured prompt, with continuous values discretized into numeric tokens. The *LLaDA-8B backbone* (Nie et al., 2026) denoises $(Y_{DT}, t_{\text{peak}}, \mathbf{w}_{\text{local}})$ jointly under bidirectional attention. The *output decoder* embeds $\hat{\mathbf{w}}_{\text{local}}$ at $\hat{t}_{\text{peak}}$ within a 512-step trajectory, filling the remainder with a near-zero baseline.

S2.2 Data Preprocessing and Output Reconstruction

All input features and output waveforms are min-max normalized to $[-0.5, 0.5]$. Following the numeric-tokenization recipe of ChatTime (Wang et al., 2025), continuous values are discretized into a vocabulary of 10,000 numeric tokens, enabling the diffusion LM to perform regression through classification over the token vocabulary. For the hierarchical decomposition, we extract a local waveform of length $T_{\text{local}}=19$ centered at the ground-truth peak time during training; $T_{\text{local}}=19$ matches the physical neutron-burst full width in our dataset, so $\mathbf{w}_{\text{local}}$ covers the entire active region while the surrounding 493 steps carry negligible signal. At inference, the predicted local waveform is embedded into a 512-step sequence at the predicted peak position, with remaining positions filled by a near-zero baseline ($\epsilon \approx 10^{-6}$).

S2.3 Training Algorithm

Algorithm 1 summarizes the overall training procedure of ICF-DLM.

S2.4 Model and Training Configuration

We implement ICF-DLM with LLaDA-8B (Nie et al., 2026) as the bidirectional Transformer backbone (hidden dim 4096, 32 layers, 32 attention heads). The vocabulary extends the base LLaDA vocabulary of 128,256 with 10,006 additional numeric and special tokens for waveform discretization. We fine-tune with Low-Rank Adaptation (LoRA) (Hu et al., 2021) at rank $r=16$, $\alpha=64$, and dropout 0.05, applied to the q/k/v/o and gate/up/down projections ($\sim 42\text{M}$ trainable, 0.5% of the full model).

Supervised fine-tuning runs for 5 epochs on 40,000 samples with batch size 32, learning rate 2×10^{-4} under a cosine schedule with 3% warmup, AdamW optimizer ($\beta_1=0.9$, $\beta_2=0.999$, weight decay 0.01), and max sequence length 512. PPO follows for 200 iterations with batch size 8, learning rate 1×10^{-5} , KL coefficient $\beta=0.01$, clip ratio $\epsilon=0.2$, batch-level advantage normalization, and reward clipping to $[-3, 3]$. At inference, we use 128 denoising steps with confidence-based unmasking and greedy decoding ($\tau=0$).

S2.5 Prompt Template

Table S3 illustrates the prompt format used for training and inference. All continuous values

Algorithm 1 ICF-DLM Training Procedure

Require: Training data $\mathcal{D} = \{(\vec{T}_i, \vec{P}_i, \mathbf{y}_i)\}_{i=1}^N$, denoising steps S , learning rate η

Ensure: Trained model parameters θ

- 1: Initialize LLaDA backbone with LoRA adapters
 - 2: **// Stage 1: Supervised Fine-Tuning**
 - 3: **for** each epoch in SFT stage **do**
 - 4: **for** each batch $(\vec{T}, \vec{P}, \mathbf{y}) \in \mathcal{D}$ **do**
 - 5: Decompose target: $\mathbf{y} \rightarrow (Y_{DT}, t_{\text{peak}}, \mathbf{w}_{\text{local}})$
 - 6: Tokenize and construct structured prompt
 - 7: Sample mask ratio $t \sim \text{Uniform}(0, 1)$
 - 8: Forward process: $\mathbf{x}_t = \text{Mask}(\mathbf{x}_0, t) \triangleright$ mask t -fraction of tokens
 - 9: Predict masked tokens: $\hat{\mathbf{x}}_0 = f_{\theta}(\mathbf{x}_t, \vec{T}, \vec{P})$
 - 10: $\mathcal{L}_{\text{SFT}} = -\sum_{i \in \mathcal{M}} \log p_{\theta}(x_0^{(i)} | \mathbf{x}_t)$
 - 11: Update $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}_{\text{SFT}}$
 - 12: **end for**
 - 13: **end for**
 - 14: **// Stage 2: RL with Physics-Driven Rewards**
 - 15: **for** each epoch in RL stage **do**
 - 16: **for** each batch $(\vec{T}, \vec{P}, \mathbf{y}) \in \mathcal{D}$ **do**
 - 17: Generate $(\hat{Y}_{DT}, \hat{t}_{\text{peak}}, \hat{\mathbf{w}}_{\text{local}})$ via S -step denoising
 - 18: $r_{\text{mag}} \leftarrow \frac{1}{1 + |\hat{Y}_{DT} - Y_{DT}^*| / Y_{DT}^*}$
 - 19: $r_{\text{peak}} \leftarrow \exp(-|\hat{t}_{\text{peak}} - t^*|)$
 - 20: $r_{\text{wave}} \leftarrow -\log(1 + \|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2)$
 - 21: $r = r_{\text{mag}} + r_{\text{peak}} + r_{\text{wave}} + r_{\text{constraint}}$
 - 22: Update θ via PPO with reward r
 - 23: **end for**
 - 24: **end for**
 - 25: **return** θ
-

are normalized to $[-0.5, 0.5]$ and discretized into 10,000 tokens.

S3 Baseline Implementations

Classical baselines. We implement RNN, LSTM, and GRU as 2-layer bidirectional networks with hidden size 256 and dropout 0.1. The Transformer baseline is a standard encoder-decoder with hidden dim 512, 8 attention heads, 6 encoder/decoder layers, and feedforward dim 2048. FEDformer uses the official implementation with default hyperparameters (Zhou et al., 2022).

LLM-based baselines. TimeLLM (Jin et al., 2024)

Table S3: Prompt template for ICF-DLM. Values are normalized and tokenized using the `###value###` format.

Component	Format
Input Prompt	
Task	Predict L (yield), T (peak), W (waveform)
Target Params	$T_0=0.644$, $T_1=1.445$, $T_2=0.564$, $T_3=1.360$, $T_4=0.455$
Pulse Shape	###-0.4961### ###-0.4903### ###-0.4961### ... (256 tokens)

uses LLaMA-3-8B with a reprogramming layer for time-series adaptation, following the original recipe.

S4 Denoising-Trajectory Analysis

The matched-AR comparison in the main text shows ICF-DLM improving over an AR backbone. We examine whether the diffusion backbone uses bidirectional revision during decoding. We instrument greedy decoding on all $N=50$ test shots and log, at every step, the argmax over numeric tokens at the $\langle t \rangle$ position. Table S4 reports the trajectory statistics. 60% of test shots show some peak revision, 44% show ≥ 10 -step revision, and 34% show ≥ 20 -step revision, with revisions contributing an average PTE reduction of 2.48 steps over the denoising trajectory. A control that forces $\langle t \rangle$ to commit at step 0 and freezes it for the remaining 127 steps confirms this gap, consistent with bidirectional revision contributing to peak placement.

Table S4: Denoising-trajectory statistics on the test set ($N=50$). See §S4 for the protocol.

Statistic	Value
Shots with any peak revision ($\text{span} \geq 1$)	30 / 50 (60%)
Shots with significant revision ($\text{span} \geq 10$)	22 / 50 (44%)
Shots with strong revision ($\text{span} \geq 20$)	17 / 50 (34%)
PTE reduction across the denoising trajectory	-2.48 steps
PTE increase under the frozen- $\langle t \rangle$ control	+2.48 steps
Revisions that move $\langle t \rangle$ toward GT	68%
Mean PTE improvement on revised shots	-6.53 steps

S5 Yield-Regime Stratification

We stratify per-shot errors on the held-out test set by ground-truth Y_{DT} magnitude (Table S5). PTE is stable across the three yield bins (low, mid, high), indicating that peak localization does not degrade for low-yield shots. YE decreases monotonically from the low-yield bin to the high-yield bin, reflecting that the denominator of the relative-error metric

grows with yield magnitude.

Table S5: Yield-stratified per-shot error on the test set.

Bin	n	PTE	YE (%)	MSE
overall	50	9.2	35.1	0.064
low ($Y_{DT} < 1.5$)	8	7.1	64.9	0.184
mid ($1.5 \leq Y_{DT} < 5$)	23	10.7	33.6	0.043
high ($Y_{DT} \geq 5$)	19	8.2	24.5	0.040

S5.1 Per-Component Reward Analysis

Table S6 shows the breakdown of reward components during RL training, demonstrating that all three objectives improve over training.

Table S6: Reward component breakdown (mean $\pm$ std over last 50 training steps).

Stage	r_{peak}	r_{mag}	r_{wave}
Initial (SFT)	0.42 ± 0.15	0.38 ± 0.12	-1.24 ± 0.31
After RL	0.68 ± 0.11	0.56 ± 0.09	-0.72 ± 0.18

S5.2 Comparison with Y_{DT} Surrogate

Under the same SFT-only training setting (40,000 samples, 5 epochs), a dedicated Y_{DT} surrogate model that only predicts the scalar energy descriptor achieves 83.6% Y_{DT} yield error. In contrast, our method under the same SFT-only setting not only produces full waveform predictions but also achieves better Y_{DT} accuracy (68% YE, cf. SFT only row in Table 2) as a byproduct of the hierarchical decomposition.

S6 Case Studies

S6.1 Successful Predictions

Figure S1 shows eight representative cases spanning Y_{DT} values from 2.6 to 5.7 under diverse pulse shapes. Predicted Y_{DT} deviates from ground truth by below 3% and peak timing aligns with $\Delta t=0$ across all cases, indicating that the hierarchical decomposition generalizes across pulse-shape regimes.

S6.2 Failure Mode Analysis

Prediction quality degrades in the tails of the yield distribution. Shots with $Y_{DT} < 0.5$ or $Y_{DT} > 5.0$ carry the largest relative error. Table S5 shows the same trend across yield bins. Both tails are thinly covered by the training data. Targeted augmentation or importance-weighted training may close this gap.

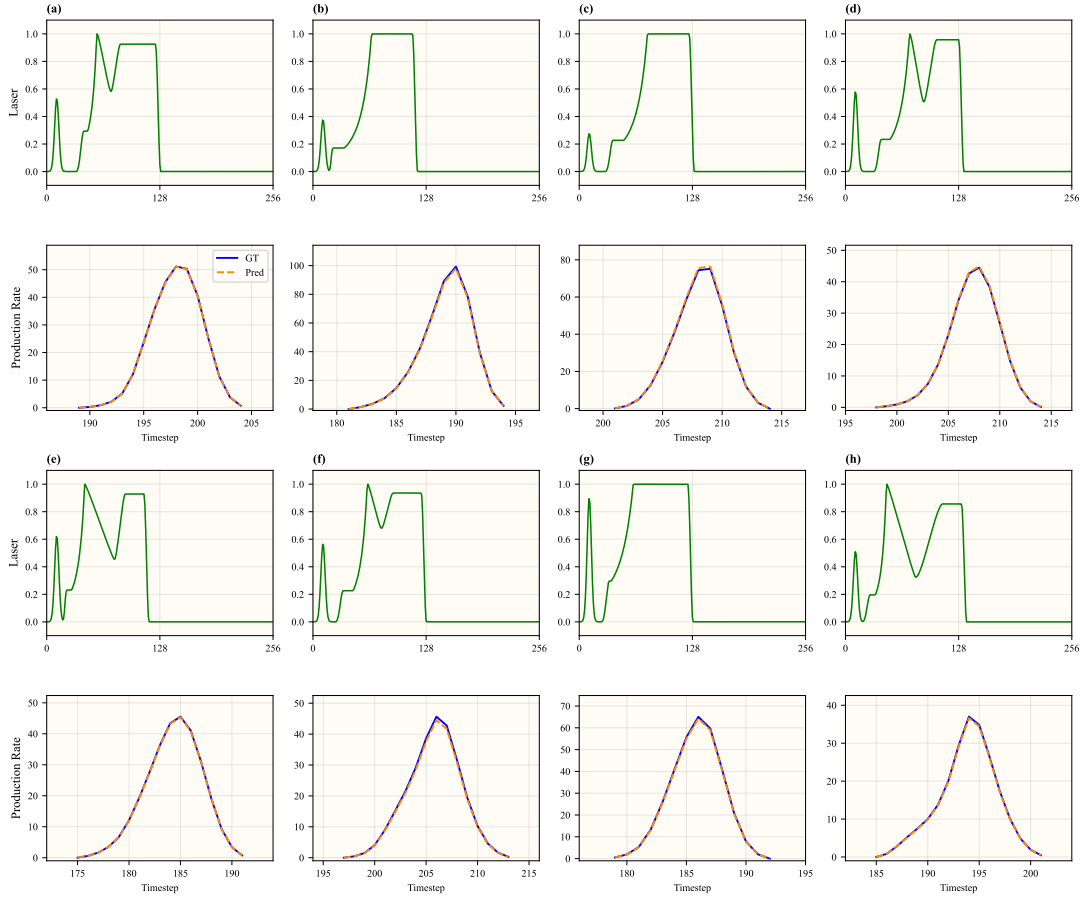


Figure S1: Successful prediction cases on simulation data. Top rows show laser input waveforms; bottom rows show Y_{DT} output with ground truth (blue) and ICF-DLM prediction (orange dashed). Labels indicate Y_{DT} values (GT→Pred) and timing error (Δt).

S7 Physical Interpretation

The hierarchical decomposition in ICF-DLM mirrors the workflow employed by ICF physicists when analyzing radiation-hydrodynamics simulation outputs. Domain experts typically begin by examining total neutron yield to assess overall implosion performance. They then identify bang time to understand compression dynamics and hot-spot formation. Finally, they analyze the detailed burn rate profile around peak emission to diagnose burn propagation and confinement. By encoding this expert workflow into the model architecture, we provide strong inductive biases that improve both prediction accuracy and physical interpretability.

S7.1 Learned Representations

Analysis of attention patterns reveals that the model learns physically meaningful relationships. Peak timing prediction attends strongly to laser pulse rise time and peak power, consistent with the known physics of shock timing in ICF implosions. Yield

prediction correlates with integrated laser energy and target coupling efficiency indicators, reflecting the energy balance considerations central to ICF performance. Local waveform generation shows strong dependence on the predicted Y_{DT} and t_{peak} values, confirming that the hierarchical conditioning mechanism effectively propagates information across prediction stages.

S8 Artifact Licenses

LLaDA-8B (Nie et al., 2026) is licensed under MIT; LLaMA-3-8B (Grattafiori et al., 2024) is licensed under Llama 3 Community License Agreement; LoRA (Hu et al., 2021), PPO (Schulman et al., 2017), and the baseline implementations (Zhou et al., 2022; Yuan and Qiao, 2024; Jin et al., 2024) are licensed under Apache-2.0 or MIT. Our use is consistent with their intended research purposes.