Co4ICF: Co-evolving Physics-Informed Surrogate and RL-based Pulse Optimizer for Inertial Confinement Fusion

Jiatong Zhao*
zhaojiatong@sjtu.edu.cn
School of Physics and Astronomy
Zhiyuan College
Shanghai Jiao Tong University
Shanghai, China

Tengyue Zhang*
zhangty_23@sjtu.edu.cn
School of Physics and Astronomy
Zhiyuan College
Shanghai Jiao Tong University
Shanghai, China

Yuhan Wang*
wang.yuhan@sjtu.edu.cn
School of Physics and Astronomy
Zhiyuan College
Shanghai Jiao Tong University
Shanghai, China

Fuyuan Wu
fuyuan.wu@sjtu.edu.cn
School of Physics and Astronomy
Shanghai Jiao Tong University
Shanghai, China

Junchi Yan†
yanjunchi@sjtu.edu.cn
School of Artificial Intelligence
Shanghai Jiao Tong University
Shanghai, China

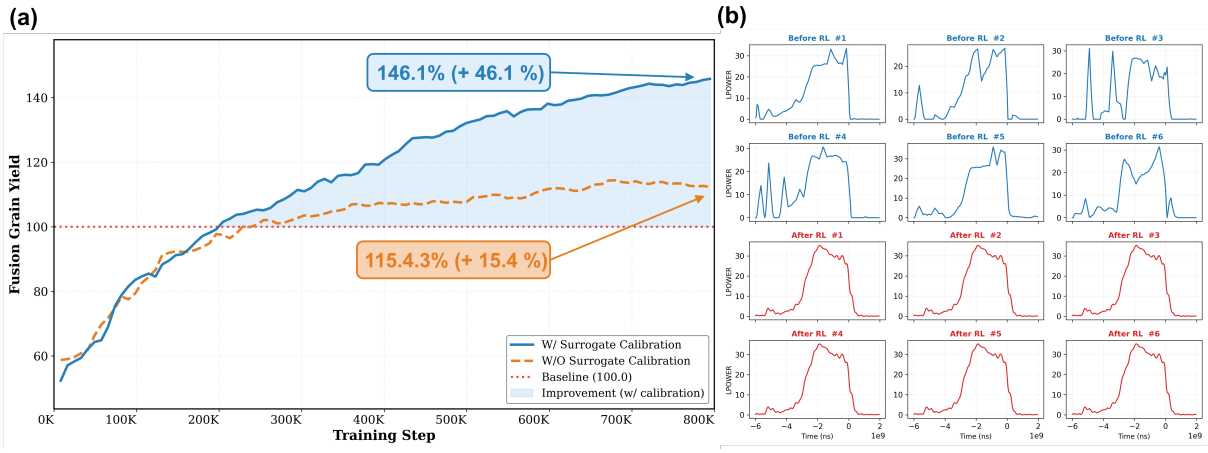

Figure 1: Performance of Co4ICF in laser pulse optimization. (a) Fusion Yield: The Co4ICF framework (blue solid line) dynamically calibrates the surrogate, reaching 146.1% normalized yield in the 1D-MULTI optimization loop, i.e., a 46.1% improvement over the designed baseline (red dotted line). This outperforms the static surrogate approach (orange dashed line, +15.4%), which suffers from distribution shift. (Re-evaluated by 1D-MULTI; see Table 3 for final direct 2D-MULTI evaluation.) (b) Pulse Optimization: Comparison between the initial random pulse samples (top rows) and the final RL-optimized waveforms (bottom rows), showing the pulse structures found by the policy.

Abstract

Offline-trained surrogates for Inertial Confinement Fusion (ICF) suffer a well-known failure mode that iterative optimizers drive inputs into out-of-distribution (OOD) regions where predictions become unreliable. Here we present **Co4ICF**, a co-evolving framework that couples a physics-informed surrogate with a PPO-based pulse optimizer. The surrogate is iteratively fine-tuned on policy-induced trajectories, correcting extrapolation errors as the optimizer shifts the input distribution; the optimizer queries this evolving surrogate as a fast environment. In the 1D MULTI environment, Co4ICF achieves **146.1%** normalized yield based on current laser design baseline; as a post-hoc cross-fidelity check, the optimized pulse further attains **246.9%** normalized yield when directly evaluated in 2D-MULTI without any 2D training or fine-tuning. Budget-matched ablations support that the gains are not explained solely by additional simulation data and are consistent with the co-evolving mechanism playing a key role. We release a large-scale MULTI-IFE simulation dataset to support future benchmarking.

*These authors contributed equally to this research.
†Corresponding author.

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818998

CCS Concepts

• **Applied computing** → **Physics**; • **Computing methodologies** → **Modeling methodologies**; *Machine learning*; *Artificial intelligence.*

Keywords

Inertial Confinement Fusion, Surrogate Modeling, Reinforcement Learning, Pulse Optimization

ACM Reference Format:

Jiatong Zhao, Tengyue Zhang, Yuhan Wang, Fuyuan Wu, and Junchi Yan. 2026. Co4ICF: Co-evolving Physics-Informed Surrogate and RL-based Pulse Optimizer for Inertial Confinement Fusion. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3770855.3818998

1 Introduction

High-fidelity numerical simulations are a cornerstone of Inertial Confinement Fusion (ICF) research, but expensive simulations and manual parameter tuning limit their use in design optimization. Recent AI for Fusion research has begun to alleviate these bottlenecks by applying offline pretrained deep neural surrogates [5, 7, 22] and RL-based optimization policies [3, 4]. Given the high dimensional design space, it is difficult for an offline-trained surrogate to span a broad solution manifold, so the performance can degrade sharply under OOD inputs during iterative optimization [20]. RL methods in contrast require massive on-policy interaction for policy updates, while using high-fidelity simulators as the training environment is not computationally practical. A further concern is physical consistency, that surrogates and policies trained purely on data can violate conservation laws or known scaling relations [10, 15].

To address these bottlenecks, we present **Co4ICF**: a co-evolving ICF design framework that couples a physics-informed surrogate with a PPO pulse optimizer. As illustrated in Fig. 2, Co4ICF couples two modules: (i) a physics-informed *surrogate* that replaces expensive radiation hydrodynamics simulations (MULTI), and (ii) an RL-based laser pulse *designer* that parameterizes and searches laser pulses under feasibility and energy constraints. The core mechanism is a *co-evolving loop*: the designer is updated at high frequency using the surrogate as a cheap environment, while the surrogate is periodically fine-tuned at lower frequency on policy-induced trajectories to correct extrapolation errors caused by distribution shift.

The insight behind this loop is that optimizer-induced OOD drift, conventionally treated as a failure mode, can be recast as a training signal: the very distribution shift that degrades a static surrogate instead provides on-policy data for its refinement. In the 1D-MULTI optimization loop, Co4ICF reaches **146.1%** normalized yield relative to the designed-pulse baseline while delivering a **990**× rollout-time reduction for high-frequency policy updates. As a post-hoc cross-fidelity check, the final optimized pulse is directly evaluated in 2D-MULTI and reaches **246.9%** normalized yield under the same baseline normalization, without using any 2D samples for training or fine-tuning. Budget-matched ablations confirm that these gains are not explained solely by additional simulation data and are consistent with the co-evolving mechanism playing a key role. We provide full access to Co4ICF dataset and implementation: dataset at: https://huggingface.co/datasets/Oyhs/Co4ICFDataset; code at: https://github.com/Co4ICF/co4icf.

In short, our main contributions are:

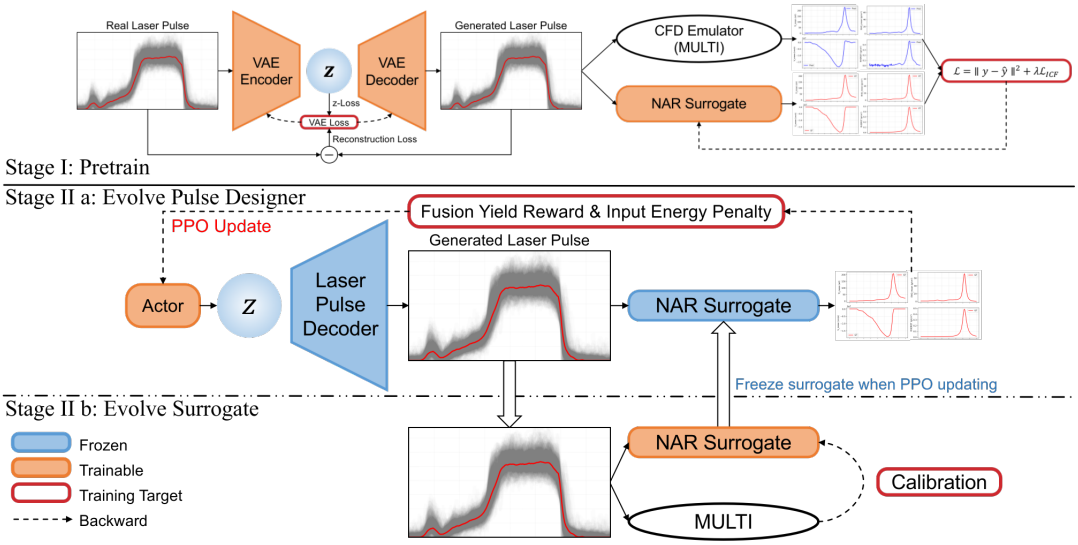

Figure 2: Co4ICF framework. Stage I (Pretrain): A VAE decoder learns a latent representation of laser pulses; the surrogate is trained to mimic MULTI's dynamics with physics-informed regularization. Stage II (Evolve): The actor samples latent codes, decodes them into pulses, and queries the NAR surrogate for rewards. High-frequency PPO updates optimize the actor; low-frequency surrogate fine-tuning on MULTI-verified on-policy samples corrects distribution shift.

(1) We propose a *co-evolving* framework that jointly updates an ICF surrogate and a pulse designer, addressing optimizer induced distribution shift in iterative learned-model-based loops.
(2) We develop a *PPO-based* pulse optimizer that searches in a low-dimensional VAE latent pulse manifold and queries the physics-regularized surrogate as a fast environment under energy and feasibility constraints.
(3) Our pipeline reaches **146.1%** normalized yield in the 1D-MULTI optimization loop and **246.9%** normalized yield in post-hoc direct 2D-MULTI evaluation, while accelerating high frequency policy rollouts by ∼**990**× within 1D search space.
(4) We release a large-scale 1D MULTI simulation dataset and the implementation to facilitate reproducible benchmarking.

2 Background

2.1 Inertial Confinement Fusion and Simulation

Inertial Confinement Fusion (ICF) is one of the major approaches toward realizing controlled nuclear fusion energy. ICF uses high-energy drivers (like lasers or pulsed power) to rapidly compress a fuel capsule to extreme densities and temperatures, triggering a thermonuclear burn wave before the target disassembles. In December 2022, the National Ignition Facility (NIF) achieved target gain $G > 1$ for the first time, producing 3.15 MJ of fusion yield from 2.05 MJ of laser energy, and subsequent experiments [9] have replicated this result.

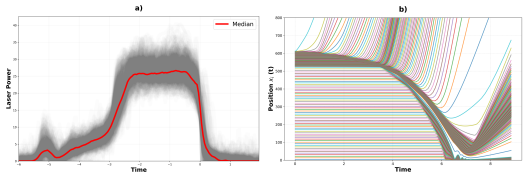

Figure 3: Inputs and outputs of the MULTI-IFE simulation. (a) Distribution of input laser pulses across the dataset. (b) an example 1D-MULTI output: each curve corresponds to a Lagrangian mass layer, showing ablation-driven compression (decreasing radius) followed by ignition.

Given the extreme nonlinearities and prohibitive experimental costs, large-scale numerical simulation is indispensable for interpreting implosion dynamics and optimizing pulse designs. This work uses MULTI-IFE (also "MULTI"), an open-source radiation hydrodynamics code [16]. MULTI is a Lagrangian code designed for Inertial Fusion Energy studies; it captures multigroup radiation diffusion, separate electron/ion temperatures, and tabulated equations of state while remaining computationally efficient. The three-stage implosion dynamics captured by MULTI are detailed in Appendix B.1.

2.2 Related Work

Physics-informed surrogates. Such constraints improve data efficiency and generalization in PDE surrogates. Physics-informed Neural Networks (PINNs) [15] enforce governing equations as soft penalties, and this idea has been adapted to ICF surrogate modeling. Early ICF surrogates [1, 7] focused on efficient mapping and physical consistency. Recent work uses facility-specific architectures [5] and Transformers [14, 22]. Transfer learning [8] bridges the simulation–experiment gap with sparse experimental data, culminating in predictive capabilities for NIF ignition [17]. However, a shared limitation is that these surrogates are trained offline and can degrade sharply under OOD inputs, which is precisely the regime that iterative optimizers actively explore.

ICF pulse optimization. On the optimization side, early ICF design relied on heuristic and evolutionary strategies [12, 19], later augmented with genetic algorithms [24] and DNN surrogates [23]. Bayesian Optimization (BO) has since become a standard approach, from multi-fidelity simulation frameworks [21] to automated experimental campaigns [6]. RL remains scarce in ICF compared to Tokamak control [4], with Capuano et al. [3] as a rare application for pulse shaping. A key difficulty common to all iterative methods is that training on a static, offline surrogate leads to *model exploitation*: the optimizer discovers OOD inputs that score high under the surrogate but are physically invalid, motivating approaches where the surrogate and optimizer adapt jointly.

Evolving surrogate and optimizer. The idea of alternating model refinement with optimization is well established: the Dyna architecture in model-based RL interleaves real experience with model updates [18], Surrogate-Assisted Evolutionary Algorithms periodically retrain the surrogate on evaluated candidates [13], and offline Model-Based Optimization methods address distribution shift through conservative objectives or data-manifold constraints [11, 20]. A common thread is that these methods either penalize predictive uncertainty or restrict search to the offline data manifold to avoid OOD exploitation.

3 Co4ICF

We present **Co4ICF**, a closed-loop framework that jointly optimizes a pulse designer and a physics-informed surrogate.

3.1 Pipeline

Overview. Co4ICF is a closed-loop framework with a *pulse designer* and a *physics-informed surrogate* (Fig. 4). Let $\mathbf{t} \in \mathbb{R}^{d_t}$ denote static target/capsule parameters and $\mathbf{p} \in \mathbb{R}^T$ denote a laser power waveform discretized on a fixed grid. The *designer* operates in a low-dimensional latent action space: an *actor*, or policy network $\pi_\omega(\cdot \mid \mathbf{t})$ outputs a distribution over latent codes $\mathbf{z} \in \mathbb{R}^{d_z}$, and a pretrained *decoder* $\mathbf{D}_\psi$ maps $\mathbf{z}$ to a feasible waveform $\mathbf{p} = \mathbf{D}_\psi(\mathbf{z})$. Given $(\mathbf{t}, \mathbf{p})$, the *surrogate* S_θ predicts time-resolved trajectories $\hat{\mathbf{y}} = S_\theta(\mathbf{t}, \mathbf{p})$ from which we compute a scalar optimization objective $r = R(\hat{\mathbf{y}})$ that encodes ignition reward together with feasibility/energy penalties. The designer proposes candidate pulses and the surrogate provides fast, physically consistent evaluations. To mitigate OOD drift and reward hacking, we interleave *high-frequency* actor updates with *low-frequency* surrogate refinement on newly queried on-policy trajectories.

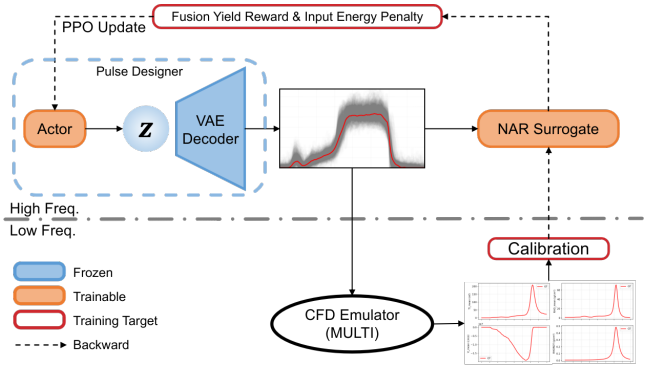

Figure 4: Co-evolving Stage of Co4ICF. The PPO actor queries the surrogate at high frequency for policy updates; every K steps, on-policy pulses are evaluated with MULTI and the surrogate is fine-tuned on the augmented dataset.

Stage I: offline pretraining. We use offline MULTI data $\mathcal{D}_0$. Each sample contains target parameters, a laser pulse, and a MULTI trajectory: $(\mathbf{t}_i, \mathbf{p}_i, \mathbf{y}_i)$ (Section 3.2). First, we pretrain the VAE on a broad pulse subset to learn a feasible pulse manifold, and retain its decoder $\mathbf{D}_\psi$ as a fixed mapping from the latent prior to waveform space. This converts waveform search into a lower-dimensional control problem and constrains exploration to the learned manifold. In parallel, we pretrain the surrogate S_θ to regress MULTI trajectories. Beyond standard supervised regression, we introduce inequality-based physics regularization to discourage non-physical trajectories and improve extrapolation stability (Section 3.3).

Stage II: co-evolving optimization with dual-frequency updates. After pretraining, we alternate between fast policy improvement against the surrogate and periodic surrogate refinement on on-policy samples. In the high-frequency loop, we treat S_θ as an inexpensive environment: at each PPO iteration, the actor samples $\mathbf{z} \sim \pi_\omega(\cdot \mid \mathbf{t})$, decodes it into $\mathbf{p} = \mathbf{D}_\psi(\mathbf{z})$, queries S_θ for $\hat{\mathbf{y}} = S_\theta(\mathbf{t}, \mathbf{p})$, and computes the reward $r = R(\hat{\mathbf{y}})$; PPO then updates the policy using batches of surrogate-generated rollouts. In the low-frequency loop, we refine S_θ every K PPO updates: we sample pulses from the current policy, re-evaluate a budgeted subset with MULTI, and fine-tune the surrogate on the union of the offline data and these newly labeled on-policy samples.

3.2 Dataset

A major challenge in AI for fusion is the lack of a common benchmark. We construct a synthesized simulation dataset that combines multiple input sampling strategies to broaden coverage while preserving physical plausibility. The dataset contains three subsets (Fig. 5). The *Real* subset contains measured laser pulses from ICF experiments, i.e., partially optimized solutions in a limited design space. The *Clean* subset samples a VAE trained only on *Real*, yielding *in-distribution* pulses on the learned manifold; this dataset VAE is distinct from the optimization VAE. The *Other* subset uses noise augmentation to cover *out-of-distribution* patterns. Each pulse is simulated with MULTI to obtain time-resolved implosion trajectories and associated physical quantities. Details are in Appendix A.

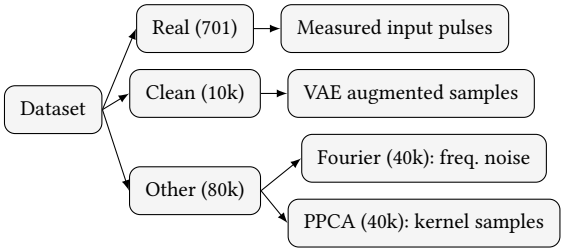

Figure 5: Dataset composition. *Real*: 701 measured pulses; *Clean*: 10k VAE-augmented in-distribution samples; *Other*: 80k OOD pulses from Fourier noise and PPCA sampling.

3.3 Surrogate with Physics-Informed Regularization

The surrogate maps laser pulse inputs to time-resolved implosion trajectories, replacing slow MULTI calls with fast forward evaluation. The input is a 160-step laser history $\mathbf{p}$ plus three static capsule parameters $\mathbf{t}$. We concatenate them into $\mathbf{x} = \text{Concat}(\mathbf{t}, \mathbf{p})$. The output contains five 160-step trajectories: fuel density, velocity, temperature, ablation-boundary index, and boundary radius.

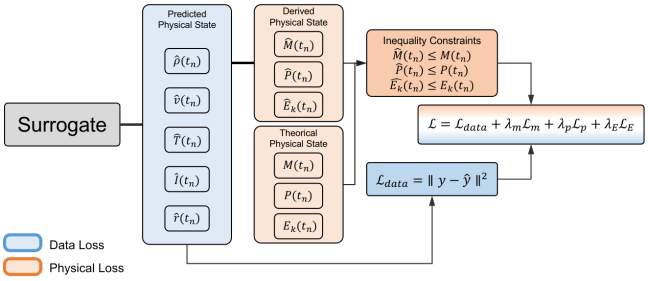

Figure 6: Physics-Informed Loss. The surrogate predicts trajectories from which boundary index, radius, and velocity are extracted. Steady-ablation budgets then yield inequality constraints on mass, momentum, and energy via squared hinge penalties.

A regression-trained surrogate can fit the data yet still violate conservation laws, especially in OOD regions sought by the optimizer. We therefore add a differentiable *Physics-Informed Loss* with inequality constraints derived from the steady ablation model. Concretely, a differentiable mass-from-index operator converts the predicted boundary index into an enclosed-mass estimate. Steady-ablation budgets then give conservative upper bounds on remaining mass, momentum, and energy. Squared hinge penalties enforce these bounds. Valid predictions are left unchanged. We use upper bounds rather than equality constraints because the budgets are approximate: they depend on simplified models and proxy quantities. The one-sided hinge preserves inequality semantics. Full derivations are in Appendix B.

3.4 PPO-based Optimization

Our goal in laser pulse optimization is to design an input pulse that maximizes ignition performance while maintaining feasibility. We train the policy using Proximal Policy Optimization (PPO).

Target scope. While the VAE decoder and surrogate were pretrained on data with varying target/capsule parameters, the RL optimization uses a single, fixed target configuration. This matches practical ICF workflows (e.g., Double-Cone Ignition), where target and laser pulses are optimized alternately; fixing the target isolates the pulse shaping problem.

Reward design. We define the reward from the Lawson criterion $\rho RT \geq 1.5\ \mathrm{g \cdot keV \cdot cm^{-2}}$ [2]. We use the peak ρRT product over the predicted trajectory as the terminal reward, plus a penalty on laser energy exceeding the experimental budget:

$$r(\mathbf{p}) = \max_{t \in [0, t_{\max}]} \hat{\rho}\hat{R}(t)\hat{T}(t) - \alpha \cdot \max(0, E_{\text{pulse}} - E_{\text{threshold}}). \quad (1)$$

The training reward is a ρRT proxy rather than fusion yield itself, but the two are positively correlated on the optimized pulses: higher ρRT pushes toward ignition, where self-sustaining alpha-particle heating triggers a nonlinear surge in yield. The "normalized yield" reported in Table 3 is MULTI-verified yield Y divided by the experimental baseline Y_{base}, expressed as a percentage.

Bandit formulation and PPO objective. Pulse design is cast as a one-step episodic decision process: at each episode the policy π_θ samples a latent action $\mathbf{z} \in \mathbb{R}^{d_z}$, decodes it into a pulse $\mathbf{p} = \mathbf{D}_\psi(\mathbf{z})$ via the fixed VAE decoder, and receives reward $r(\mathbf{p})$. The objective is $\eta(\theta) = \mathbb{E}_{\mathbf{z}\sim\pi_\theta}[r(\mathbf{D}_\psi(\mathbf{z}))]$. We optimize π_θ with PPO's clipped surrogate objective:

$$\mathcal{L}^{\text{clip}}(\theta) = \mathbb{E}_{\mathbf{z}\sim\pi_{\text{old}}}\Big[\min\Big(\rho_\theta(\mathbf{z})\,\hat{A}(\mathbf{z}), \\ \text{clip}(\rho_\theta(\mathbf{z}), 1-\epsilon, 1+\epsilon)\,\hat{A}(\mathbf{z})\Big)\Big], \quad (2)$$

where $\rho_\theta(\mathbf{z}) = \pi_\theta(\mathbf{z})/\pi_{\text{old}}(\mathbf{z})$ and the bandit-style advantage is $\hat{A}(\mathbf{z}) = r(\mathbf{D}_\psi(\mathbf{z})) - V_\phi(s)$ with a learned value baseline V_ϕ. PPO is preferable to simpler bandit methods because (1) it natively handles high-dimensional continuous action spaces, (2) the clipped objective bounds policy updates and prevents drift into adversarial OOD regions, and (3) the actor-critic architecture reduces variance. Detailed derivations and discussion are provided in Appendix C.

4 Experiments

We evaluate **Co4ICF** from two perspectives: (i) surrogate accuracy and physical consistency under distribution shift, and (ii) end-to-end laser pulse optimization under a fixed MULTI budget.

4.1 Experimental Setups

Dataset and splits. We use the MULTI dataset from Section 3.2. To avoid leakage, we split by shot ID before any augmentation: the VAE and PPCA augmentation models are fit *only* on the *Real* training shots, and test-set shots are never used in any augmentation step. Surrogate results are reported on an ID test split (Real shots held out before augmentation) and an OOD split (Fourier-perturbed shots, generated independently of training data).

Surrogate pretraining and evaluation. The designer's VAE and the surrogate are trained separately. The VAE learns a latent representation using *PPCA* subset pulses, while the surrogate is pretrained on the union of *Clean* and *PPCA* data. We evaluate on an **ID** split (*Real*) and an **OOD** split (*Fourier*), measuring peak MAE and Pearson correlation.

Optimization protocol. For pulse optimization, all methods use the same fixed target configuration, 1D surrogate backend, and MULTI labeling budget. Co4ICF performs high-frequency PPO updates on the surrogate and periodically refines the surrogate with 1D-MULTI on-policy labels. Final pulse designs from all optimizers are evaluated by the same direct 2D-MULTI protocol.

4.2 Surrogate

Table 1 summarizes surrogate performance on ID and OOD splits.

Physics constraints improve OOD. Adding physics-informed inequality constraints consistently improves generalization on the harder *Fourier* split across NAR backbones. For the Transformer encoder, OOD peak-x MAE drops from 8.020 to 6.302. Peak-y MAE drops from 0.310 to 0.252, and Pearson correlation rises from 0.855 to 0.912. There is a mild trade-off on the ID *Real* split, where conservative bounds slightly reduce in-distribution accuracy; but OOD robustness improves considerably. The trend holds across architectures. Physical regularization suppresses non-physical extrapolations and reduces error amplification under optimizer-induced OOD inputs. Fig. 7 further tracks the L_2 loss ratio during surrogate pretraining. The curves show that physics regularization stabilizes training and mitigates overfitting, yielding representations that transfer better to unseen pulses.

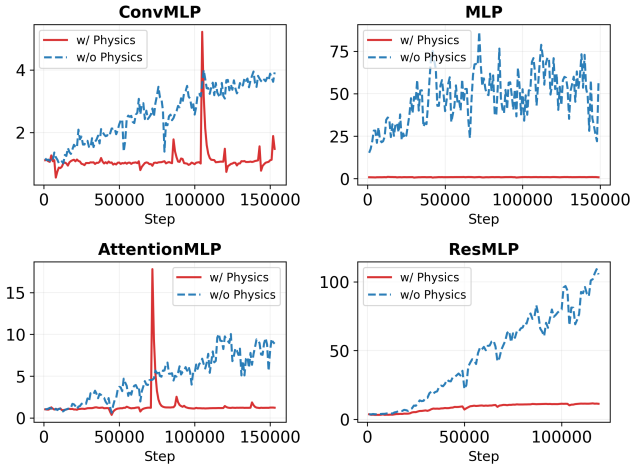

Figure 7: L_2 loss ratio during surrogate pretraining.

Transformer encoder is the strongest NAR surrogate. It offers the strongest balance between accuracy and robustness among NAR models. With physics constraints, it has the lowest OOD peak-x MAE (6.302), the lowest OOD peak-y MAE (0.252), and high OOD Pearson correlation (0.912). Given the short horizon (T=160) and

Table 1: Surrogate architecture comparison under ID and OOD evaluation.

Surrogate	ID (*Real*)						OOD (*Fourier*)					
	Pk x↓		Pk y↓		Pearson↑		Pk x↓		Pk y↓		Pearson↑	
	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o
Non-autoregressive (NAR)												
Vanilla MLP	2.628	2.316	0.078	0.089	0.971	0.998	9.094	10.528	0.273	0.386	0.925	0.899
ResMLP	7.431	7.922	0.383	0.469	0.930	0.958	25.798	32.128	1.118	1.131	0.636	0.558
1DConv	4.614	2.276	0.181	0.095	0.961	0.996	9.434	9.126	0.431	0.482	0.901	0.869
Transformer Enc.	2.450	2.194	0.083	0.111	0.969	0.997	6.302	8.020	0.252	0.310	0.912	0.855
Autoregressive (AR)												
Transformer Dec.	N/A	8.318	N/A	0.502	N/A	0.762	N/A	36.875	N/A	1.966	N/A	0.451

w/ and **w/o**: with / without physics-informed constraints (Section 3.3 & Appendix B). Notably, we do not apply physical constraints to the AR model, since the constraint formulation is not directly comparable between AR and NAR predictors.
Cell color: applied to **w/** only; green = better than w/o, red = worse than w/o (following ↓ / ↑).
Pk x/y: peak (x, y) MAE of the fusion-grain peak location.
Underlined values: the best in the column.

the need for high throughput, we use it as the default surrogate backend.

NAR vs. AR comparison. We compare with an AR decoder. The AR Transformer underperforms on both splits. Although autoregressive decoding respects temporal causality, it is more sensitive to compounding rollout errors. Its lower throughput also limits practical training and tuning (Table 2). Our explanation is that the simulator context has only 160 time steps, which is short relative to typical long-horizon sequence tasks. With limited context, step-wise temporal modeling helps less, while sequential rollouts still accumulate errors and increase latency. This agrees with prior forecasting work [25]. In such settings, simple linear or MLP-style baselines can match Transformer variants on many benchmarks.

Surrogate inference speedup. Table 2 reports timing for 1,000 forward evaluations. MULTI takes 319.83 s on 16 CPU threads. On CPU, the Transformer encoder takes 14.76 s (21.67× speedup). On GPU, it takes 4.58 s (69.83× speedup). The Transformer decoder is slower than MULTI on CPU (561.67 s) and only marginally faster on GPU (216.32 s). The NAR encoder is therefore the most efficient evaluation engine for iterative pulse search.

Table 2: Wall-clock time for 1k forward evaluations.

Backend	Time (s)↓	Speedup
Transformer Encoder (CPU)	14.76	21.67
Transformer Encoder (GPU)	**4.58**	**69.83**
Transformer Decoder (CPU)	561.67	0.57
Transformer Decoder (GPU)	216.32	1.48
MULTI (CPU)	319.83	1.00

Measurements use Ubuntu 24.04 LTS, an AMD Ryzen 9 7940H CPU (16 threads), and an NVIDIA GeForce RTX 4060 Max-Q GPU (8 GB). MULTI uses 16 CPU threads.
Speedup is computed as Time(MULTI CPU)/Time(Backend).

4.3 Co-evolving Optimization

4.3.1 Main Performance. We first test whether surrogate refinement improves the 1D pulse-optimization trajectory. In Fig. 1a, periodic refinement reaches 146.1% after 1D-MULTI re-evaluation. The frozen-surrogate variant plateaus at 115.4%. This gap indicates that updating the surrogate on on-policy samples helps after the policy moves away from the offline training distribution. We use this 1D trajectory as the main diagnostic in this subsection; final designs are evaluated by direct 2D-MULTI in Section 4.3.2.

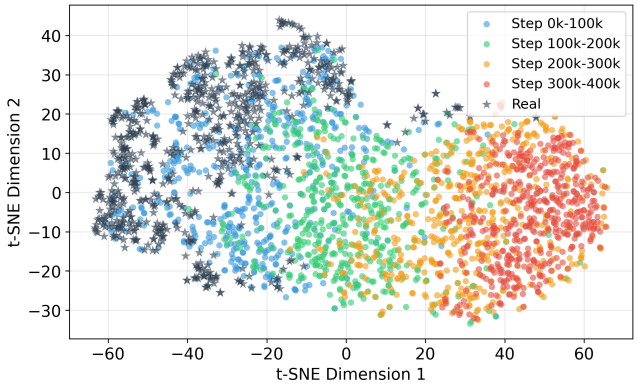

Figure 8: t-SNE visualization of laser pulses during Co4ICF optimization. Colored dots: policy-sampled pulses at different stages (0–100k, 100k–200k, 200k–300k, 300k–400k); gray stars: experimental pulses. The policy progressively moves beyond the experimental pulse cluster in the embedding space.

The optimized waveforms provide a qualitative check that the improvement is not driven by isolated reward spikes. Fig. 1b compares randomly initialized pulse samples with final policy samples. After RL with surrogate refinement, the designer produces smoother ramp–plateau–shutoff structures among the optimized

samples, which is consistent with the higher 1D-MULTI reward curve in Fig. 1a.

The policy also moves beyond the measured-pulse cluster within the decoder-defined search space. Fig. 8 embeds policy-sampled 160-step pulses with PCA followed by t-SNE. Early samples remain close to the Real cluster, whereas later samples expand into broader embedding regions. This progression suggests that the actor moves beyond imitation while staying constrained by the learned decoder.

Policy training speedup. Runtime drops sharply in the PPO loop. Across 3,967 PPO iterations, MULTI-based training takes 19,800 s. The surrogate loop takes 20 s, giving a **990**× inner-loop speedup. This timing covers PPO policy updates only. It excludes offline data generation, surrogate pretraining, low-frequency relabeling, and 2D-MULTI verification. Overall speedup depends on cycle count and labeling budget M. The main saving is replacing per-iteration MULTI calls with surrogate queries.

4.3.2 Direct 2D Evaluation and Ablation. We next test whether 1D-optimized pulses remain strong under direct 2D-MULTI evaluation. All optimizers use the same 1D surrogate and MULTI labeling budget during search. 2D-MULTI is used only for post-hoc evaluation of final designs. Thus, no 2D samples are used to train, fine-tune, or select the PPO policy.

Table 3: Direct 2D-MULTI evaluation of final pulse designs under the same 1D surrogate and MULTI labeling budget. All reported physical quantities and normalized yields are computed by 2D-MULTI, which is used only for post-hoc evaluation. Normalized Yield is the 2D-MULTI fusion yield divided by the baseline pulse's 2D-MULTI yield.

Case	$\bar{\rho}$	ρR	T	Norm. Yield↑
Baseline	7.60	0.907	163.63	100.0%
BO	7.20	1.100	233.98	173.5%
GA	7.30	0.829	218.10	121.9%
W/O Update (Static)	6.52	0.482	388.72	126.3%
W/O Update (Enlarged)	6.49	0.512	422.09	174.8%
W/ Update (Co4ICF)	**7.10**	**1.063**	344.60	**246.9%**

Table 3 shows that Co4ICF achieves the highest normalized 2D yield: 246.9%, compared with 173.5% for BO, 121.9% for GA, and 126.3% for frozen-surrogate PPO. The table also gives a budget-matched ablation for the co-evolving surrogate. W/O Update (Enlarged) retrains the frozen surrogate with the same number of extra samples as Co4ICF, but draws them from the original data distribution instead of the current policy. This enlarged frozen surrogate improves over W/O Update (Static), from 126.3% to 174.8%, but remains well below W/ Update (Co4ICF). Together, these comparisons suggest that the gain comes from coupling optimization with on-policy surrogate refinement, rather than from optimizer choice or extra data alone.

4.3.3 Distribution-Shift Diagnostics. The budget-matched ablation shows the final effect of on-policy refinement. To diagnose this mechanism, we track the peak error between surrogate predictions and MULTI-evaluated trajectories during PPO updates. This diagnostic measures whether the surrogate remains reliable as the policy shifts the input distribution away from the offline training set.

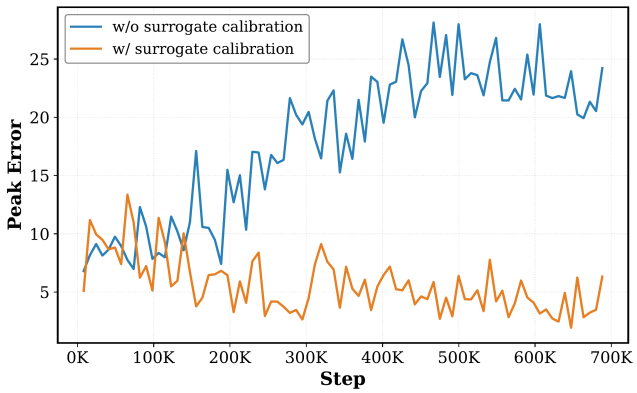

Figure 9: Surrogate–MULTI peak error over PPO training steps. Without calibration, the error grows unbounded as the policy drives inputs OOD; with periodic calibration, the error stays bounded.

Fig. 9 shows that the frozen surrogate becomes increasingly unreliable during optimization. Its peak error grows and eventually spikes, indicating optimizer-induced distribution shift. Periodic calibration keeps the surrogate–MULTI gap bounded, explaining why W/ Update outperforms both W/O Update variants in Table 3.

5 Conclusions

We presented **Co4ICF**, a co-evolving framework that couples a physics-informed surrogate with an RL-based pulse designer for ICF design under limited high-fidelity budgets. The dual-frequency loop combines high-frequency PPO updates with low-frequency surrogate fine-tuning to correct local extrapolation errors and reduce reward hacking. Co4ICF achieves 246.9% normalized yield in direct 2D-MULTI evaluation relative to the experimental baseline. It does so without 2D training or fine-tuning, while retaining a 990× inner-loop speedup. Budget-matched ablations indicate that these gains are not explained solely by additional simulation data.

More broadly, Co4ICF suggests a recipe for simulation-informed design. Treat optimizer-induced shift as a training signal, and keep the learned simulator aligned with decision-relevant regions.

6 Limitations and Broader Impact

Limitations. The framework relies on MULTI-IFE (1D & 2D) for high-fidelity feedback [16], so it cannot capture asymmetric instabilities (e.g., Rayleigh–Taylor) that matter in real 3D implosions. Search is constrained to the VAE manifold, which keeps pulses feasible but risks missing physically valid shapes outside it.

Broader impact. The co-evolving paradigm is not specific to ICF and may transfer to other domains where iterative optimization interacts with a learned model (e.g., molecular design, materials

discovery). However, ICF research inherently bridges civilian fusion energy and weapons physics; methodologies for optimizing implosion performance could, in principle, be applied to weapons-relevant contexts. Researchers adapting this paradigm should adhere to relevant export control and non-proliferation frameworks.

Acknowledgments

This work was supported by the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (2025-GZL-RGZN-BTBX-02024), and National Natural Science Foundation of China Program (125B100032, 92370201).

References

[1] Rushil Anirudh, Jayaraman J Thiagarajan, Peer-Timo Bremer, and Brian K Spears. 2020. Improved surrogates in inertial confinement fusion with manifold and cycle consistencies. *Proceedings of the National Academy of Sciences* 117, 18 (2020), 9741–9746.

[2] Stefano Atzeni and Jürgen Meyer ter Vehn. 2004. *The Physics of Inertial Fusion: Beam–Plasma Interaction, Hydrodynamics, Hot Dense Matter.* Oxford University Press, Oxford, UK. doi:10.1093/acprof:oso/9780198562641.001.0001

[3] Francesco Capuano, Davorin Peceli, and Gabriele Tiboni. 2025. Shaping Laser Pulses with Reinforcement Learning. *arXiv preprint arXiv:2503.00499* (2025).

[4] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* 602, 7897 (2022), 414–419.

[5] Rahman Ejaz, Varchas Gopalaswamy, A Lees, C Kanan, D Cao, and R Betti. 2024. Deep learning-based predictive models for laser direct drive at the Omega Laser Facility. *Physics of Plasmas* 31, 5 (2024).

[6] 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 (2025), 013009.

[7] Kelli Denise Humbird. 2019. *Machine learning guided discovery and design for inertial confinement fusion.* Ph. D. Dissertation. Texas A&M University.

[8] 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 (2021).

[9] Omar A Hurricane. 2024. How ignition and target gain> 1 were achieved in inertial fusion. *High Energy Density Physics* 53 (2024), 101157.

[10] 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 (2021), 422–440.

[11] Minsu Kim, Jiayao Gu, Ye Yuan, Taeyoung Yun, Zixuan Liu, Yoshua Bengio, and Can Chen. 2025. Offline Model-Based Optimization: Comprehensive Review. arXiv:2503.17286 [cs.LG] https://arxiv.org/abs/2503.17286

[12] Z Li, ZQ Zhao, XH Yang, GB Zhang, YY Ma, H Xu, FY Wu, FQ Shao, and J Zhang. 2023. Hybrid optimization of laser-driven fusion targets and laser profiles. *Plasma Physics and Controlled Fusion* 66, 1 (2023), 015010.

[13] Shulei Liu, Handing Wang, Wei Peng, and Wen Yao. 2024. Surrogate-assisted evolutionary algorithms for expensive combinatorial optimization: a survey. *Complex & Intelligent Systems* 10, 4 (Aug. 2024), 5933–5949. doi:10.1007/s40747-024-01465-5

[14] Matthew L Olson, Shusen Liu, Jayaraman J Thiagarajan, Bogdan Kustowski, Weng-Keen Wong, and Rushil Anirudh. 2024. Transformer-powered surrogates close the ICF simulation-experiment gap with extremely limited data. *Machine Learning: Science and Technology* 5, 2 (2024), 025054.

[15] 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 (2019), 686–707.

[16] Rafael Ramis and Jürgen Meyer-ter Vehn. 2016. MULTI-IFE—A one-dimensional computer code for Inertial Fusion Energy (IFE) target simulations. *Computer Physics Communications* 203 (2016), 226–237.

[17] Brian K Spears, Scott Brandon, Dan T Casey, John E Field, Jim A Gaffney, Kelli D Humbird, Andrea L Kritcher, Michael KG Kruse, Eugene Kur, Bogdan Kustowski, et al. 2025. Predicting fusion ignition at the National Ignition Facility with physics-informed deep learning. *Science* 389, 6761 (2025), 727–731.

[18] Richard S Sutton. 1991. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin* 2, 4 (1991), 160–163.

[19] Tao Tao, Guannan Zheng, Qing Jia, Rui Yan, and Jian Zheng. 2023. Laser pulse shape designer for direct-drive inertial confinement fusion implosions. *High Power Laser Science and Engineering* 11 (2023), e41.

[20] Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. 2022. Design-Bench: Benchmarks for Data-Driven Offline Model-Based Optimization. arXiv:2202.08450 [cs.LG] https://arxiv.org/abs/2202.08450

[21] Jingyi Wang, N Chiang, Andrew Gillette, and J Luc Peterson. 2024. A multi-fidelity Bayesian optimization method for inertial confinement fusion design. *Physics of Plasmas* 31, 3 (2024).

[22] Zixu Wang, Yuhan Wang, Junfei Ma, Fuyuan Wu, Junchi Yan, Xiaohui Yuan, Zhe Zhang, and Jie Zhang. 2025. Predictive Hydrodynamic Simulations for Laser Direct-drive Implosion Experiments via Artificial Intelligence. *arXiv preprint arXiv:2507.16227* (2025).

[23] S Wei, F Wu, Y Zhu, J Yang, L Zeng, X Li, and J Zhang. 2024. A Machine Learning Method for the Optimization Design of Laser Pulse in Fast Ignition Simulations. *Journal of Fusion Energy* 43, 1 (2024), 6.

[24] Fuyuan Wu, Xiaohu Yang, Yanyun Ma, Qi Zhang, Zhe Zhang, Xiaohui Yuan, Hao Liu, Zhengdong Liu, Jiayong Zhong, Jian Zheng, et al. 2022. Machine-learning guided optimization of laser pulses for direct-drive implosions. *High Power Laser Science and Engineering* 10 (2022), e12.

[25] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2022. Are Transformers Effective for Time Series Forecasting? arXiv:2205.13504 [cs.AI] https://arxiv.org/abs/2205.13504

A Dataset and Data Augmentation

This section describes the dataset composition and the three data augmentation strategies (VAE, PPCA, Fourier-domain noise) used to expand training coverage beyond the limited set of experimentally measured pulses.

A.1 Dataset Overview

The dataset contains about 90k normalized pairs, stored in HDF5 format by shot ID. Each quantity is interpolated onto a uniform grid of 160 time steps (0.05 ns resolution) and stored as a fixed-length vector. Each input consists of a laser power pulse and three discrete target parameters.

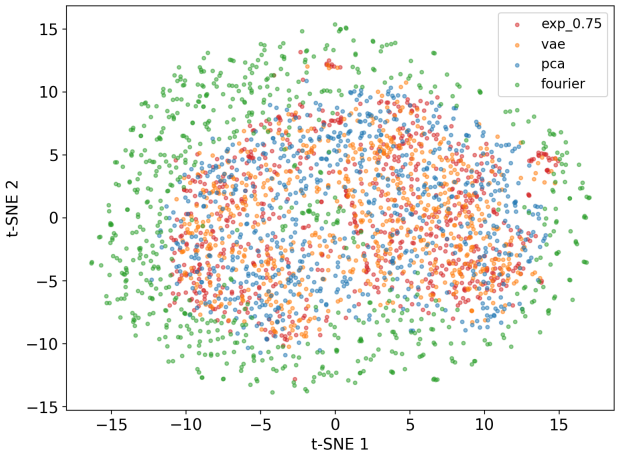

Figure 10: t-SNE visualization of input laser pulses across dataset subsets. Coverage increases from Real (701 pulses) to VAE-augmented, PPCA, and Fourier-noise samples.

A.2 VAE Sampling

We employ a variational autoencoder (VAE) to learn a compact generative representation of measured laser pulses for offline data augmentation (`VAE_aug`). `VAE_aug` is trained *only* on the *Real* subset, producing samples that follow the same empirical distribution

as measured pulses. This dataset VAE is *not* the VAE used in the downstream optimization stage.

A.2.1 Model and Training Objective. Let $p \in \mathbb{R}^T$ denote a normalized laser power pulse interpolated onto a uniform grid with $T = 160$ time steps. The VAE consists of an encoder $q_\phi(z|p)$ and a decoder $p_\theta(p|z)$, where $z \in \mathbb{R}^d$ is a latent code. We assume a standard Gaussian prior $p(z) = \mathcal{N}(0, I)$ and train the VAE by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{\text{VAE}}(p; \phi, \theta) = \mathbb{E}_{z\sim q_\phi(z|p)}\left[\|p - \hat{p}\|_2^2\right] + \beta\,\text{KL}\left(q_\phi(z \mid p)\|, \mathcal{N}(0, I)\right), \tag{3}$$

$$\hat{p} = D_\theta(z), \tag{4}$$

where $D_\theta(\cdot)$ denotes the decoder, and β is a weighting factor controlling the strength of the KL regularization.

A.2.2 Sampling Procedure. After pretraining, we draw latent vectors from the prior and decode them:

$$z \sim \mathcal{N}(0, I), \qquad \tilde{p} = \Pi_{\geq 0}(D_\theta(z)), \tag{5}$$

where $\Pi_{\geq 0}(\cdot)$ denotes a non-negativity projection (implemented as $\max(\cdot, 0)$) to ensure physically valid laser power.

A.3 PPCA Sampling

To provide a linear generative baseline complementary to the nonlinear VAE, we utilize Probabilistic Principal Component Analysis (PPCA). Similar to the VAE strategy, the PPCA model is fitted exclusively on the *Real* subset.

A.3.1 Model and Training Objective. We model the observed pulse $p \in \mathbb{R}^T$ as a linear transformation of a lower-dimensional latent variable $z \in \mathbb{R}^{d_{pca}}$ with additive Gaussian noise:

$$p = Wz + \mu + \epsilon, \tag{6}$$

where $\mu \in \mathbb{R}^T$ is the data mean, $W \in \mathbb{R}^{T\times d_{pca}}$ is the factor loading matrix, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ represents isotropic noise.

A.3.2 Sampling Procedure.

$$z \sim \mathcal{N}(0, I), \qquad \tilde{p} = \Pi_{\geq 0}\left(Wz + \mu + \epsilon_{\text{sample}}\right), \tag{7}$$

where ϵ_{sample} is drawn from $\mathcal{N}(0, \sigma^2 I)$, and $\Pi_{\geq 0}(\cdot)$ enforces the physical non-negativity constraint.

A.4 Fourier-Domain Exponential-Decay Noise Augmentation

We perturb measured pulses in the Fourier domain to generate OOD yet physically plausible laser pulses. For a real-valued input pulse $p(t) \in \mathbb{R}^T$, we compute its one-sided real FFT

$$P_k = \text{RFFT}(p)_k \in \mathbb{C} \tag{8}$$

then inject additive complex Gaussian noise whose magnitude decays exponentially with frequency index:

$$\tilde{P}_k = P_k + A \exp\left(-\frac{k}{\tau}\right)\epsilon_k, \quad \epsilon_k \sim \mathcal{N}(\mu, 1) + i\mathcal{N}(\mu, 1) \tag{9}$$

where k is the discrete frequency index, τ is the decay factor, and A sets the overall perturbation strength. The exponential envelope biases perturbations toward low-frequency components. This largely preserves coarse pulse structure while producing diverse pulse variations. As noted in the main text, the exponential cutoff suppresses high-frequency noise and respects bandwidth and slew-rate limits of real ICF lasers.

Finally, we transform back via

$$\tilde{p}(t) = \max(\text{IRFFT}(\tilde{P}), 0) \tag{10}$$

to enforce continuity and non-negativity.

B Physics-Informed Inequality Loss: Full Derivations

This section provides complete derivations of the physics-informed inequality constraints used to regularize the surrogate during pretraining (Section 3.3).

B.1 ICF Implosion in Three Stages

An ICF implosion proceeds through three stages with different dominant mechanisms. These stages motivate the inequality constraints below.

Stage 1: Laser ablation and rocket effect. An intense laser pulse irradiates the fuel capsule surface. The outer material heats and expands outward as a plasma corona; by momentum conservation, this ablation drives an inward *rocket reaction* that accelerates the remaining shell. The mass ablation rate $\dot{m}_\alpha$ is set by laser intensity I_L and wavelength λ_L through Eq. (14). Because ablation removes mass from the shell, enclosed mass decreases monotonically during this stage. The ablation pressure also imparts radial momentum, analogous to rocket thrust from mass ejection.

Stage 2: Implosion and compression. After the main drive ends, the inward-moving shell coasts and converges toward the center, compressing the fuel to extreme densities ($\bar{\rho} \sim 10^2$–10^3 g/cm^3). During this inertial coasting phase, no further momentum is injected; the shell kinetic energy is converted into internal energy as the fuel stagnates. The implosion velocity v_c and remaining shell mass m together determine the kinetic energy budget available for compression. Conservation laws demand that the enclosed momentum and kinetic energy cannot exceed the total impulse and energy delivered during Stage 1—these are the *momentum* and *energy confinement* constraints derived below.

Stage 3: Stagnation and thermonuclear burn. When the imploding shell decelerates against the compressed fuel core, kinetic energy is thermalized and the fuel reaches peak temperature and density. If the Lawson criterion $\rho RT \geq 1.5$ g · keV · cm^{-2} is met, a thermonuclear burn wave propagates before hydrodynamic disassembly quenches it. Fusion yield depends on how efficiently implosion kinetic energy is converted into thermal energy at stagnation, which is set by the optimized pulse shape.

This three-stage picture explains why conservation constraints matter. A surrogate that predicts too much enclosed mass or momentum is physically impossible. The inequalities below encode mass, momentum, and energy bookkeeping constraints.

B.2 Setup and Notation

We consider a 1D Lagrangian capsule discretized into N mass layers ($N = 160$ in MULTI), indexed by $i \in \{1, \dots, N\}$, with fixed initial layer masses m_i. Let t_n denote the uniformly sampled time grid ($n = 1, \dots, T$, $T = 160$) and Δt the timestep. The total initial mass is

$$M_0 = \sum_{i=1}^{N} m_i. \tag{11}$$

After de-normalization, the surrogate predicts: (i) boundary shock index $\hat{I}(t_n)$, (ii) boundary radius $\hat{R}_I(t_n)$, (iii) mass-weighted average velocity $\hat{v}_c(t_n)$, together with other state histories.

B.3 Differentiable Mass Operator

Define the cumulative initial mass

$$M(k) = \sum_{i=1}^{k} m_i, \quad k \in \{0, 1, \dots, N\},\ M(0) = 0. \tag{12}$$

We define a differentiable *mass-from-index* operator $\mathcal{M}(\cdot)$ via linear interpolation:

$$\hat{m}(t_n) = \mathcal{M}(\hat{I}(t_n)) = (1 - \delta_n) M(\lfloor \hat{I}_n \rfloor) + \delta_n M(\lceil \hat{I}_n \rceil) \tag{13}$$

where $\delta_n = \hat{I}_n - \lfloor \hat{I}_n \rfloor$, and $\hat{I}_n = \mathrm{clip}(\hat{I}(t_n), 0, N)$.

B.4 Steady Ablation Model

Following the ablation rate formula [2]:

$$\dot{m}_\alpha(t) = 1.3{\times}10^6 \Big(\frac{A}{2Z}\Big)^{2/3} \Big(\frac{\lambda_L}{0.35\,\mu\mathrm{m}}\Big)^{-4/3} \Big(\frac{I_L(t)}{10^{15}\,\mathrm{W/cm^2}}\Big)^{1/3} \tag{14}$$

When only laser power $P(t)$ is available, we use the boundary radius to approximate intensity:

$$I_L(t) \approx \frac{P(t)}{4\pi \hat{R}_I^2(t)}. \tag{15}$$

The total ablated mass rate is then

$$\dot{M}_{\mathrm{abl}}(t) = 4\pi \hat{R}_I^2(t)\, \dot{m}_\alpha(t). \tag{16}$$

In discrete time, the cumulative ablated mass and the remaining-mass *upper bound* are

$$\Delta M_{\mathrm{abl}}(t_n) = \dot{M}_{\mathrm{abl}}(t_n)\Delta t,\ M_{\mathrm{lim}}(t_n) = \max\Big(0,\ M_0 - \sum_{k=1}^{n} \Delta M_{\mathrm{abl}}(t_k)\Big). \tag{17}$$

B.5 Physical Confinement Constraints

B.5.1 Momentum confinement. Define the predicted enclosed momentum as

$$\hat{p}(t_n) = \hat{m}(t_n)\, |\hat{v}_c(t_n)|. \tag{18}$$

We approximate the exhaust speed:

$$u_{\mathrm{ex}}(t) \approx 2c_s(t) \approx 2\sqrt{\frac{k_B T_I^*(t)}{m_{\mathrm{ion}}}}, \tag{19}$$

The impulse budget up to t_n is

$$p_{\mathrm{lim}}(t_n) = \sum_{k=1}^{n} \Delta M_{\mathrm{abl}}(t_k)\, u_{\mathrm{ex}}(t_k). \tag{20}$$

The momentum constraint is enforced by a squared hinge penalty:

$$\mathcal{L}_p = \frac{1}{T} \sum_{n=1}^{T} \Big[\max\big(0,\ \hat{p}(t_n) - p_{\mathrm{lim}}(t_n)\big) \Big]^2. \tag{21}$$

B.5.2 Energy confinement. Define the predicted kinetic energy of the enclosed mass:

$$\hat{E}_{\mathrm{kin}}(t_n) = \frac{1}{2} \hat{m}(t_n) \hat{v}_c^2(t_n). \tag{22}$$

Using the remaining-mass bound $M_{\mathrm{lim}}(t_n)$:

$$u_{\mathrm{imp}}(t_n) = \sum_{k=1}^{n} u_{\mathrm{ex}}(t_k)\, \frac{\Delta M_{\mathrm{abl}}(t_k)}{M_{\mathrm{lim}}(t_k) + \varepsilon}, \qquad \varepsilon > 0. \tag{23}$$

The corresponding energy upper bound is

$$E_{\mathrm{lim}}(t_n) = \frac{1}{2} M_{\mathrm{lim}}(t_n)\, u_{\mathrm{imp}}^2(t_n). \tag{24}$$

$$\mathcal{L}_E = \frac{1}{T} \sum_{n=1}^{T} \Big[\max\big(0,\ \hat{E}_{\mathrm{kin}}(t_n) - E_{\mathrm{lim}}(t_n)\big) \Big]^2. \tag{25}$$

B.5.3 Mass confinement.

$$\mathcal{L}_m = \frac{1}{T} \sum_{n=1}^{T} \Big[\max\big(0,\ \hat{m}(t_n) - M_{\mathrm{lim}}(t_n)\big) \Big]^2. \tag{26}$$

B.6 Overall Objective

$$\mathcal{L} = \mathcal{L}_{\mathrm{data}} + \lambda_m \mathcal{L}_m + \lambda_p \mathcal{L}_p + \lambda_E \mathcal{L}_E. \tag{27}$$

In practice, these constraints matter most during ablation and compression. When the boundary index collapses to $\hat{I}(t) \approx 0$ after ignition, $\hat{m}(t)$ becomes small and the penalties naturally deactivate. Because the constraints operate on full trajectories, they apply only to the non-autoregressive surrogate. Adding them to an autoregressive rollout would conflate physics violations with rollout error accumulation.

C PPO Objective Derivation for the One-Step Setting

We derive the PPO-style surrogate objective used in Section 3.4 for the one-step episodic setting. The policy π_θ samples a latent action $\mathbf{z} \in \mathbb{R}^{d_z}$, which is decoded into a pulse $\mathbf{p} = \mathbf{D}_\psi(\mathbf{z})$ by a fixed decoder $\mathbf{D}_\psi$. The scalar reward is $r(\mathbf{p})$, hence the expected return is

$$\eta(\theta) \triangleq \mathbb{E}_{\mathbf{z} \sim \pi_\theta}\Big[r\big(\mathbf{D}_\psi(\mathbf{z})\big) \Big]. \tag{28}$$

C.1 Bandit Advantage and Policy Difference Identity

Given a reference (behavior) policy π_{old}, we define the one-step advantage as

$$A_{\text{old}}(\mathbf{z}) \triangleq r\Big(\mathbf{D}_\psi(\mathbf{z})\Big) - \mathbb{E}_{\mathbf{z}'\sim\pi_{\text{old}}}\Big[r\Big(\mathbf{D}_\psi(\mathbf{z}')\Big)\Big]\,. \tag{29}$$

Lemma C.1 (Policy difference identity in the one-step setting). *For any policy π_θ and reference policy π_{old},*

$$\eta(\theta) = \eta(\pi_{old}) + \mathbb{E}_{\mathbf{z}\sim\pi_\theta}[A_{old}(\mathbf{z})]\,. \tag{30}$$

Proof. By definition,

$$\begin{aligned}\eta(\theta) - \eta(\pi_{\text{old}}) &= \mathbb{E}_{\mathbf{z}\sim\pi_\theta}\Big[r\Big(\mathbf{D}_\psi(\mathbf{z})\Big)\Big] - \mathbb{E}_{\mathbf{z}'\sim\pi_{\text{old}}}\Big[r\Big(\mathbf{D}_\psi(\mathbf{z}')\Big)\Big]\\ &= \mathbb{E}_{\mathbf{z}\sim\pi_\theta}\Big[r\Big(\mathbf{D}_\psi(\mathbf{z})\Big) - \mathbb{E}_{\mathbf{z}'\sim\pi_{\text{old}}}\Big[r\Big(\mathbf{D}_\psi(\mathbf{z}')\Big)\Big]\Big]\\ &= \mathbb{E}_{\mathbf{z}\sim\pi_\theta}[A_{\text{old}}(\mathbf{z})]\,,\end{aligned} \tag{31}$$

where the second line uses that the π_{old}-expectation is a constant independent of $\mathbf{z}$. □

C.2 Importance-Sampling Surrogate Objective

Using importance sampling,

$$\mathbb{E}_{\mathbf{z}\sim\pi_\theta}[f(\mathbf{z})] = \mathbb{E}_{\mathbf{z}\sim\pi_{\text{old}}}\left[\frac{\pi_\theta(\mathbf{z})}{\pi_{\text{old}}(\mathbf{z})} f(\mathbf{z})\right]. \tag{32}$$

Applying (32) to (30) yields

$$\eta(\theta) = \eta(\pi_{\text{old}}) + \mathbb{E}_{\mathbf{z}\sim\pi_{\text{old}}}[\rho_\theta(\mathbf{z})\, A_{\text{old}}(\mathbf{z})]\,, \qquad \rho_\theta(\mathbf{z}) \triangleq \frac{\pi_\theta(\mathbf{z})}{\pi_{\text{old}}(\mathbf{z})}. \tag{33}$$

Eq. (33) motivates the PPO-style surrogate objective used in the main text.