

ARTIFICIAL INTELLIGENCE FOR THE ELECTRON ION COLLIDER
MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MA, U.S.A.
27–29 OCTOBER 2025

ScatterPrism: convergence for generative simulation and inverse problems in particle and nuclear physics

Zeyu Xia^{a,*}, Tyler Kim^a, Trevor Reed^b, Judy Fox^b, Geoffrey Fox^a
and Adam Szczepaniak^c

^a*Department of Computer Science, University of Virginia,
85 Engineer's Way, Charlottesville, VA 22904, U.S.A.*

^b*School of Data Science, University of Virginia,
1919 Ivy Road, Charlottesville, VA 22903, U.S.A.*

^c*Department of Physics, Indiana University Bloomington,
727 East Third Street, Bloomington, IN 47405, U.S.A.*

E-mail: zeyu.xia@virginia.edu

ABSTRACT: High-fidelity simulations and complex inverse problems, such as detector modeling and unfolding, are computationally intensive bottlenecks across subatomic physics, yet essential for accurate physical interpretation. While Conditional Flow Matching (CFM) offers a robust acceleration approach, we demonstrate its standard training loss is fundamentally misleading. Specifically, utilizing a Jefferson Lab Nuclear Physics (NP) kinematic dataset ($\gamma p \rightarrow \rho^0 p \rightarrow \pi^+ \pi^- p$), we expose that CFM loss plateaus prematurely, obscuring ongoing physical refinement. To verify this disconnect is a dataset-agnostic pathology, we introduce ScatterPrism, an efficient generative surrogate evaluated against both the NP data and synthetic stress tests modeling challenging 1D distribution topologies. Coupling these benchmarks, we establish that physics-informed metrics continue improving long after standard loss converges. Consequently, we propose a multi-metric diagnostic protocol to ensure true kinematic fidelity without data memorization. Driven by NP challenges relevant to the forthcoming Electron-Ion Collider (EIC), this unified machinery has strong potential to extend to High-Energy Physics (HEP) applications, such as jet modeling. Furthermore, the framework holds promise for broader domains requiring rigorous generative reliability, including medical imaging, astrophysics, and quantitative finance.

KEYWORDS: Analysis and statistical methods; Data processing methods; Simulation methods and programs; Software architectures (event data models, frameworks and databases)

ARXIV EPRINT: [2604.01313](https://arxiv.org/abs/2604.01313)

*Corresponding author.



Contents

1	Introduction	1
2	Related work	2
3	Methodology	3
3.1	Datasets and feature representation	3
3.2	Conditional flow matching framework	3
3.3	Physics-informed metrics	5
4	Results	5
4.1	Synthetic benchmark validation	5
4.2	Performance on MC-POM dataset	5
5	Discussion and conclusion	7
A	Network architecture and training details	9
B	Evaluation metrics	10
C	Synthetic benchmark analysis	12
D	Extended unconditional generation validation	13
E	Extended detector unfolding validation	15
F	Computational performance	16

1 Introduction

Modern experimental physics analyses rely on large-scale Monte Carlo simulation datasets to attain the statistical precision required for high-fidelity measurements. While parton-level event generation (e.g., Pythia [1], MadGraph [2]) is comparatively cheap per event, the subsequent full detector-response simulation (typically GEANT4 [3]) dominates compute and scales unfavorably with event volume and detector complexity. This challenge is rooted in contemporary Nuclear Physics (NP) programs, such as Jefferson Lab experiments (e.g., CLAS12 [4], GlueX [5]) and the forthcoming Electron-Ion Collider (EIC), where modeling particle transport and detector responses dominates computational budgets. These barriers are shared by High-Energy Physics (HEP) initiatives like the Large Hadron Collider (LHC), establishing a universal need for faster AI-based surrogate models across both communities.

Detector unfolding — recovering true event-level observables from detector-level measurements degraded by detector effects, such as finite instrumental resolution — further magnifies these computational demands, motivating the shift toward generative deep learning. Deep generative models, particularly Conditional Flow Matching (CFM), offer stable, simulation-free training. However,

using the JLab NP photoproduction dataset, we demonstrate that standard CFM training loss is an unreliable indicator of true physical convergence.

This limitation, characterized as spectral bias [6], obscures ongoing physical refinement. To resolve this dataset-agnostic pathology, we introduce ScatterPrism, a CFM framework tailored for high-fidelity kinematic event generation and detector unfolding. Our primary contributions are:

1. *Convergence diagnostics and validation suite.* We identify the premature plateau of standard CFM loss, which obscures ongoing physical refinement, and establish a rigorous multi-metric protocol to accurately track true convergence, verify generative fidelity, and prevent data memorization; the constituent metrics are introduced and motivated in section 3.3.
2. *ScatterPrism Framework.* We introduce ScatterPrism, a configurable CFM-based tool, and validate its capabilities for event generation and conditional detector unfolding on a realistic Jefferson Lab dataset ($\gamma p \rightarrow \rho^0 p \rightarrow \pi^+ \pi^- p$) relevant to the forthcoming EIC.
3. *Synthetic stress tests.* We provide controlled 1D benchmarks (gaussian, high-cut, multi-peak, high-frequency, delta, uniform, and exponential) to isolate generative capabilities and diagnose topological failure modes prior to deployment on real physics data.

2 Related work

Two computational pillars dominate subatomic-physics data pipelines: forward generative simulation [7] and detector unfolding [8]. Both are bottlenecked by computationally expensive GEANT-class algorithms, motivating parallel machine-learning surrogates. On the forward simulation side, architectures have evolved from Generative Adversarial Network (GAN)-based shower simulators (CaloGAN [9]) and phase-space samplers [10] to high-fidelity normalizing flows like CaloFlow [11]. Recent benchmarks systematically validate these fast-simulation surrogates against full generation engines like GEANT4 [12].

On the inverse side, traditional binned unfolding methods like Iterative Bayesian Unfolding (IBU) suffer from dimensionality curses. OmniFold [13] circumvented this via unbinned neural classifiers, initiating a shift toward deep generative unfolding. Modern approaches include conditional invertible networks (cINNs) [14] and Schrödinger-bridge formulations [15]. These domains increasingly converge on a shared toolkit: while diffusion models [16] define transport through stochastic transitions and scale well using denoising or score-matching objectives, Conditional Flow Matching [17] directly regresses the transport vector field that generates a probability path under a deterministic flow. Building on this, ScatterPrism differentiates itself from likelihood-based cINNs [14] by learning a CFM velocity field end-to-end without iterative refinement.

These foundations rapidly populate workflows in both fields, with modern generative architectures yielding analysis-ready unfolding for complex final states [18]. While generative Artificial Intelligence (AI) is vital for EIC simulations [19], current implementations often assume generic metrics safely indicate modeling quality. A crucial gap remains: no prior work systematically addresses the disconnect between training loss convergence and true kinematic fidelity of CFM models, a concern relevant to both NP and HEP. We address this gap on the NP side using a low-multiplicity JLab photoproduction dataset; the diagnostic methodology is designed to transfer to HEP settings, though further validation is left to future work.

3 Methodology

3.1 Datasets and feature representation

MC-POM dataset. The MC-POM (Monte Carlo Pomeron) dataset models exclusive photoproduction $\gamma p \rightarrow \rho^0 p \rightarrow \pi^+ \pi^- p$. This serves as a representative low-multiplicity NP topology. We focus on forward kinematics with low momentum transfer ($|t| < 1 \text{ GeV}^2$), where pomeron exchange dominates. In this regime, the P-wave $\rho(770)$ resonance is prominent in the $M(\pi^+ \pi^-)$ mass spectrum. We utilize a fixed dataset of 8M events, partitioned into an 8:1:1 training/validation/test split (6.4M / 0.8M / 0.8M events) used consistently across all generation and unfolding experiments.

Events initially consist of 24-dimensional vectors encoding the four-momenta of all involved particles ($p_\gamma^\mu, p_1^\mu, p_2^\mu, \pi^\pm$) and derived variables ($t, M_{\pi\pi}, \cos \theta, \phi$) evaluated in the $\pi^+ \pi^-$ helicity rest frame. To eliminate redundancy, we project the data into a 10-dimensional phase space. Excluding the recoil proton (fixed by four-momentum conservation) and extracting the spatial momenta (p_x, p_y, p_z) of the remaining four particles yields 12 components. Dropping the identically zero p_y components of the incident photon and target proton results in the final 10 dimensions. Each feature was standardized to zero mean and unit variance, then scaled by 5.0; this empirically outperformed factors of 1.0 and 2.0. The mismatch with the unit-variance prior $\mathcal{N}(0, I)$ separates source and target supports along $x_t = t x_1 + (1-t) x_0$, amplifying the deterministic velocity signal and improving training stability. The inverse transform recovers physical units.

For detector unfolding, we simulate resolution effects via independent Gaussian smearing of each Cartesian momentum component k of the $\pi^\pm$ tracks, with standard deviation $k^2 \cdot \sigma_{\text{smear}}$ for $\sigma_{\text{smear}} \in \{0.5, 1.0, 2.0\}$. The recoil proton four-momentum is subsequently algebraically inferred ($p_2^\mu = p_\gamma^\mu + p_1^\mu - p_{\pi^+}^\mu - p_{\pi^-}^\mu$). Because pion smearing directly breaks energy-momentum conservation, this inferred state drops off the exact invariant mass shell. A primary advantage of ScatterPrism is its ability to ingest this physically ‘broken’ conditional data and learn the implicit constraints required to project it back onto the exact ground-truth manifold. Throughout training, all physics-informed metrics are monitored on the held-out validation split, and final results in section 4.2 are reported on the held-out test split. We compute the nearest-neighbor ratio R_{NN} , evaluated against the training manifold, as an explicit guard against memorization.

Synthetic mock datasets. To isolate modeling challenges, we construct a diverse suite of 1D synthetic benchmarks (gaussian, high-cut, multi-peak, high-frequency, delta, uniform, and exponential) across configuration presets. These controlled environments enable rigorous ablation studies of mode collapse and fine-grained resolution prior to real physics deployment. Detailed formulations and extended results are in appendix C.

3.2 Conditional flow matching framework

Flow Matching (FM) formulates the generation process as modeling a time-dependent vector field (also referred to as a velocity field) $v_\theta(x, t)$, whose induced Ordinary Differential Equation (ODE) transports a simple prior probability measure $p_0 = \mathcal{N}(0, I)$ to a complex target data distribution p_1 . Since directly regressing the marginal vector field is computationally intractable, we adopt Conditional Flow Matching (CFM) [17]. By conditioning on individual data samples to construct simple, independent probability paths, CFM provides a tractable regression objective that, in expectation, perfectly recovers the underlying marginal vector field.

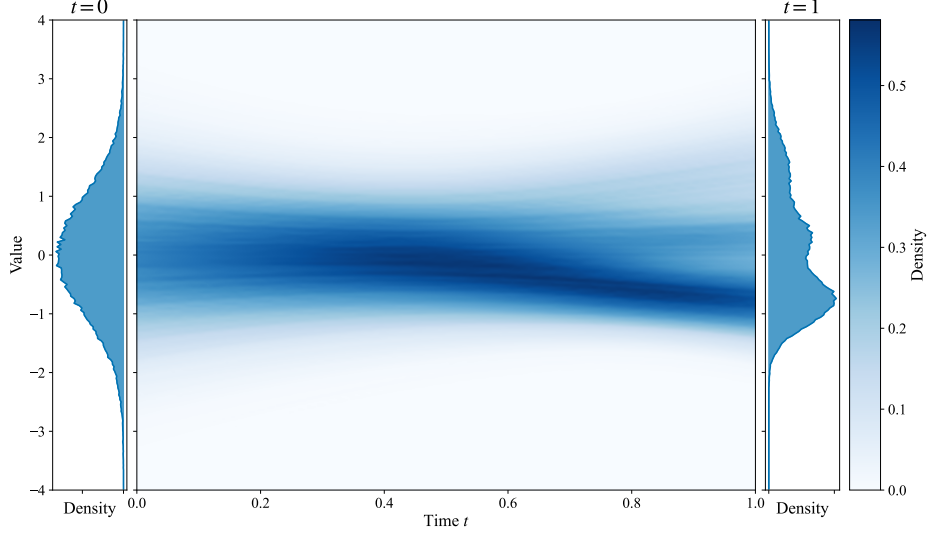


Figure 1. An illustration of the CFM generation process. The model learns to transform a simple base Gaussian distribution (left) to a complex three-peak mixed-Gaussian target distribution (right) by learning the intermediate velocity field (middle).

Given a target data sample $x_1 \sim p_{\text{data}}$ and noise $x_0 \sim \mathcal{N}(0, I)$, CFM constructs interpolated samples along a linear conditional path:

$$x_t = (1 - t)x_0 + tx_1, \quad t \in [0, 1]. \quad (3.1)$$

The conditional vector field generating this path is the constant derivative: $u_t = \frac{dx_t}{dt} = x_1 - x_0$. The CFM loss then trains a neural network to regress against this target velocity:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, x_0, x_1} [\|v_\theta(x_t, t) - (x_1 - x_0)\|^2]. \quad (3.2)$$

An intuitive visual representation of this learned velocity mapping, transporting samples from a base Gaussian noise state through intermediate trajectories toward a multi-modal deterministic target, is provided in figure 1.

We use two distinct network variants under this shared CFM objective:

Unconditional generation network. For generative simulation, we train an unconditional velocity network $v_\theta(x_t, t)$ that takes as input the current state x_t concatenated with a Fourier embedding of t , autonomously mapping pure Gaussian noise into the target physics distribution.

Conditional unfolding network. For detector unfolding, we train a conditional velocity network $v_\theta(x_t, t | c)$ parameterized by the detector-level measurement c . In this context, c explicitly encodes the resolution-degraded kinematic observables — specifically, the Gaussian-smeared spatial momenta of the $\pi^\pm$ tracks representing finite instrumental precision. This approach is conceptually analogous to conditional invertible-network unfolding [14] but utilizes a simulation-free CFM objective. The conditioning vector c is processed through a learned two-layer SiLU-activated Multi-Layer Perceptron (MLP) and concatenated alongside the x_t state and time embedding. During inference, this condition

c is held constant throughout the entire ODE integration, allowing the network to deterministically recover particle-level kinematics from localized, smeared observations.

Both variants utilize residual backbone architectures [20] with SiLU activations. At inference, generation is performed by integrating the learned ODE from $t = 0$ to $t = 1$ via an adaptive Dormand-Prince solver using deterministic paths, establishing a deterministic mapping between the prior and target distributions. Comprehensive hyperparameter configurations of both networks are provided in appendix A.

3.3 Physics-informed metrics

To rigorously evaluate model performance, we monitor the following physics-informed metrics in addition to standard training metrics, such as loss, during the validation and testing phases:

1. *Marginals*: We report the χ^2 statistic and the Wasserstein-1 distance (W_1) between generated and true univariate distributions.
2. *Pairwise joints*: We report the 2D binned χ^2 statistic (χ_{2D}^2) over all feature pairs, testing whether the model captures bivariate dependencies beyond individual marginals.
3. *Global correlation structure*: We report the correlation matrix distance $D_{\text{corr}} = \|\text{corr}(\text{truth}) - \text{corr}(\text{gen})\|_F$, the Frobenius norm of the Pearson correlation-matrix difference, measuring the holistic reproduction of linear dependencies across all channels.
4. *Memorization*: We report the nearest-neighbor distance ratio $R_{\text{NN}} = \bar{d}_{\text{gen} \rightarrow \text{train}} / \bar{d}_{\text{train} \rightarrow \text{train}}$, where $\bar{d}$ denotes the mean L^2 nearest-neighbor distance. A ratio $R_{\text{NN}} \approx 1$ indicates generalization, whereas $R_{\text{NN}} \ll 1$ flags memorization.

In tabular summaries, we present χ^2 , W_1 , χ_{2D}^2 , D_{corr} , and R_{NN} . For dynamic tracking (figure 2(a)), we monitor training loss, Number of Function Evaluations (NFE) (lower values indicate straighter flows for adaptive solvers), W_1 , and D_{corr} . Detailed mathematical formulations are in appendix B.

4 Results

4.1 Synthetic benchmark validation

To isolate modeling challenges, we first validated the architecture on synthetic 1D distributions; detailed results with complex topologies are provided in appendix C.

4.2 Performance on MC-POM dataset

A crucial observation during MC-POM training is that the convergence of the standard CFM velocity loss does not align with true physical fidelity. As figure 2(a) shows, the CFM loss plateaus rapidly after ~ 20 epochs. In contrast, physics-informed metrics (W_1 , D_{corr}), evaluated on the held-out validation split, improve steadily until epoch 600. Thus, CFM loss alone cannot guarantee accurate kinematic reconstruction, necessitating decoupled physical validation metrics.

Table 1 summarizes quantitative performance on the MC-POM dataset for generation and unfolding tasks. All evaluations report metrics from the best-performing checkpoint. Distributional metrics (including the unnormalized 50-bin χ^2 , W_1 , χ_{2D}^2 , D_{corr}) are computed on the held-out test split using

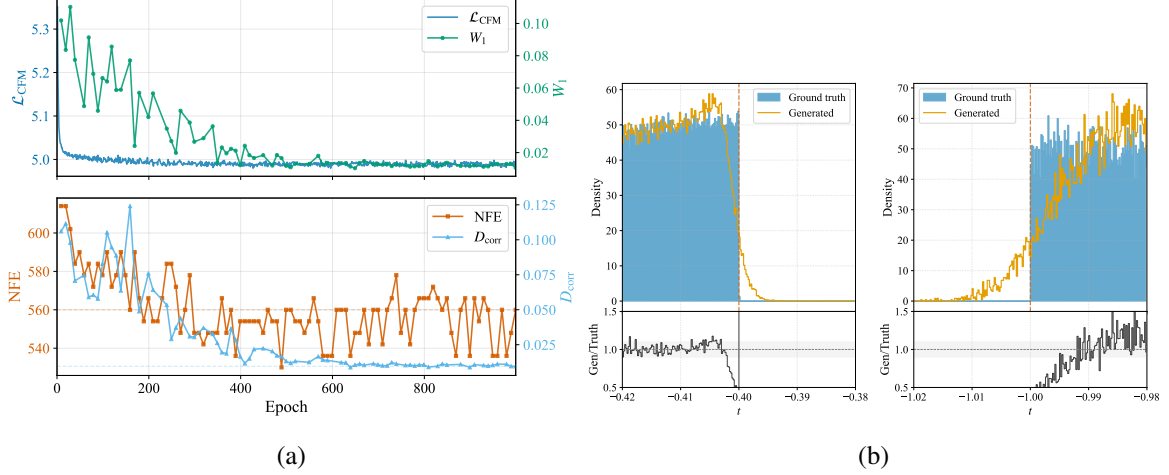


Figure 2. Diagnostics on the MC-POM generation task. (a) Training metrics tracked over time, comparing the CFM loss against physics-informed indicators. (b) Close-up comparison of generated and ground-truth distributions in two sharp cut-off regions of the t -channel.

0.8M generated events (matched 1:1 to the test split size), whereas R_{NN} compares 80K generated events against the 6.4M-event training split to diagnose memorization.

For unconditional generation, the model achieves high-fidelity sampling over the entire phase space, with $R_{\text{NN}} \approx 1.00$ confirming generalization. Figure 3 shows strong agreement between generated and ground-truth distributions (correlation matrix in appendix D, figure 14). Exact-zero channels and axis units (GeV, GeV^2 , radians) are omitted for clarity. A close-up of the t -channel (figure 2(b)), computed from the generated 10D momenta, reveals minor deviations only near hard kinematic cutoffs — a known limitation of continuous flows.

For detector unfolding, the model deterministically maps smeared observations back to particle-level truth. Table 1 shows unfolding metrics remain comparable across smearing scales; small variations are likely attributable to training stochasticity. Figure 4 confirms degraded variables at $\sigma_{\text{smeared}} = 1.0$ are restored to high-fidelity distributions (extended validations in appendix E). R_{NN} is omitted for unfolding because proximity to the training manifold is the desired objective there, not a memorization failure mode.

Table 1. MC-POM generation and detector unfolding performance across various smearing scales σ_{smeared} . A ratio $R_{\text{NN}} \approx 1$ confirms generalization without data memorization.

Task	σ_{smeared}	$\chi^2 \downarrow$	$W_1 \downarrow$	$\chi^2_{2\text{D}} \downarrow$	$D_{\text{corr}} \downarrow$	R_{NN}
Generation	—	33503.2	8.96×10^{-4}	33962.4	1.53	1.01
	2.0	33364.4	1.51×10^{-4}	29245.0	1.44	—
Unfolding	1.0	33355.6	1.60×10^{-4}	40226.7	1.48	—
	0.5	33346.0	1.28×10^{-4}	38270.9	2.11	—

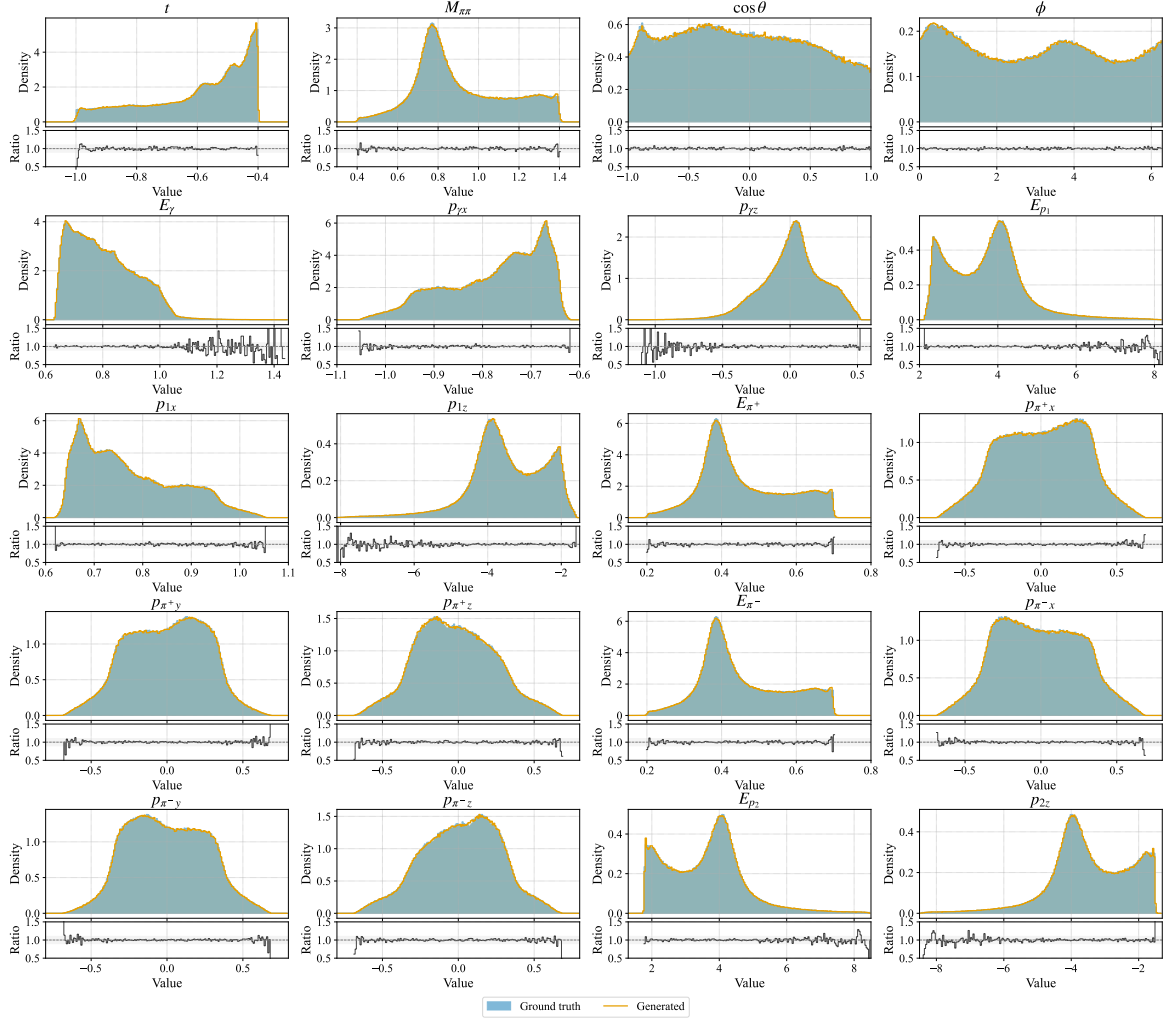


Figure 3. Comparison of generated kinematic distributions produced by the CFM model against the ground truth on the JLab MC-POM dataset.

5 Discussion and conclusion

For practitioners applying generative models across NP and HEP, our results show that standard CFM velocity loss convergence can mislead and does not ensure physical-observable fidelity. Because physical distributions need longer training to stabilize, decoupled evaluation using domain-specific observables is essential. ScatterPrism — validated on synthetic and JLab datasets — demonstrates that CFM provides a robust, deterministic framework for capturing phase spaces and unfolding detector kinematics.

By learning flexible mappings between detector observables and particle truths, ScatterPrism provides an AI-driven, unbinned alternative to expensive Monte Carlo simulations. Initiated by the convergence pathologies exposed in the JLab NP dataset, our methodology reliably captures low cross-section topologies (e.g., near-threshold J/ψ photoproduction, exotic mesons) vital for the EIC. Furthermore, our synthetic benchmarks demonstrate that this diagnostic framework is inherently dataset-agnostic. Consequently, identical machinery is positioned to extend to complex HEP final states,

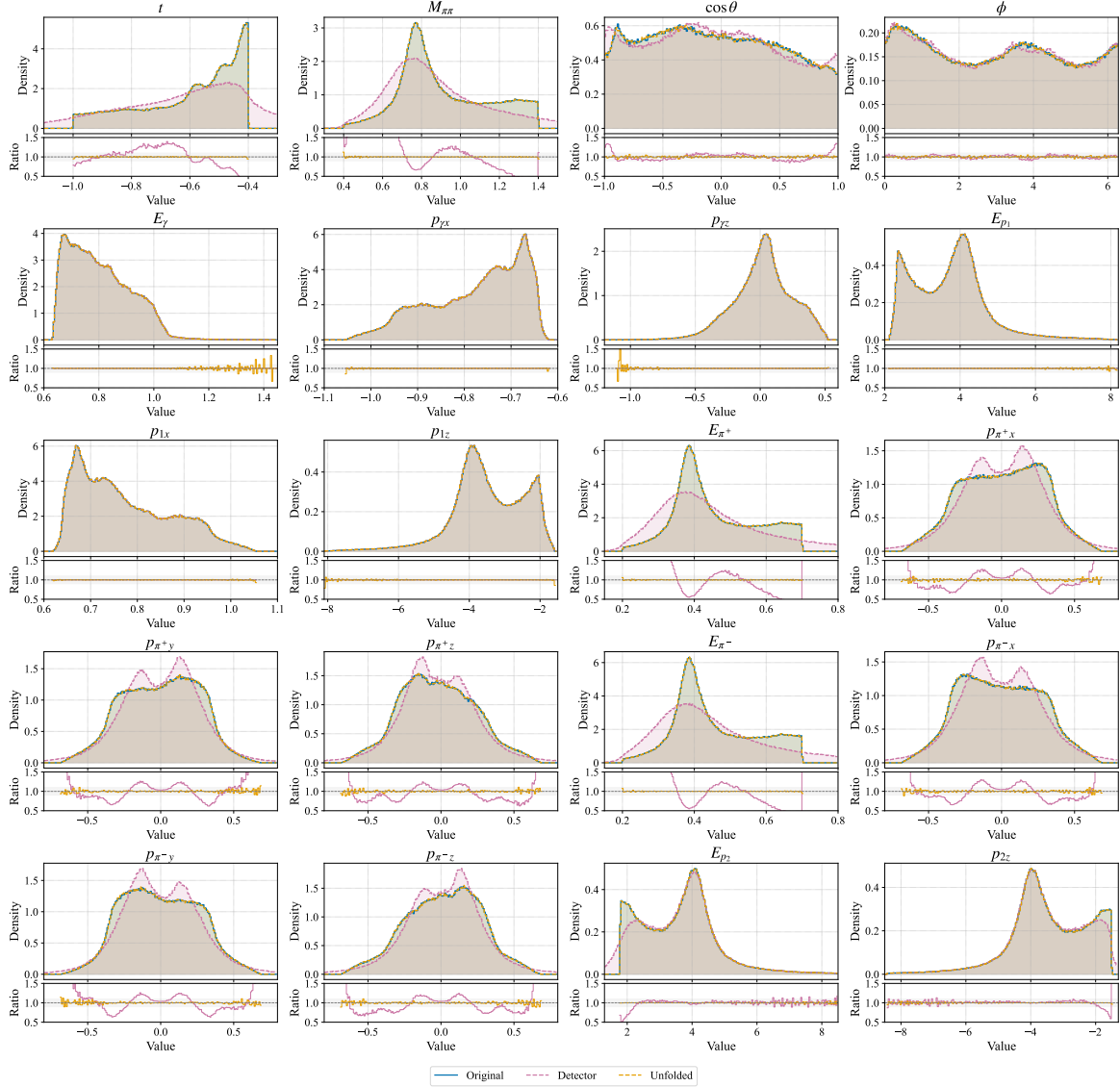


Figure 4. Unfolding ($\sigma_{\text{smear}} = 1.0$) mapping detector-level distributions back to particle-level truth, compared against the ground truth and smeared detector-level inputs.

such as jets and high-multiplicity multi-particle decays. This unified approach enables rapid systematic iterations without repeated simulation campaigns (see computational throughput in appendix F).

ScatterPrism prioritizes modularity and reproducibility via PyTorch Lightning [21], Hydra [22], and Weights & Biases [23]. We utilized standard CFM, achieving high-fidelity structural reconstruction across joints and marginals without the high cost of calculating optimal couplings via Optimal Transport Conditional Flow Matching (OT-CFM) on large datasets.

Beyond nuclear physics, this methodology naturally extends to domains such as molecular dynamics, medical imaging, and astrophysics [24, 25], which frequently require mapping noisy observations back to foundational truth — mirroring detector unfolding. Our multi-metric validation provides a blueprint for ensuring generative models satisfy physical constraints rather than merely minimizing probability divergences.

Having established the necessity of physics-informed convergence diagnostics in CFM, future deployments will natively integrate GEANT-based detector simulations. The lightweight architecture supports uncertainty quantification, architectural ablations, and benchmarking. Future pipelines will add tests for out-of-distribution generalization and explicit physics-informed loss functions [26–28] to eliminate invalid generations and improve unfolding precision.

Acknowledgments

We thank Xuweiyi Chen (University of Virginia) for sharing his valuable experience in developing the model. We also thank Huilin Huang (University of Virginia) for her financial support. This work was partially supported by the National Science Foundation under POSE award 2346173.

Data and Code Availability Statement. Code, models, and datasets are available on GitHub (<https://github.com/xiazeyu/ScatterPrism>) and Zenodo (<https://doi.org/10.5281/zenodo.20407373>, <https://doi.org/10.5281/zenodo.20391126>).

Artificial Intelligence Disclosure. The authors utilized Gemini 3.1 Pro, Claude Opus 4.5/4.6/4.7 to refine prose and assist with code/documentation. All AI-generated content was thoroughly reviewed, verified, and edited. The authors take full responsibility for the content, accuracy, and integrity of this publication.

A Network architecture and training details

Building upon the methodology in section 3.2, both network variants share a residual MLP architecture. Table 2 lists the primary hyperparameters. The unconditional network evaluates the concatenated state $[x_t \parallel e_t] \in \mathbb{R}^{D+d_t}$, where e_t is the Fourier time embedding. For the conditional variant, the detector measurement $c \in \mathbb{R}^D$ is embedded into $e_c \in \mathbb{R}^{128}$, expanding the input to $[x_t \parallel e_t \parallel e_c] \in \mathbb{R}^{D+d_t+128}$.

Table 2. Network architecture and training hyperparameters.

Architecture	Value	Training	Value
Input dimension D	10	Optimizer	AdamW
Hidden dimensions	$[512]^6$ (6 layers)	Learning rate	10^{-4} (decay 0.5, patience 50)
Time embedding dim.	64	Weight decay	10^{-5}
Activation function	SiLU	Batch size	20K
Total parameters	$\sim 2.7\text{M}$	Total epochs	1000

Time conditioning. The scalar time $t \in [0, 1]$ is embedded using 32 geometrically-spaced frequencies up to a maximum frequency $\omega_{\max} = 64$. A learned linear projection maps this Fourier feature into a $\mathbb{R}^{64}$ vector, providing the network with high-frequency temporal components necessary for resolving sharp changes in the velocity field.

Residual blocks. The six hidden widths in table 2 expand into an input linear projection, followed by five residual blocks, followed by an output linear projection. Each residual block applies a two-layer SiLU-activated sequence, $\text{SiLU}(\text{Linear}(\text{SiLU}(\text{Linear}(h)))) + h$, promoting stable gradient flow and facilitating the learning of identity mappings where the vector field is approximately constant.

Optimization details. We optimize parameters with AdamW. A ReduceLROnPlateau scheduler tracks the epoch-level validation CFM loss (`val/loss_epoch`), decaying the initial 10^{-4} learning rate by a factor of 0.5 upon detecting plateaus over a 50-epoch patience threshold (capped minimally at 10^{-7}). For efficiency, validation metrics are computed each epoch on the full held-out validation split using a fixed-step fourth-order Runge-Kutta (RK4) integrator to accelerate the training loop; the adaptive Dormand-Prince (DOPRI5) solver described below is used for all final-prediction reporting. The NFE traces in figure 2(a) are recomputed post hoc with DOPRI5 on a fixed 50K-event subset per checkpoint. The globally optimal checkpoint persisting into final tabular summaries monitors the pure generative physical reconstruction (`val/chi2_mean`) over the raw CFM velocity loss, as discussed in detail throughout section 4.2.

Generation and inference details. For the Dormand-Prince (DOPRI5) ODE solver introduced in section 3.2, absolute and relative tolerances are uniformly set to 10^{-7} for unconditional generation and relaxed to 10^{-3} for unfolding tasks. On a single NVIDIA RTX A6000 GPU, one training iteration on a 20K-event batch completes in ~ 24 ms (~ 820 K samples/s). Inference throughput reaches ~ 3.0 K samples/s for strict unconditional generation and ~ 83 K samples/s for conditional unfolding, with a detailed breakdown provided in section F.

Implementation and environment details. The ScatterPrism framework is implemented in Python (≥ 3.13) utilizing PyTorch ($\geq 2.10.0$) and PyTorch Lightning (2.6.4) [21] for hardware-agnostic training. Core underlying dependencies include `torchdyn` (1.0.6) for numerical ODE integration. The V100 node used PyTorch 2.10.0 with CUDA 12.8, while the A6000 and dedicated CPU nodes used PyTorch 2.12.0 with CUDA 13.0.

All model training and computational evaluations were primarily executed on the UVA Rivanna High-Performance Computing Cluster. GPU acceleration was performed using a single NVIDIA Tesla V100-SXM2 tensor core GPU (32 GB VRAM; utilizing 4 cores of an Intel Xeon Gold 6230 @ 2.10 GHz, SMT off) and an NVIDIA RTX A6000 GPU (48 GB VRAM; utilizing 4 cores of an AMD EPYC 7352, SMT off). CPU-only benchmarking was performed on a dedicated dual-socket AMD EPYC 9454 node (2×48 physical cores with SMT disabled, 96 cores total). Exact SLURM submission scripts and environment configurations used for these benchmarks are provided in the accompanying artifacts.

B Evaluation metrics

To rigorously assess the fidelity of the generated kinematic distributions against the ground truth, we utilize a structured set of evaluation metrics spanning marginal distributions, multivariate structures, and network memorization characteristics:

Marginals. The quality of individual feature distributions is measured primarily using the χ^2 statistic and the Wasserstein-1 distance (W_1):

- **Mean χ^2 (val/chi2_mean):** Evaluated over 1D binned histograms of individual features. To prevent artifacts from outlier limits, exactly 50 uniform bins are dynamically bound between the minimum and maximum values of the true expected distribution. The generated histograms are normalized to match the total event count of the truth distribution before evaluation. Formally, for a single feature, the statistic is:

$$\chi^2 = \sum_{i=1}^{50} \frac{(O_i - E_i)^2}{E_i} \quad (\text{B.1})$$

where O_i is the normalized generated count and E_i is the expected true count in the i -th bin.

- **Mean Wasserstein distance (val/wasserstein_mean):** Measures the minimum mass-transport distance required to transform the generated 1D marginal distributions to the true distributions. To explicitly circumvent the manual binning limits of χ^2 , this metric evaluates the raw, unbinned 1D distributions via their cumulative distribution functions (CDFs) $F(x)$ and $G(x)$:

$$W_1 = \int_{-\infty}^{\infty} |F(x) - G(x)| dx \quad (\text{B.2})$$

Pairwise joints. To evaluate bivariate dependencies between kinematic variables, we examine two-dimensional distributions:

- **Mean 2D χ^2 (val/chi2_2d_mean):** Extends the χ^2 formulation to two-dimensional cross-sections across all 45 unique pairwise feature combinations, returning the arithmetic mean. Each axis is divided into 20 uniform bins (yielding exactly 400 rectangular bins per pair), determined entirely by the ground-truth coordinate extremes. This rigorously tests whether the generator captures underlying multivariate dependencies beyond independent marginals.

Global correlation structure. To verify that generated events accurately reproduce multi-dimensional kinematic constraints, we assess holistic structural fidelity using:

- **Correlation matrix distance D_{corr} (val/correlation_distance):** A holistic measure of how effectively the network reconstructs global linear relationships. It is evaluated as the Frobenius norm of the difference between the sample Pearson correlation matrices:

$$D_{\text{corr}} = \|\text{corr}(\text{truth}) - \text{corr}(\text{gen})\|_F \quad (\text{B.3})$$

Memorization. We test for generative generalization using comparative L^2 nearest-neighbor mappings:

- **Nearest-neighbor distance ratio R_{NN} (nn/memorization_ratio):** Serves as our principal guard against model over-fitting. It calculates the ratio of the mean nearest-neighbor distance from generated samples to the training set ($\bar{d}_{\text{gen} \rightarrow \text{train}}$) versus the mean nearest-neighbor distance natively found within the training set itself ($\bar{d}_{\text{train} \rightarrow \text{train}}$):

$$R_{\text{NN}} = \frac{\bar{d}_{\text{gen} \rightarrow \text{train}}}{\bar{d}_{\text{train} \rightarrow \text{train}}} \quad (\text{B.4})$$

A ratio $R_{\text{NN}} \approx 1$ indicates high-fidelity generalization, whereas $R_{\text{NN}} \ll 1$ flags strict dataset memorization. Notably, for certain synthetic mock datasets, the native distance $\bar{d}_{\text{train} \rightarrow \text{train}}$ can be exceptionally small, causing R_{NN} to appear abnormally large; this is an expected geometric artifact rather than an indication of model collapse.

Additional auxiliary scalars monitored behind the R_{NN} evaluation include:

- **nn/D_gen_to_train_mean:** The arithmetic mean L^2 Euclidean distance from generated events to their single closest element in the true training manifold.
- **nn/D_train_to_train_mean:** The native baseline mean L^2 distance computed by matching points from a random sub-sample of the training set to the remainder of the training elements.
- **nn/D_gen_to_train_min** and **nn/D_train_to_train_min:** The corresponding absolute minimum scalar values computed for the distributions above.

C Synthetic benchmark analysis

To isolate and understand the fundamental generative capabilities of the architecture independent of physical kinematics, we conducted extensive evaluations on a suite of synthetic 1D mock datasets. These benchmarks are specifically designed to stress-test the model against complex topological structures commonly encountered in physics, such as multi-modal overlapping resonances, sharp kinematic cut-offs, and high-frequency perturbative noise.

Each synthetic dataset uses the same 8M-event corpus and 8:1:1 train/val/test split as MC-POM (section 3.1). χ^2 and W_1 are computed on the 0.8M-event test split, and R_{NN} compares 80K generated events against the 6.4M-event training split. Table 3 summarizes the resulting fidelity and memorization metrics across these diverse 1D topologies. Across the majority of datasets, the model achieves exceptionally low Wasserstein (W_1) distances and minimal χ^2 deviations, indicating robust macro-level distribution reconstruction. Crucially, the nearest-neighbor diagnostics confirm that the framework generalizes without duplicating training events: R_{NN} stays within a narrow band of 0.54 (NARROW-WIDE-OVERLAP) to 15.39 (EXPONENTIAL-DECAY) across all realistic topologies, with $D_{g \rightarrow t}$ remaining the same order of magnitude as the baseline $D_{t \rightarrow t}$. Values mildly above unity indicate generated samples sit slightly farther from any training point than the typical training neighbor — consistent with smooth generalization rather than memorization (where $R_{\text{NN}} \ll 1$). The single large positive outlier, $R_{\text{NN}} \approx 3650$ for UNIFORM-FLAT, is a denominator-driven artifact: a perfectly flat training density packs neighboring events so closely that $D_{t \rightarrow t}$ approaches zero, inflating the ratio independent of any model failure.

Visual evidence corroborating these quantitative metrics is presented in figures 5–10. Upon inspecting the generated distribution histograms, we observed highly accurate shape reconstructions across all topological presets. For successful complex cases, such as the TRIPLE-MIXED scenario (figure 7), the network seamlessly learns to balance multiple overlapping Gaussian distributions with varying heights, widths, and proximity, accurately capturing their specific relative population fractions. The TALL-FLAT-FAR benchmark (figure 8) confirms the model’s ability to resolve widely separated peaks with disparate amplitudes. In the highly challenging NOISE-10SPIKES topology (figure 9), the model precisely resolves sharp, fine-grained structural perturbations rather than artificially blurring

Table 3. Quantitative evaluation on synthetic 1D benchmarks. The table reports marginal fidelity (χ^2 , W_1) and memorization characteristics ($D_{g \rightarrow t}$, $D_{t \rightarrow t}$, R_{NN}) across diverse topological structures.

Dataset	Epoch	$\chi^2 \downarrow$	$W_1 \downarrow$	$D_{g \rightarrow t}$	$D_{t \rightarrow t}$	R_{NN}
BIMODAL-ASYM	999	117.8	1.66×10^{-3}	6.65×10^{-7}	2.01×10^{-7}	3.31
EXPONENTIAL-DECAY	339	88.7	3.23×10^{-3}	1.58×10^{-5}	1.03×10^{-6}	15.39
GAUSS-CUTOFF	639	73.8	7.38×10^{-4}	8.73×10^{-6}	1.20×10^{-6}	7.28
NARROW-WIDE-OVERLAP	449	89.6	6.35×10^{-3}	5.30×10^{-7}	9.74×10^{-7}	0.54
NOISE-3SPIKES	789	97.7	2.20×10^{-3}	5.74×10^{-7}	3.56×10^{-7}	1.61
NOISE-10SPIKES	439	91.0	3.89×10^{-3}	8.05×10^{-7}	5.64×10^{-7}	1.43
TALL-FLAT-FAR	539	95.0	5.27×10^{-3}	2.76×10^{-7}	2.37×10^{-7}	1.17
TRIPLE-FLAT-SPREAD	519	160.6	1.15×10^{-2}	2.74×10^{-7}	1.34×10^{-7}	2.04
TRIPLE-MIXED	389	118.5	8.36×10^{-3}	1.59×10^{-6}	1.91×10^{-7}	8.35
UNIFORM-FLAT	399	101.2	3.05×10^{-3}	1.23×10^{-5}	3.36×10^{-9}	3650.36*

*Denominator-driven artifact: the perfectly uniform training density packs neighboring events so densely that $D_{t \rightarrow t} \rightarrow 0$, inflating R_{NN} . The model generalizes correctly — this does not indicate memorization.

them into a single smoothed envelope. Furthermore, figure 11 demonstrates the deterministic mapping of the learned velocity field in action, illustrating how the CFM vector paths intuitively transport the diffuse base Gaussian noise directly into a condensed, localized delta-function target without severe unphysical scatter.

However, perfectly resolving these dense spatial features fundamentally requires decoupled validation schemas. Common physical modeling failure modes before full convergence are depicted in figure 12, revealing the inherent edge-case sensitivities of flow-based generation. When undertrained, the generated manifolds frequently exhibit smeared mass boundaries along strict kinematic limits, or they produce spurious, bridged population points between completely separate disjoint peaks. Figure 13 actively tracks the resolution of these artifacts by charting the progressive structural refinement over the different training stages. These sequential density profiles underscore that while the general macroscopic envelope of the distribution is identified rapidly, extended parameter optimization is strictly required for the velocity field to harden and reliably resolve targeted physical nuances.

D Extended unconditional generation validation

Beyond analyzing individual 1D marginal distributions, predicting complex multivariate correlations is crucial for validating physical simulations. Figure 14 presents the comparative Pearson correlation matrices for both the generated and ground-truth kinematics in the unconditional MC-POM generation task. The high degree of concordance across the entire parameter space demonstrates the model’s capacity to naturally reconstruct global linear relationships and couple interdependent physical limits, confirming high-fidelity multidimensional learning.

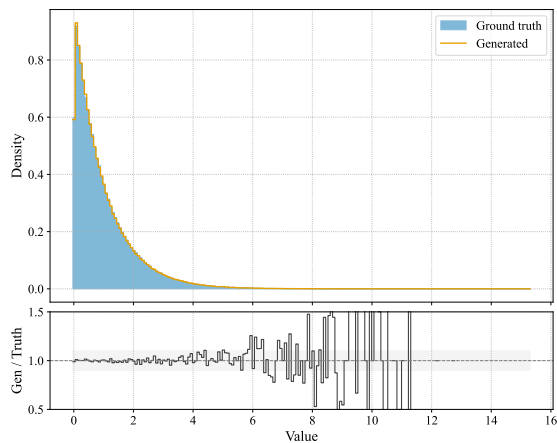


Figure 5. Accurate representation of a sharp exponential decay distribution, smoothly capturing the steep initial drop-off.

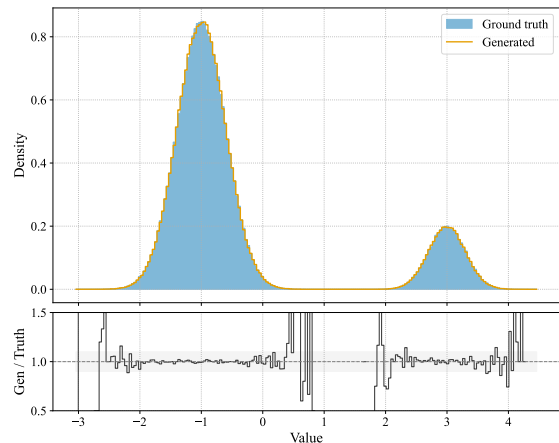


Figure 6. High-fidelity reconstruction of an asymmetric bimodal topology, preserving the distinct independent peak heights.

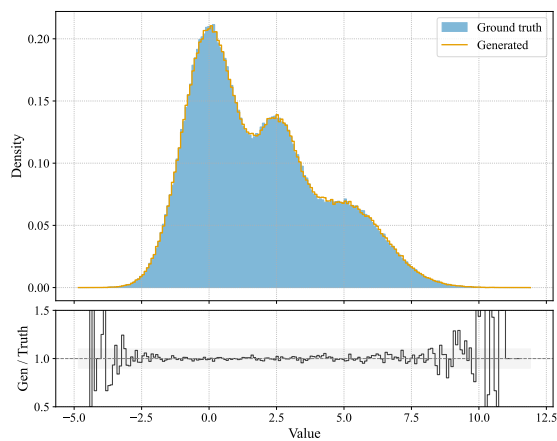


Figure 7. Morphological reproduction of a three-peak mixed Gaussian distribution, matching strict boundaries.

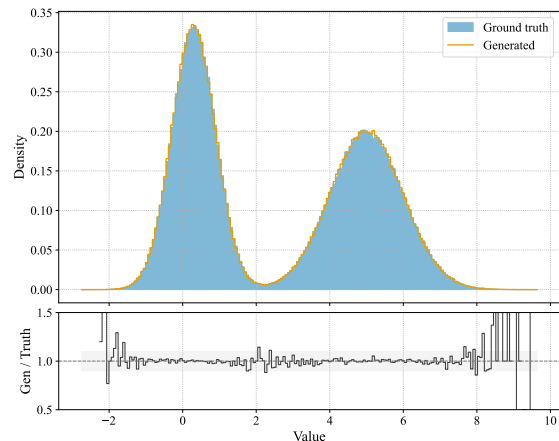


Figure 8. Reconstruction of widely separated multi-modal peaks with disparate amplitudes in the tall-flat-far dataset.

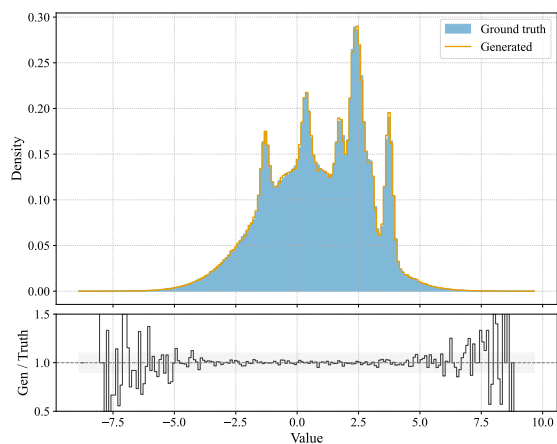


Figure 9. Detailed capture of high-frequency noise spikes, successfully resolving dense structural perturbations.

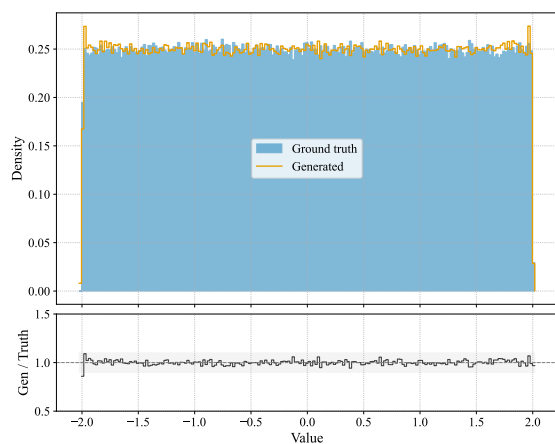


Figure 10. Generation on the uniform-flat distribution, illustrating boundary smearing at sharp discontinuous density edges.

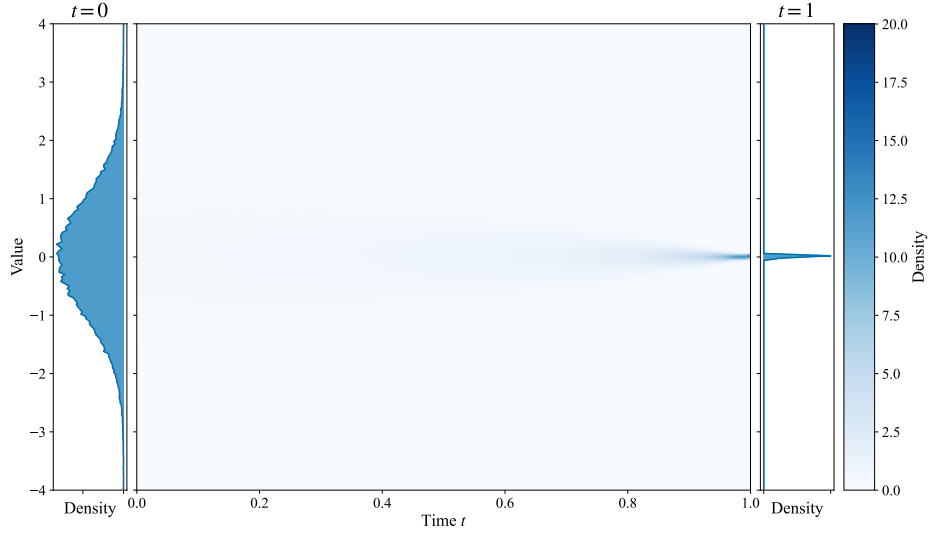


Figure 11. Velocity field density demonstrating the deterministic mapping of base noise vectors into a sharp, localized delta function target.

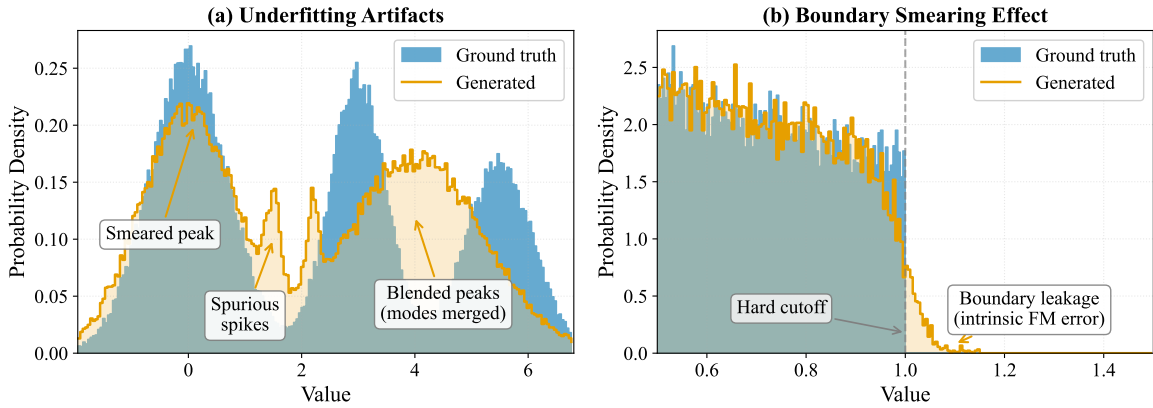


Figure 12. Common generative failure modes before full convergence. (a) Underfitting artifacts exhibit smeared, spurious, and blended peaks. (b) Boundary smearing effects show slight macroscopic edge deviations along strict numerical cutoffs.

E Extended detector unfolding validation

To further assess the robustness of conditional generation across varying degrees of signal degradation, we profile the unfolding consistency under diverse smearing intensities. Figure 15 demonstrates the model’s structural recovery when initialized with severe synthetic resolution degradation ($\sigma_{\text{smear}} = 2.0$), while figure 16 illustrates the near-perfect, high-fidelity phase space restoration achieved from optimally calibrated, low-uncertainty detector signals ($\sigma_{\text{smear}} = 0.5$). Together, these mappings highlight the stable deterministic pathways formed by the conditional CFM architecture regardless of the initial smearing scale.

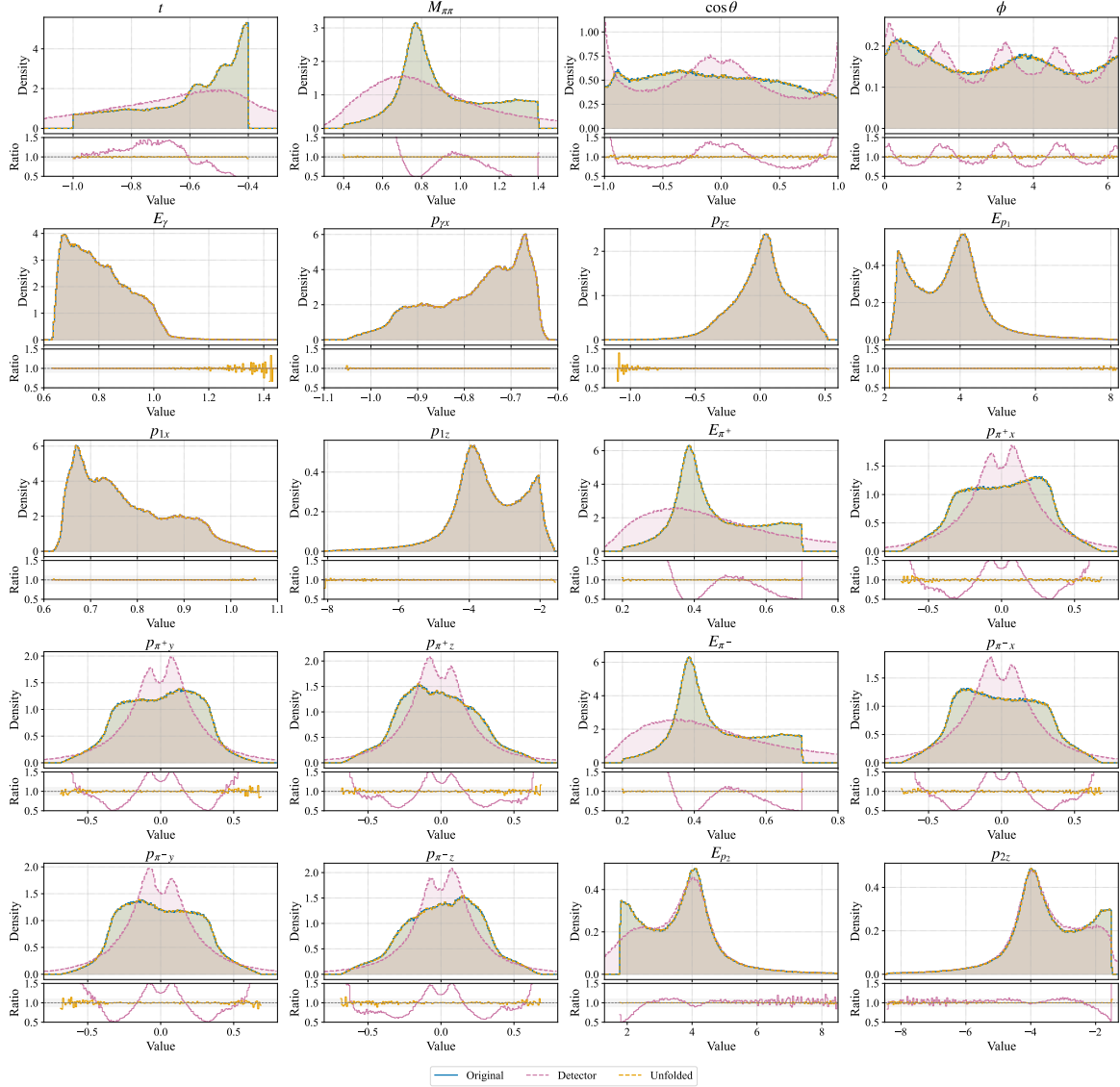


Figure 15. Detector-level unfolding reconstruction initialized from severe simulated momentum smearing ($\sigma_{\text{smear}} = 2.0$). Despite heavy initial degradation, the generative model successfully localizes and recovers the macroscopic physical distributions.

generation (an $\sim 27\times$ speed disparity consistent across GPU hardware). This disparity arises primarily from the ODE solver tolerances: unconditional generation uses strict tolerances ($\text{atol} = \text{rtol} = 10^{-7}$), whereas unfolding employs relaxed tolerances ($\text{atol} = \text{rtol} = 10^{-3}$), requiring far fewer function evaluations per integration. Notably, the relaxed tolerances introduce no measurable degradation in unfolding fidelity (table 1), suggesting that the conservative generation tolerances could be substantially loosened to achieve comparable speedups without compromising distributional accuracy — a promising avenue for future optimization.

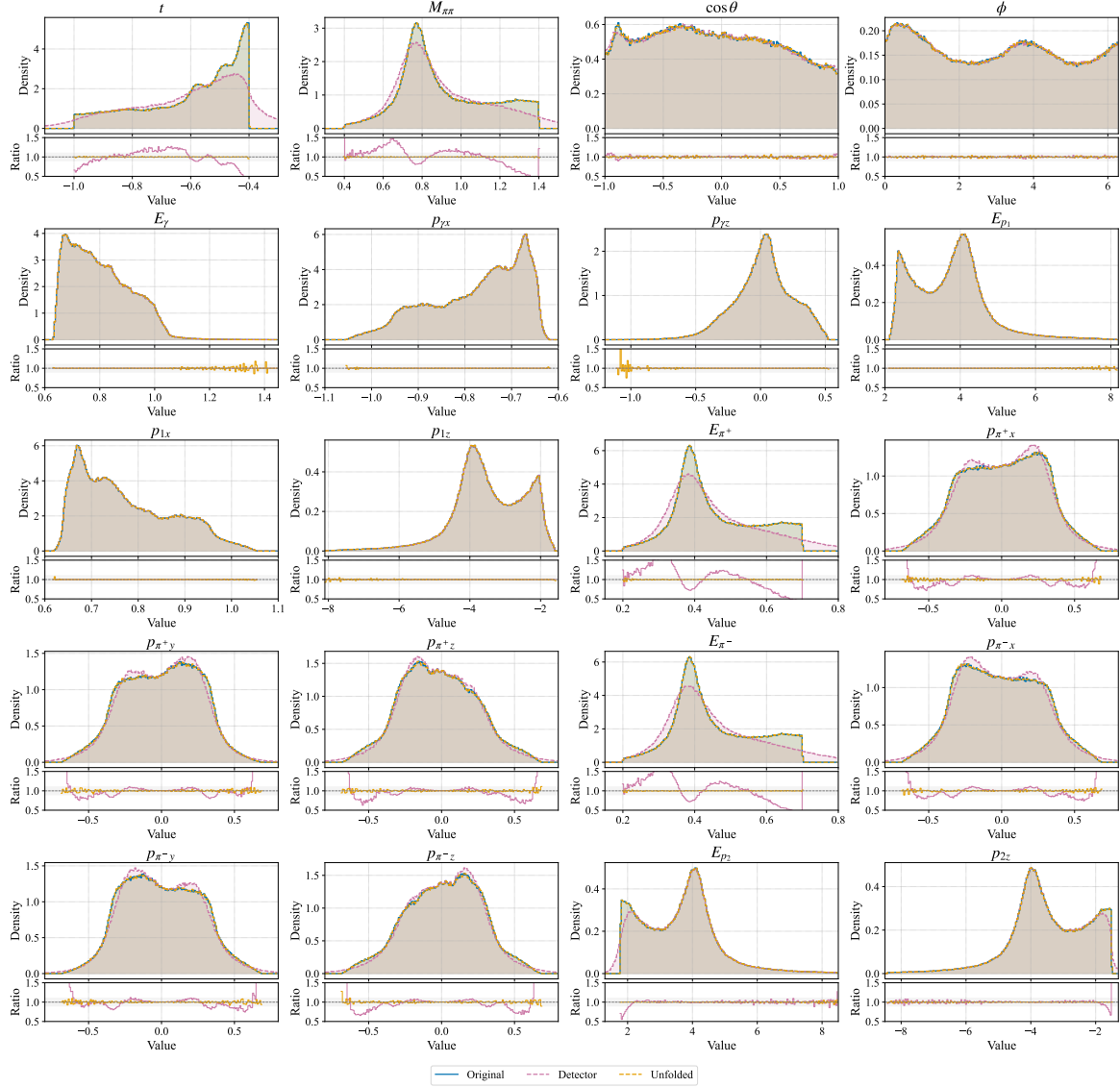


Figure 16. High-fidelity detector unfolding mapped from detector uncertainties ($\sigma_{\text{smeas}} = 0.5$). The generated recovery profiles display exceptional alignment with the particle-level parameters.

Table 4. Compute speed for unconditional generation and conditional unfolding tasks across different hardware configurations.

Task	Hardware	Training		Inference	
		ms / iter ↓	samples / s ↑	s / run ↓	events / s ↑
Generation	CPU (2× AMD EPYC 9454)	2756.7 ± 48.5	7.26×10^3	635.6 ± 1.7	7.87×10^1
	GPU (NVIDIA Tesla V100-SXM2)	31.7 ± 0.9	6.32×10^5	19.6 ± 0.2	2.55×10^3
	GPU (NVIDIA RTX A6000)	24.3 ± 0.1	8.24×10^5	16.5 ± 0.1	3.04×10^3
Unfolding	CPU (2× AMD EPYC 9454)	2907.9 ± 21.0	6.88×10^3	29.8 ± 0.04	1.68×10^3
	GPU (NVIDIA Tesla V100-SXM2)	32.5 ± 0.1	6.15×10^5	0.718 ± 0.001	6.97×10^4
	GPU (NVIDIA RTX A6000)	25.9 ± 0.1	7.72×10^5	0.600 ± 0.003	8.33×10^4

References

- [1] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, *SciPost Phys. Codeb.* **2022** (2022) 8 [[arXiv:2203.11601](#)].
- [2] J. Alwall et al., *MadGraph 5: Going Beyond*, *JHEP* **06** (2011) 128 [[arXiv:1106.0522](#)].
- [3] GEANT4 collaboration, *Geant4 — A Simulation Toolkit*, *Nucl. Instrum. Meth. A* **506** (2003) 250.
- [4] V.D. Burkert et al., *The CLAS12 Spectrometer at Jefferson Laboratory*, *Nucl. Instrum. Meth. A* **959** (2020) 163419.
- [5] GLUEX collaboration, *The GLUEX beamline and detector*, *Nucl. Instrum. Meth. A* **987** (2021) 164807 [[arXiv:2005.14272](#)].
- [6] B. Wang and C. Pehlevan, *An Analytical Theory of Spectral Bias in the Learning Dynamics of Diffusion Models*, *Adv. Neural Inf. Process. Syst.* **38** (2025) 95865 [[arXiv:2503.03206](#)].
- [7] B. Hashemi and C. Krause, *Deep generative models for detector signature simulation: A taxonomic review*, *Rev. Phys.* **12** (2024) 100092 [[arXiv:2312.09597](#)].
- [8] N. Huetsch et al., *The landscape of unfolding with machine learning*, *SciPost Phys.* **18** (2025) 070 [[arXiv:2404.18807](#)].
- [9] M. Paganini, L. de Oliveira and B. Nachman, *CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks*, *Phys. Rev. D* **97** (2018) 014021 [[arXiv:1712.10321](#)].
- [10] A. Butter, T. Plehn and R. Winterhalder, *How to GAN LHC Events*, *SciPost Phys.* **7** (2019) 075 [[arXiv:1907.03764](#)].
- [11] C. Krause and D. Shih, *Fast and accurate simulations of calorimeter showers with normalizing flows*, *Phys. Rev. D* **107** (2023) 113003 [[arXiv:2106.05285](#)].
- [12] F.Y. Ahmad, V. Venkataswamy and G. Fox, *A Comprehensive Evaluation of Generative Models in Calorimeter Shower Simulation*, [arXiv:2406.12898](#).
- [13] A. Andreassen et al., *OmniFold: A Method to Simultaneously Unfold All Observables*, *Phys. Rev. Lett.* **124** (2020) 182001 [[arXiv:1911.09107](#)].
- [14] M. Backes, A. Butter, M. Dunford and B. Malaescu, *An unfolding method based on conditional invertible neural networks (cINN) using iterative training*, *SciPost Phys. Core* **7** (2024) 007 [[arXiv:2212.08674](#)].
- [15] S. Diefenbacher et al., *Improving generative model-based unfolding with Schrödinger bridges*, *Phys. Rev. D* **109** (2024) 076011 [[arXiv:2308.12351](#)].
- [16] J. Ho, A. Jain and P. Abbeel, *Denoising Diffusion Probabilistic Models*, *Adv. Neural Inf. Process. Syst.* **33** (2020) 6840 [[arXiv:2006.11239](#)].
- [17] Y. Lipman et al., *Flow Matching for Generative Modeling*, in the proceedings of *The Eleventh International Conference on Learning Representations*, (2023) [<https://openreview.net/forum?id=PqvMRDCJT9t>] [[arXiv:2210.02747](#)].
- [18] A. Shmakov et al., *Full event particle-level unfolding with variable-length latent variational diffusion*, *SciPost Phys.* **18** (2025) 117 [[arXiv:2404.14332](#)].
- [19] C. Allaire et al., *Artificial Intelligence for the Electron Ion Collider (AI4EIC)*, *Comput. Softw. Big Sci.* **8** (2024) 5 [[arXiv:2307.08593](#)].
- [20] K. He, X. Zhang, S. Ren and J. Sun, *Deep Residual Learning for Image Recognition*, in the proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Las Vegas, NV, U.S.A. (2016), p. 770–778 [[DOI:10.1109/CVPR.2016.90](#)] [[arXiv:1512.03385](#)].

- [21] W. Falcon and The PyTorch Lightning team, *PyTorch Lightning*, Zenodo, DOI:10.5281/zenodo.20317430 (2026).
- [22] *Facebookresearch/Hydra*, Meta Research (2026) [<https://github.com/facebookresearch/hydra>].
- [23] L. Biewald, *Experiment Tracking with Weights and Biases*, (2020) [<https://www.wandb.com/>].
- [24] M.K. Islam et al., *Cosmo3DFlow: Wavelet Flow Matching for Spatial-to-Spectral Compression in Reconstructing the Early Universe*, [arXiv:2602.10172](https://arxiv.org/abs/2602.10172).
- [25] Z. Xia and S. Cheng, *PyTorchFire: A GPU-accelerated wildfire simulator with Differentiable Cellular Automata*, *Environ. Model. Softw.* **188** (2025) 106401 [[arXiv:2502.18738](https://arxiv.org/abs/2502.18738)].
- [26] J.-H. Bastek, W. Sun and D. Kochmann, *Physics-Informed Diffusion Models*, in the proceedings of *Thirteenth International Conference on Learning Representations*, (2025), p. 3360–3385 [https://proceedings.iclr.cc/paper_files/paper/2025/hash/096347b4efc264ae7f07742fea34af1f-Abstract-Conference.html] [[arXiv:2403.14404](https://arxiv.org/abs/2403.14404)].
- [27] G. Baldan, Q. Liu, A. Guardone and N. Thuerey, *Physics vs Distributions: Pareto Optimal Flow Matching with Physics Constraints*, in the proceedings of *The Fourteenth International Conference on Learning Representations*, (2026) [<https://openreview.net/forum?id=tAf1KI3d4X>] [[arXiv:2506.08604](https://arxiv.org/abs/2506.08604)].
- [28] J. Tauberschmidt, S. Fellenz, S.J. Vollmer and A.B. Duncan, *Physics-Constrained Fine-Tuning of Flow-Matching Models for Generation and Inverse Problems*, in the proceedings of *The Fourteenth International Conference on Learning Representations*, (2026) [<https://openreview.net/forum?id=khBHJz2wcV>] [[arXiv:2508.09156](https://arxiv.org/abs/2508.09156)].