Improving the efficiency and precision of data-driven surrogates for Monte-Carlo particle transport in the ALICE experiment

Łukasz Dubiel[0009−0008−3921−4176], Emilia Majerz[0009−0005−2034−0410], and Witold Dzwinel[0000−0001−8321−5928]

AGH University of Krakow, Poland
lukdubiel@student.agh.edu.pl, {majerz,dzwinel}@agh.edu.pl

Abstract In the ALICE experiment at CERN, detailed calorimeter response simulations using Monte Carlo methods implemented in toolkits such as GEANT4 achieve high accuracy but require substantial computational resources. To enable fast simulation for large-scale studies, we investigate machine-learning-based surrogate models for the planned Forward Calorimeter (FoCal), specifically its hadronic component (FoCal-H). Several generative approaches, including GANs, normalizing flows, and diffusion models, are evaluated against GEANT4 using physics-motivated metrics that compare energy profiles or the morphology of the showers. In addition, we introduce a hybrid simulation strategy in which a fast, physics-inspired approximation generates a coarse response that is subsequently refined by a compact generative model. This approach achieves a favorable balance between simulation fidelity and computational efficiency, which makes it suitable for high-rate ALICE analyses.

Keywords: High Energy Physics · Calorimeter Simulation · Surrogate Modeling · Generative Deep Learning · Hybrid approximation · Fast Simulation

1 Introduction

CERN (the European Organization for Nuclear Research) serves as a global hub for high-energy physics, operating the world's largest particle accelerator, the Large Hadron Collider (LHC). These experiments rely on sophisticated detectors, such as calorimeters, to measure the energy of particles emerging from the collisions. To understand the behavior of these devices under specific conditions, as well as for calibration or quality control purposes, high-fidelity simulations of particle detectors are now studied. Traditionally, such simulations rely on the Monte Carlo method, as implemented in toolkits such as GEANT4 [1]. Although this first-principles approach yields realistic results, it is also computationally intensive. Consequently, generating the enormous number of events required for modern collider experiments becomes a major bottleneck. A promising strategy to meet this challenge is to train a data-driven surrogate model to emulate the calorimeter response in place of the full GEANT simulations. Recent advances

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

in machine learning have resulted in generative models capable of learning the probability distributions of particle showers. Well-studied architectures include Generative Adversarial Networks (GANs) [7], diffusion models [8,19] or Normalizing Flows (NFs) [9].

One such measuring device requiring fast simulations is the planned Forward Calorimeter (FoCal), a high granularity calorimeter designed to extend the capabilities of the ALICE experiment into the forward rapidity region (pseudorapidity $\eta \approx 3.3 - 5.3$) [2]. Here, we examine the use of generative frameworks for simulating the FoCal hadronic (FoCal-H) calorimeter.

In summary, the main contributions of this work include:

- We apply state-of-the-art generative frameworks, such as GANs, NFs, and diffusion models, to the task of FoCal-H simulation.
- We introduce a hybrid surrogate model that combines a physics-driven approximation with a lightweight generative enhancer, achieving a favorable trade-off between simulation fidelity and inference latency.
- We demonstrate that incorporating a physics-informed positional loss may significantly improve the modeling of complex multi-shower responses.
- We quantitatively analyze the trade-off between realism and throughput across different generative families, showing that diffusion models maximize shower fidelity while approximation-based approaches enable high-rate simulation.

2 Forward Calorimeter

The primary function of FoCal is to measure the energy and spatial distribution of high-energy particles. FoCal consists of two main modules: electromagnetic (FoCal-E) and hadronic (FoCal-H). FoCal-E is constructed with alternating layers of tungsten absorbers and high-resolution silicon sensors, typically using CMOS pixel sensors or silicon pad sensors. FoCal-H is positioned behind FoCal-E, and its purpose is to absorb and measure the energy of hadrons. It uses thicker absorbers and larger sensors, providing the depth necessary to contain hadronic showers. A schematic overview of the FoCal detector can be found in [2].

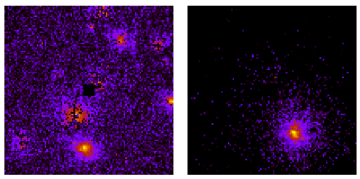

Figure 1: Example of detector activation produced by the decay products of a particle (left, dataset 1) and by the corresponding undecayed particle (right, dataset 2).

In our studies, we focus on FoCal-H. It produces responses which can be seen as an image of a 3D camera. However, for simplicity, they can be reduced to 2D

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

via projection onto the front wall of FoCal. The visualization of such projected responses is shown in Fig. 1, which presents one sample per two studied datasets, where a more intense color corresponds to a higher energy deposited in a given cell. Each detector response was simulated using three input parameters, namely the particle energy E, the pseudo-rapidity η and the azimuthal angle ϕ. The last two allow inferring the trajectory of a particle, so it is possible to predict where a given particle would hit in the most straightforward case, i.e., without any decays; yet, some particles are prone to self-decay, e.g. pions. Consequently, detector responses are possible in two variants. When an undecayed particle hits the detector (dataset 2), a single shower corresponds to a well-defined splash on the image. On the other hand, decaying particles may cause a more chaotic response with multiple mini-showers (dataset 1).

3 Related Work

The use of deep learning for fast calorimeter simulations has become an actively researched area. Many methods were proposed during the Fast Calorimeter Simulation Challenge (CaloChallenge) [12]. GANs were the first to be adopted, with pioneering work such as LAGAN [16] and CaloGAN [17]. Although early models struggled with high-fidelity reproduction of complex correlations, recent advancements such as CaloShowerGAN [6] and Mixture of Generative Experts [4] have significantly improved performance. GANs remain one of the fastest inference engines, but often suffer from training instability and mode collapse.

NFs offer tractable likelihoods and stable training. CaloNormFlow [14] demonstrated superior fidelity over GANs, while CaloINN [5] achieved top performance in the CaloChallenge by combining Invertible Neural Networks with VAEs. Approaches such as Layer-to-Layer Flow [3] further adapted these invertible architectures for multilayer detector geometries. Complementing the coupling-layer approaches, Inverse Autoregressive Flows (IAFs) have also demonstrated strong performance on calorimeter data [13]. Recent work has shown the effectiveness of IAFs combined with physics-based modeling for fast simulation of the ALICE Zero Degree Calorimeter (ZDC) [15].

Denoising Diffusion Probabilistic Models (DDPMs) [8] have established a state-of-the-art in generation quality. Applications to the ZDC detector [10] show that diffusion models outperform GANs in Wasserstein metrics. However, their iterative sampling process makes them orders of magnitude slower. Techniques such as Denoising Diffusion Implicit Models (DDIMs) [19] or Latent Diffusion Models [18] attempt to mitigate this computational cost, often at the expense of slight quality degradation.

Comprehensive research on the ZDC detector, presented in [22], highlights the trade-offs in current generative approaches. The authors observed that fast models often compromise the quality of the simulation - VAEs tend to produce blurred energy depositions and GANs struggle with mode collapse in sparse regions. These limitations motivate the exploration of more expressive density estimators, such as diffusion models and NFs. Complementarily, [21] shows sig-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

nificant capabilities of a newer approach with Flow Matching. Another recent study introduces a novel method for fast simulation using supermodeling with a mixture of generative experts [4] that achieves a significant boost in data fidelity without the deterioration of generation speed. Machine learning has also been applied to related detector tasks, including improving the precision of fast-sampled timing detectors [11].

Finally, inspired by research on medical image reconstruction [20], hybrid approaches, where a generative model acts as an enhancer for low-quality approximations, offer a promising direction to balance speed and fidelity.

4 Methods

In this work, several generative architectures are studied and compared, including GANs, diffusion models, NFs, and a hybrid one that incorporates image enhancement techniques.

4.1 Generative Baselines

Generative Adversarial Networks GANs, introduced in [7], are a class of generative models that learn to produce synthetic data samples by establishing a game between two neural networks: the generator and the discriminator. The generator G maps latent space vectors $z \approx p(z)$ to the data space x, with the objective of producing outputs that are indistinguishable from real data. The discriminator D, on the other hand, learns to distinguish between real samples drawn from the true data distribution $p_{data}(x)$ and those generated by G. The two networks are trained simultaneously using the following *minimax* objective:

$$\min_G \max_D \; \mathbb{E}_{x\sim p_{\text{data}}(x)} \left[\log D(x)\right] + \mathbb{E}_{z\sim p(z)} \left[\log(1 - D(G(z)))\right]. \tag{1}$$

Normalizing Flows NFs are a family of generative models that transform a simple probability distribution (e.g. standard Gaussian) into a complex one via a sequence of invertible and differentiable mappings. The core idea is based on the *change of variables formula* in probability theory.

Let $\mathbf{z} \in \mathbb{R}^d$ be a latent variable with known density $p_Z(\mathbf{z})$, and let $\mathbf{x} = f(\mathbf{z})$ be the observed variable where $f : \mathbb{R}^d \to \mathbb{R}^d$ is invertible and differentiable. Then the density of $\mathbf{x}$ is given by:

$$p_X(\mathbf{x}) = p_Z(\mathbf{z}) \left| \det \left(\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \right)^{-1} \right| = p_Z(\mathbf{z}) \left| \det \left(\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \right) \right|^{-1}. \tag{2}$$

The training objective of the complete transformation consisting of K invertible mappings: $\mathbf{z}_0 \sim p_Z(\mathbf{z}_0)$, $\mathbf{z}_K = f_K \circ f_{K-1} \circ \cdots \circ f_1(\mathbf{z}_0)$ is the maximum likelihood estimation (MLE) in the data $\{\mathbf{x}_i\}$:

$$\max_\theta \sum_{i=1}^{N} \log p_X(\mathbf{x}_i; \theta) = \sum_{i=1}^{N} \left[\log p_Z(f_\theta^{-1}(\mathbf{x}_i)) + \log \left| \det \left(\frac{\partial f_\theta^{-1}(\mathbf{x}_i)}{\partial \mathbf{x}_i} \right) \right| \right]. \tag{3}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

4.2 Conditional Diffusion with DDIM

Diffusion-based generative models define an explicit, tractable generative process by reversing a fixed noise injection procedure. The core idea is to define a forward process $q(x_{1:T} \mid x_0)$ that maps data (x_0) to noise through a Markov chain, and to learn a reverse process $p_\theta(x_{0:T})$ capable of mapping noise to realistic samples. The forward (noising) process is a fixed Markov chain that gradually adds Gaussian noise:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{4}$$

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)\mathbf{I}), \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s. \tag{5}$$

As $t \to T$, the sample approaches an isotropic Gaussian distribution. The schedule $\{\beta_t\}$, where $\alpha_t = 1 - \beta_t$, determines the rate of noise injection.

A neural network $\epsilon_\theta(x_t, t, y)$ approximates the noise added in step t. The training objective is to minimize the lower bound of the evidence (ELBO). After simplification, the standard objective becomes a noise prediction loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,\mathbf{I}), t} \left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2 \right]. \tag{6}$$

Sampling requires many steps (often $T = 1000$), leading to a high computational cost. DDPMs yield high-quality samples because of their expressive reverse distribution.

DDIMs define a deterministic non-Markovian process: $x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \eta_t^2} \epsilon_\theta(x_t, t, y) + \eta_t \epsilon_t$, where $\eta_t = 0$ gives a completely deterministic path. This allows reducing the number of steps to as few as 10–50 while preserving sample quality.

4.3 Positional loss

Here, we introduce an additional physics-based regularization during training. The idea is to guide the model toward the expected outcome derived from physical theory, in this case, the expected hit position.

The expected hit position (c_x, c_y) is derived by projecting the particle's trajectory onto the front wall of the detector. Starting from the definition of pseudorapidity, $\eta = -\ln(\tan(\theta/2))$, the polar angle can be determined as $\theta = 2\arctan(e^{-\eta})$. For a detector located at a longitudinal distance z from the interaction point, the radial distance r from the beam axis in the transverse plane is given by $r = z\tan(\theta)$. Projecting this radius using the azimuthal angle ϕ yields the Cartesian coordinates: $(c_x, c_y) = (r\cos(\phi), r\sin(\phi))$.

Comparing the expected hit coordinates with the ones present in the generated responses introduces the formula for the *positional* loss/distance. The observed hit position can be calculated as the center of mass (energy) of the generated responses. The *positional* loss/distance is calculated as the average L^1 norm between the expected and actual hit positions.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

4.4 Hybrid Architecture: Approximation & Enhancement

For stable particles, detector responses exhibit a characteristic single-peak profile - maximum intensity at the center with radially decreasing values. This raises the question of whether such a pattern can be modeled without using complex neural networks. With methods introduced in Section 4.3, it is possible to calculate the center of such a shower. Then, by taking larger and larger concentric circles centered at the peak and summing the energy deposited within the area enclosed by them, a characteristic of a shower energy distribution can be defined. The same applies to multi-shower responses, but it requires calculating the cumulative energy for each shower center. For a given response R this characteristic can be formulated as follows:

$$\zeta(r, R) = \sum_{j,i:(j-c_y(R))^2+(i-c_x(R))^2<r^2} R_{ji}, \tag{7}$$

where r is the radius of the circle centered at the expected hit location $c = (c_x, c_y)$. Consequently, for a given R, the result of ζ is a nondecreasing function. To cover the entire image of size (105, 105) for every possible hit position, it is necessary to calculate ζ up to $r = 148$. This process is visualized in Fig. 2.

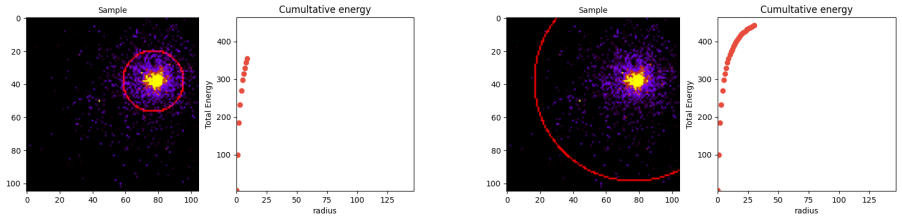

Figure 2: The cumulative energy sum at an early (left) and a later (right) stage.

Thus, image generation reduces to characteristic generation. Given a vector of particle parameters, the task is to predict ζ for all r. For this, some shallow conditional decoder architectures can be used. The target is to minimize:

$$\mathcal{L}(\theta) = \mathbb{E}_{c,\zeta}\left[\|\zeta - \zeta_\theta(c)\|^2\right]. \tag{8}$$

The predicted ζ allows for reconstruction of the approximate response by distributing the energy $e = \zeta_r - \zeta_{r-1}$ equally across all pixels of the ring with radius r. This image can then be fed to a generative framework, e.g. a diffusion model, to improve its quality [20]. Here, we approach the enhancement using conditional diffusion or GAN as previously described, but with the models' input enriched with a low-quality image.

In the GAN case, we propose that the conditional vector fed to the discriminator contains only the particle features, without the initial approximation of the detector response. The early experiments have shown strong dominance of the discriminator in most setups, and thus adding more information seemed to be unnecessary. In the case of diffusion, conditioning can be performed by merging

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

the noisy input image and the low-quality approximation along the channel dimension, and training the UNet to recognize 1-channel noise based on 2-channel input; the inference utilizes a similar logic. For GANs, the merging is done when the vector upsampled from the latent space reaches the target response size.

5 Experiments and Results

Here, we present the main results of our studies. We begin this Section with the datasets and metrics descriptions, followed by training setup details. The results for different generative models are presented in Tabs. 1, 2, and then discussed in Sections 5.4 to 5.8. For each dataset and architecture, we present the generated samples in Fig. 3. In our studies, we verified multiple model configurations for each generative framework, and here we include the results for the best-performing setups. Due to the significant computational cost of the architectures evaluation, hyperparameter tuning was conducted heuristically. For diffusion models, we adapted stable configurations established in previous related studies [22]. For GANs, NFs, and approximation models, we performed an empirical search over standard parameter ranges.

5.1 Datasets

We perform our studies using two datasets introduced in Section 2. Dataset 1 (DS1) consists of 18,346 samples, whereas dataset 2 (DS2) of 29,824. We divided the datasets into separate training, validation and test sets (70:10:20).

5.2 Metrics

Response-wise metrics To quantify spatial agreement, we utilize the Wasserstein distance. The *Wass. Sum* is defined as the sum of Wasserstein distances between the respective horizontal and vertical energy profiles (marginals) of the generated and reference showers. To assess visual properties, Perceptual Loss (LPIPS [23]) is used. For the less diverse one-shower DS2, we also compute mean absolute errors (MAEs) between pixel values.

Global metrics To evaluate the general network output, we employ global distance metrics. The fidelity of the total deposited energy is measured by Total Energy Wasserstein (*Total Wass.*), which is the Wasserstein distance between the histograms of total energies of real and generated samples divided by their maximums. Furthermore, to assess the relationship between particle energy and total response energy, we employ the Joint Energy Wasserstein (*Joint Wass.*). This metric corresponds to the Sliced 2D-Wasserstein distance between the joint distributions of particle energy and total response energy, approximated via Monte Carlo projections onto random unit vectors. Standardization is applied to both dimensions independently to ensure that the metric focuses on the shape of the distributions rather than the absolute scales.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

We performed evaluations with a batch size of 1024, measuring only the generative pipeline execution with data pre-loaded into VRAM. The time values presented later are derived from the minimum measured duration of a single batch, scaled by a factor of 10. This way, we minimize the impact of the machine load from other calculations on our measurements, which can be significant in the supercomputer setting.

5.3 Training Setup

In the case of GANs, the generator and discriminator are optimized separately using the Adam optimizer with configurable momentum parameter β_1 (default $\beta_1 = 0.9$) and fixed $\beta_2 = 0.999$. No learning rate scheduling is applied. The adversarial objectives use binary cross–entropy with logits for both the discriminator and the generator, optionally augmented with a positional regularizer.

Diffusion uses AdamW with $(\beta_1, \beta_2, \epsilon, \text{wd}) = (0.87, 0.82, 2.5 \times 10^{-10}, 5.5 \times 10^{-4})$, coupled to a cosine learning rate schedule with a warm-up phase that covers 10% of the total optimization steps. The loss is the mean squared error between the predicted and ground-truth noise (with an optional positional penalty). Scheduler updates and gradient clipping are performed at each iteration.

The flow models are trained with Adam $(\beta_1, \beta_2) = (0.9, 0.999)$ and a *OneCycle* learning rate policy whose peak is equal to the nominal learning rate; the schedule is parameterized by the number of epochs and steps per epoch.

The feed–forward approximation model is trained with AdamW using the same $(\beta_1, \beta_2, \epsilon, \text{wd})$ coefficients as for the diffusion model, without the learning rate scheduling. The approximation-guided GAN and diffusion utilize the same setups as their non-guided versions.

We carried out the experiments on two Cyfronet AGH GPU clusters, *Athena* (equipped with NVIDIA A100 40 GB) and *Ares* (equipped with NVIDIA V100 32 GB), as well as on a workstation with an NVIDIA RTX A3000 6GB GPU. We chose batch sizes in the range 32–512, depending on the model family, the optimizer, and the available device memory. For some setups, especially for diffusion, the choice of smaller batches (below 96) promoted faster convergence without negative effects on stability. Each model was trained for 70–300 epochs.

5.4 GAN experiments

We tested several versions of the generator. All models map concatenated latent and conditional inputs to an initial $7 \times 7 \times 256$ tensor. To reach the target resolution, the architectures employ either transposed convolutions or nearest-neighbor interpolation, with some variants additionally incorporating residual blocks. We also used standard techniques like Batch Normalization and Leaky ReLU activations.

For DS2, the use of positional loss increases the value of the L1-based distance metric and Wasserstein distances. On the other hand, for the DS1 (the multiple showers dataset), only the GAN with this regularization gives visually satisfactory results (shown in Fig. 3, DS1 column, GAN row), while other setups

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

Table 1: Performance evaluation metric values (Wasserstein distances, perceptual and positional distances) and generation times for various generative models trained on the multi-shower dataset (DS1). Best results for each metric are in **bold**, and second-best are underlined.

Family	#params	Wass. Sum	Perceptual Dist.	Positional Dist.	Total Wass.	Joint Wass.	Time [s]
GAN	1.0M	33.5	0.504	18.6	0.148	1.76	1.22
GAN+L1	1.0M	22.3	0.472	13.0	0.476	0.829	1.22
GAN+L1	1.4M	13.6	0.439	12.5	0.289	**0.051**	2.18
Diffusion	4.3M	**13.5**	**0.354**	12.0	0.179	0.121	307.8
Diffusion+L1	4.3M	130.1	0.430	26.3	0.283	1.35	307.8
NF	51.8M	19.5	0.393	9.84	0.370	1.38	7.87
NF	270.1M	19.9	0.395	9.68	0.114	1.44	15.4
Approx	10.6k	23.1	0.638	**4.01**	0.238	0.342	22.3
GAN-Enc	162.9k	40.2	0.573	15.4	**0.012**	1.82	**0.607**
GAN-Enc	87.1k	40.3	0.564	10.8	0.179	1.82	0.640
Diff-Enc	181.3k	13.6	0.384	10.7	0.173	0.287	171.0

Table 2: Performance evaluation metric values (Wasserstein distances, perceptual and positional distances) and generation times for various generative models trained on the single-shower dataset (DS2). Best results for each metric are in **bold**, and second-best are underlined.

Family	#params	MAE	Wass. Sum	Perceptual Dist.	Positional Dist.	Total Wass.	Joint Wass.	Time [s]
GAN	1.0M	790.4	4.44	0.319	18.1	0.081	**0.007**	1.22
GAN+L1	1.0M	837.0	5.81	0.378	25.8	0.283	0.089	1.22
Diffusion	4.3M	611.9	4.37	0.266	7.62	0.176	0.016	307.8
Diffusion+L1	4.3M	1598.2	23.8	0.501	15.8	0.231	0.895	307.8
NF	176.3M	866.4	10.1	**0.226**	7.32	0.376	0.916	14.5
Approx	14.0k	457.3	**3.44**	0.447	**1.49**	0.088	0.037	25.2
GAN-Enc	162.9k	**437.4**	5.22	0.282	2.28	**0.039**	0.570	**0.607**
Diff-Enc	181.3k	548.2	4.35	0.270	4.97	0.152	0.021	171.0

struggle with capturing the diversity of multi-shower responses. The relationship between the characteristics of the conditioning particles and the localization of the shower in the response is straightforward for the single-shower DS2. For DS1, from a physical point of view, the center of distribution of splashes in the multi-shower response should overlap with the theoretical coordinates of the hit. This information, injected into the models via the positional loss, allows them to focus better on the desired areas of the images. In this sense, it seems that positional loss serves as a substitute for an attention mechanism or as an aid in learning positional information and may be an obstacle when the model learns this information well enough without this regularization. Despite promising fidelity of the energy distribution in the DS1 case, measured by the *Wass. Sum* metric, GANs struggle to capture the proper morphology, leaving also background artifacts.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

5.5 Diffusion experiments

For the diffusion model, we utilized a conditional U-Net architecture. The network processes inputs through four downsampling and upsampling stages, with feature channels scaling from 16 to 128. To incorporate particle kinematics conditioning, cross-attention mechanisms are integrated at specific resolution levels and in the middle block, alongside standard group normalization. All samples are obtained using 50 inference steps and $\eta = 0.7$.

Diffusion yields significantly better results than GANs with respect to the response-wise metrics. In contrast to the GAN case, the positional loss term provides no benefit for any of the tested datasets. We link this behavior to the presence of attention in the network architecture; however, explaining this phenomenon requires further investigation. Diffusion models also require two orders of magnitude more time for sample generation. The advantage of diffusion in terms of energy distribution comparisons with Wasserstein distances is not as clear as in similar calorimeter studies [22]. It seems that GANs better capture the distribution of the total response energy (measured by the *Total Wass.* metric) and the relation between the input particle energy and the total response energy (the *Joint Wass.* metric). This outcome is particularly evident for DS2.

However, diffusion models produce much better visual results, which is also reflected by a low *Perceptual Distance*. Still, in both cases, the background is poorly generated compared to the original responses. The distribution of the background, deposited by dispersed low-energy particles, is almost uniform, whereas in the original responses it is concentrated more in proximity to the center of the shower. Despite these drawbacks, diffusion exceeds GANs in the fidelity of responses represented by lower Wasserstein (*Wass. Sum*), perceptual, and positional distances.

5.6 Normalizing Flow experiments

Following the CaloINN [5] architecture, the complete transformation is constructed by stacking multiple blocks, each consisting of an ActNorm layer, a coupling layer, and a mixing layer. The second one splits the input vector into two parts: one half remains unchanged, while the other half is transformed using a spline transformation (Rational Quadratic).

NFs were the most challenging to train due to the high instability of the quadratic splines. Nevertheless, for DS2, we trained well-performing models, as seen in Tab. 2 and Fig. 3. No positive flow configuration was found for DS1; we observed high Wasserstein values and bad shower shapes. The main problem with the training stability of the NF models was their tendency to overfit along a few dimensions, making most of the response energy concentrated in a few pixels. It was possible to improve quality by adjusting the temperature of NF generation, but at the cost of a worse energy distribution in the samples, which eventually yielded blurred results.

For the well-performing DS2 case, NFs struggle to learn the global distribution of the detector outputs, as quantified by high *Total Wass.* and *Joint Wass.*

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

values. A key detail here is that the background and the cloud of low-energy cells are well generated, compared to diffusion. However, NF models have a tendency to soften the central peak of the energy shower. Fine-tuning the RQ spline parameters enabled a sharper central peak; yet, it introduced training instabilities, degraded the shower structure, and led to an overconcentration of energy within a few isolated pixels.

5.7 Approximation experiments

All "approx" models are compact models of decoder architecture built with fully connected layers. They utilize ReLU, Tanh, and Leaky ReLU activations.

This approach is naturally suitable for DS2, where only one energy peak is observed. Despite its much lower complexity compared to other models tested, it achieves low *Wass. Sum*, *Total Wass.*, and *Joint Wass.* values, comparable with GANs. They also provide second-best *MAE* values. Since this method positions the shower in the expected hit position computed from input parameters, here, the *Positional Distance* values are the lowest.

The use of this method for the multi-shower DS1 is not straightforward. The multi-shower responses, as well as single-shower ones, follow some distribution that describes what portion of energy is deposited in a given distance from the (presumed) hit location. This information can be encoded in the "approx" samples for DS1. The responses generated for DS1 are visually much different from the ones obtained for DS2 - both shown in Fig. 3, Approx row. As expected, the "approx" models exhibit higher *Perceptual Distance* than other leading architectures, a consequence of their specific generative design. This approach can be useful in situations where only some summary statistics of each detector response are saved for further simulation steps/analyses.

It is important to note that the relatively high generation time (i.e. compared to GANs) is a result of the two-step approach utilized in the response generation, with the second being performed on the CPU. The first stage - the response characteristic generation - uses GPU, like other methods. However, the actual image generation is performed iteratively on the CPU by sequentially assigning the predicted energy values to the target array, increasing the shower radius in each step, as in Fig. 2. It is assumed that proper acceleration on e.g. GPU should give a significant speed boost.

5.8 Approximation enhancement

For approximation enhancement experiments, we report only the time needed for image enhancement, i.e., excluding the time for approximate image generation (discussed in Section 5.7).

GAN enhancer experiments In principle, GAN-enhancers are very similar to the GAN models discussed in Section 5.4, but have a reduced number of intermediate channels.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

Based on the high value of *Wass Sum.* for DS1, it seems that the capacity or the convolutional architecture of the GAN is not sufficient to utilize the information encoded in "approx" samples discussed in Section 5.7. The sample images generated with this method differ greatly from the reference.

The GAN enhancer samples for the DS2 show rectangular artifacts around the shower center. However, they are visually better than the pure-GAN ones, as the enhanced samples always contain only one shower (as expected), in contrast to the pure-GAN-generated ones. This approach provides the lowest values of the *Total Wass.* and *MAE* metrics, and provides satisfactory results regarding *Perceptual Distance*. Importantly, only the "approx" models produce outputs with lower *Positional Distance* than this architecture.

Diffusion enhancer experiments Similarly to the GAN enhancer, the diffusion enhancer utilizes a scaled-down architecture featuring fewer layers, fewer channels, and a reduced number of attention heads - we prepared a U-Net with two upsampling and two downsampling blocks without attention. As before, the inference steps were fixed at 50 and η at 0.7. The diffusion enhancer surpasses other diffusion models in generation speed, giving satisfactory results in terms of the metrics tested.

For DS2, the diffusion approach did not leave rectangular patterns around the center, as shown in Fig. 3, Diff-Enc row. Moreover, the background and low-energy cloud of particles are of better quality than for diffusion, as also in the NFs case. For DS1, this approach much better utilizes the "approx" input compared to the GAN enhancer.

For both datasets, this approach slightly worsens the *Joint Wass.* and *Perceptual Distance* performances, while slightly improving the *Total Wass.* and preserving the *Wass. Sum*, compared to the pure-diffusion approach. For DS2, despite the initial intention of improving the energy distribution distances by this incorporation of low-Wasserstein "approx" samples, it is the *Positional Distance* that is visibly reduced. The better quality of diffusion-enhanced samples compared to GAN-enhanced samples is reflected by lower *Perceptual Distance* values.

6 Conclusions

Here, we investigated fast surrogates for FoCal simulation, targeting realistic shower images conditioned on particle kinematics. We utilized state-of-the-art generative frameworks such as GANs, NFs, and diffusion models, as well as hybrid approaches employing the approximation of shower position and shape with further deep-learning-based refinement. We also optionally assisted the generative process with information about the expected hit position in the form of a positional loss function.

Our core observation is the presence of a tradeoff between the simulation fidelity and time. The diffusion models are the most faithful among the tested frameworks, consequently providing samples with high perceptual similarity to

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

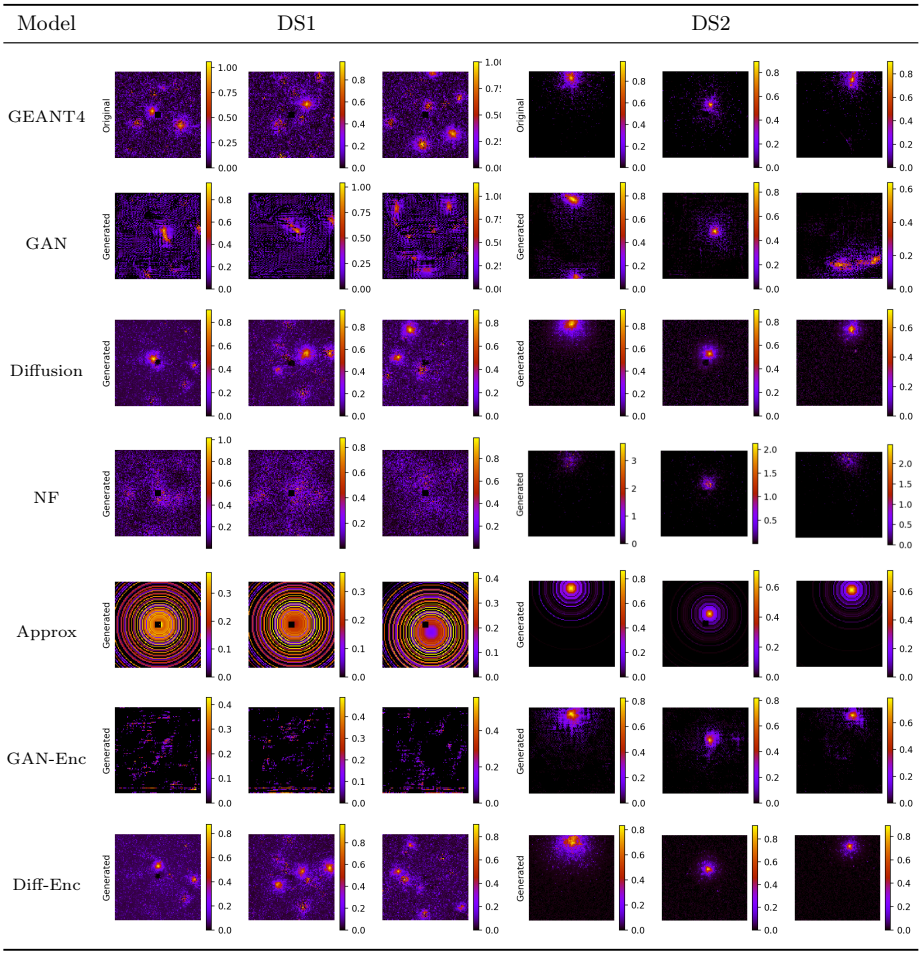

Figure 3: Example simulations generated by models discussed in this paper for DS1 and DS2. For each architecture, we used models with the lowest *Wass. Sum* metric value. The first row presents reference GEANT4 data.

the reference data. They also successfully capture the energy distribution of the responses. However, they are also the slowest. GANs, despite being orders of magnitude faster in inference, exhibit higher perceptual and positional errors, and provide visually worse samples. The NF training process is the most unstable and has not resulted in satisfactory models. The approximation pathway emerges as a strong baseline for geometry and calibration: with only tens of thousands of parameters, it achieves low positional errors and favorable global energy-response alignment, and, when paired with a lightweight enhancer, offers an attractive speed-quality tradeoff. These observations lead to the conclusion that the choice of models depends on the particular application. When fidelity

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

to shower morphology and energy profiles is paramount, diffusion is preferred, accepting higher latency at fixed step counts. For large-scale production where throughput is the bottleneck, approximation plus a tiny enhancer achieves usable realism with minimal computational demands.

Future work will prioritize the investigation of hybrid models utilizing the "approx" approach. In particular, when combined with diffusion models, this method has demonstrated the greatest potential in terms of data fidelity and model compactness. The faster GAN-based solution requires further optimization towards better quality (especially visual) of the responses. A key objective will be to optimize the implementation of "approx" models to significantly improve their computational speed.

Acknowledgments. We would like to thank Ionut Arsene, PhD (University of Oslo), Hadi Hassan, PhD (University of Tsukuba), Professor Jacek Kitowski (AGH University of Krakow), and Professor Jacek Otwinowski (Institute of Nuclear Physics PAS in Krakow) for their support. This work is co-financed and in part supported by the Ministry of Science and Higher Education (Agreements No. 2022/WK/01 and 2023/WK/07) by the program entitled "PMW" and by the Ministry funds assigned to AGH University of Krakow. We gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grants no. PLG/2024/017264 and PLG/2025/018322. We sincerely thank the reviewers for their insightful comments and constructive suggestions that helped improve our work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Geant4—a simulation toolkit. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506**(3), 250–303 (2003). https://doi.org/10.1016/S0168-9002(03)01368-8
2. Technical Design Report of the ALICE Forward Calorimeter (FoCal) (2 2024)
3. Buss, T., Gaede, F., Kasieczka, G., Krause, C., Shih, D.: Convolutional l2lflows: generating accurate showers in highly granular calorimeters using convolutional normalizing flows. Journal of Instrumentation **19**(09), P09003 (sep 2024). https://doi.org/10.1088/1748-0221/19/09/P09003
4. Będkowski, P., Dubiński, J., Szatkowski, F., Deja, K., Rokita, P., Trzcinski, T.: Expertsim: Fast particle detector simulation using mixture-of-generative-experts (08 2025). https://doi.org/10.48550/arXiv.2508.20991
5. Ernst, F., Favaro, L., Krause, C., Plehn, T., Shih, D.: Normalizing flows for high-dimensional detector simulations (2025)
6. Giannelli, M.F., Zhang, R.: Caloshowergan, a generative adversarial network model for fast calorimeter shower simulation. The European Physical Journal Plus **139**(7), 597 (Jul 2024). https://doi.org/10.1140/epjp/s13360-024-05397-4
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (06 2020). https://doi.org/10.48550/arXiv.2006.11239

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21

9. Kingma, D., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions (07 2018). https://doi.org/10.48550/arXiv.1807.03039
10. Kita, M., Dubiński, J., Rokita, P., Deja, K.: Generative diffusion models for fast simulations of particle collisions at cern (2024)
11. Kocot, M., Misan, K., Grzanka, L., Avati, V., Bossini, E., Minafra, N.: Using deep neural networks to improve the precision of fast-sampled particle timing detectors. Computer Science **25**(1) (Mar 2024). https://doi.org/10.7494/csci.2024.25.1.5784
12. Krause, C., Giannelli, M.F., Kasieczka, G., et al.: Calochallenge 2022: A community challenge for fast calorimeter simulation (2024)
13. Krause, C., Shih, D.: Accelerating accurate simulations of calorimeter showers with normalizing flows and probability density distillation. Phys. Rev. D **107**, 113004 (Jun 2023). https://doi.org/10.1103/PhysRevD.107.113004
14. Krause, C., Shih, D.: Fast and accurate simulations of calorimeter showers with normalizing flows. Phys. Rev. D **107**, 113003 (Jun 2023). https://doi.org/10.1103/PhysRevD.107.113003
15. Majerz, E., Dzwinel, W., Kitowski, J.: Inverse Autoregressive Flows for Zero Degree Calorimeter fast simulation. In: 39th Annual Conference on Neural Information Processing Systems: Includes Machine Learning and the Physical Sciences (ML4PS) (12 2025)
16. de Oliveira, L., Paganini, M., Nachman, B.: Learning particle physics by example: Location-aware generative adversarial networks for physics synthesis. Computing and Software for Big Science **1**(1) (Sep 2017). https://doi.org/10.1007/s41781-017-0004-6
17. Paganini, M., de Oliveira, L., Nachman, B.: Calogan: Simulating 3d high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. Phys. Rev. D **97**, 014021 (Jan 2018). https://doi.org/10.1103/PhysRevD.97.014021
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01042
19. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2021)
20. Webber, G., Reader, A.J.: Diffusion models for medical image reconstruction. BJR|Artificial Intelligence **1**(1), ubae013 (08 2024). https://doi.org/10.1093/bjrai/ubae013
21. Wojnar, M.: Even faster simulations with flow matching: A study of zero degree calorimeter responses. Computer Physics Communications **319**, 109936 (2026). https://doi.org/10.1016/j.cpc.2025.109936
22. Wojnar, M., Majerz, E., Dzwinel, W.: Fast simulation of the Zero Degree Calorimeter responses with generative neural networks. Computing and Software for Big Science **9**(1), 1 (2025). https://doi.org/10.1007/s41781-025-00130-x
23. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_21