Generative Neural Networks for Fast and Accurate Zero Degree Calorimeter Simulation

Maksymilian Wojnar, AGH University of Krakow
maksymilian.wojnar@agh.edu.pl

ML in PL CONFERENCE 2024

Abstract

The integration of generative neural networks into high-energy physics simulations is rapidly transforming the field, offering unprecedented efficiency and accuracy. A prominent application is the simulation of the Zero Degree Calorimeter (ZDC) in the ALICE experiment at CERN (Fig. 1 and 2). Traditionally, these simulations have relied on Monte Carlo methods, which, while highly accurate, are computationally intensive and time-consuming (Fig. 3). By employing generative networks as surrogate models, we achieve a significant reduction in computational burden while maintaining high accuracy. In this work, we utilize the latest advancements in generative neural networks, specifically focusing on flow matching (FM) and models based on vector quantization (VQ), to simulate the ZDC neutron detector. These state-of-the-art architectures enable the generation of high-fidelity data that closely mirrors real experimental results. We explore and compare the performance of the generative frameworks against established simulation methods. Our findings underscore the effectiveness of generative neural networks in providing fast yet accurate simulations, making them a valuable tool in the high-energy physics community. This poster is an extension of the work presented in [3].

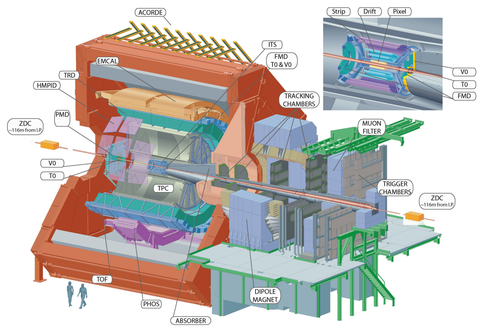

Figure 1. Schematic diagram of the ALICE central barrel detection systems in 2017 [1].

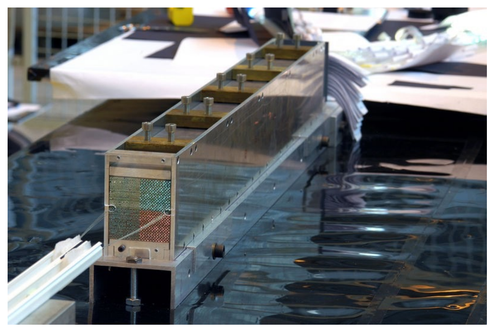

Figure 2. The ZDC neuron detector [2].

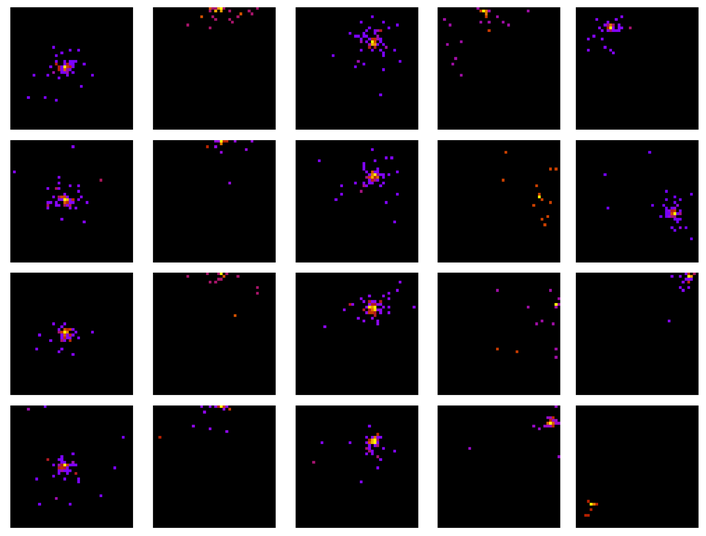

Figure 3. Example ZDC responses generated with a Mone Carlo GEANT toolkit.

Main results

This study presents a significant advancement in SOTA for fast ZDC simulation, with particular improvements in sampling speed and fidelity as measured by the Wasserstein distance and mean absolute error (MAE), as shown in Tab. 1. Our proposed latent FM achieves an impressive sampling speed of 0.008 ms per sample, outperforming all current models. In terms of fidelity, FM closely matches the SOTA Wasserstein score (1.27 vs. 1.2) but dramatically reduces generation time from 120 ms to 0.37 ms. Additionally, we optimize the previous VQ-GAN architecture, reducing its Wasserstein score from 4.58 to 2.11, with a modest increase in generation time.

The efficiency of the latent FM model establishes a new benchmark in sampling speed. Additionally, the compactness of the models further supports this efficiency, with the FM and latent FM models containing only 77k and 160k parameters, respectively.

Table 1. Performance comparison of generative frameworks.

Model	Time ↓ [ms]	Wasserstein ↓	MAE
Original data	–	0.53	16.41
FM	0.37	**1.27**	**16.99**
Latent FM	**0.008**	2.11	22.32
VQ-GAN	0.26	2.01	20.33

The metrics used in this work are defined as:

$$\text{Wasserstein-1}(w, \hat{w}) = \frac{1}{5}\sum_{i=1}^{5}\int_0^1 \left|F_{w_i}^{-1}(z) - F_{\hat{w}_i}^{-1}(z)\right| dz,$$

$$\text{MAE}(w, \hat{w}) = \frac{1}{n}\sum_{k=0}^{n}\frac{1}{5}\sum_{i=1}^{5}|w_i^k - \hat{w}_i^k|,$$

where F_q^{-1} is the inverse cumulative distribution function of the distribution q, w_i denotes the distribution of the i-th channel, n refers to the number of evaluated examples and w_j^k represents the value of the i-th channel of the k-th response.

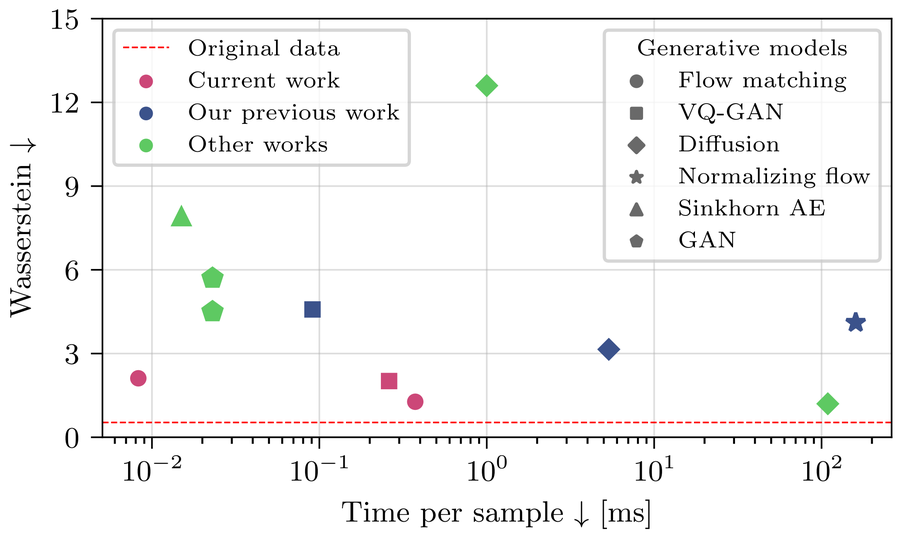

Figure 4. Comparison of Wasserstein distance and sampling time between this work, the previous work, and other methods across various generative models.

Vector quantization

Based on our previous work, we implement an improved VQ-GAN model [4], addressing limitations in the reconstruction ability of the prior model's autoencoder. To enhance this capability, we focus on the initial training phase (Fig. 5), reducing the downsize factor from x8 to x4. Following [5], we employ a larger codebook of size 512 with a lower vector dimensionality of 8, which decreases the model's parameter count (from 1M to 70k) compared to [3]. This model leverages an EMA codebook update and integrates VQ, MSE, perceptual, and adversarial loss terms (details provided in the "Flow matching" section), achieving a Wasserstein metric score of 3.09.

To improve speed, we make additional adjustments, as the learnable prior (GPT) now generates a substantially larger number of tokens. This includes using a smaller model, mixed precision training and inference, and the biggest possible batch size (Tab. 2). We apply both temperature and top-k sampling in GPT (Fig. 7) and use the same VQ-VAE for conditional variables as in [3] for generation (Fig. 6).

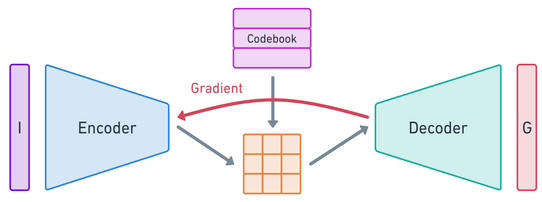

Figure 5. VQ-VAE design.

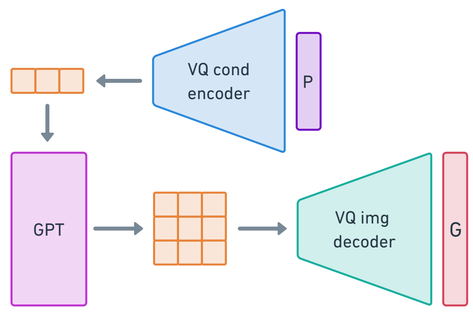

Figure 6. VQ-GAN generation procedure.

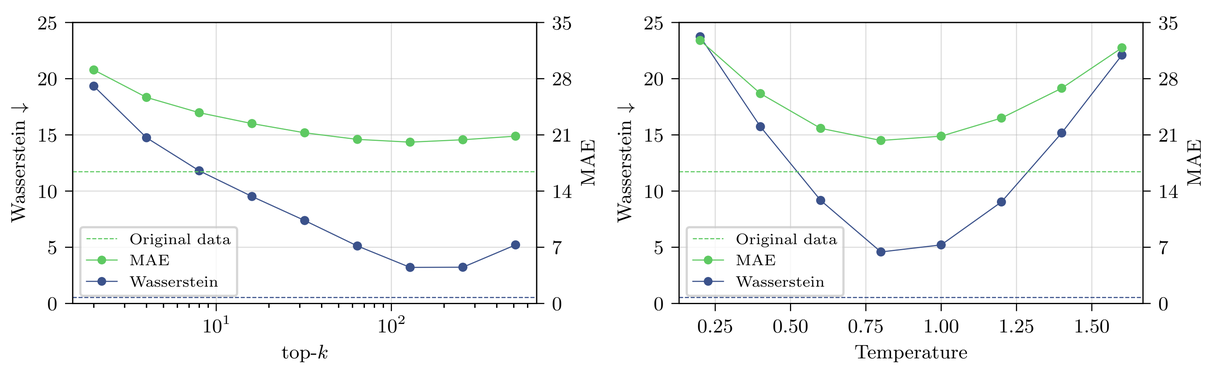

Figure 7. VQ-GAN performance with various sampling techniques.

Table 2. Improvements applied to speed up GPT.

Improvement	Time ↓ [ms]	Relative change
Previous model	0.09	–
Greater num of steps (11^2)	0.80	+777%
Smaller model (1M)	0.60	-25%
Mixed precision (F16)	0.45	-25%
Bigger batch size (16k)	**0.26**	-41%

Flow matching

FM is a family of generative models that facilitate the transition from the noise x_0 to data x_1 through a linear interpolation process:

$$x_t = (1-t)x_0 + tx_1,$$

where $t \in [0,1]$ is the interpolation time [6]. The neural network, based on a UNet architecture (Fig. 8), is trained to learn the normalized transition velocity:

$$v_\theta(x_t) = x_1 - x_0,$$

which then can be used to generate samples incrementally, applying the Euler method as:

$$x_{t+1} = x_t + \Delta t \cdot v_\theta(x_t).$$

The associated loss function is defined as:

$$\mathcal{L}(\theta; x_t, v_t) = |v_\theta(x_t) - v_t|^2.$$

The network architecture follows the Stable Diffusion [7] autoencoder and includes attention, enabling conditioning on the input vector. For this study, we implement a compact UNet with 77k parameters and a linear noise schedule. We apply additional improvements to speed up the model (Tab. 3) and we adjust the number of steps (Fig. 10).

The VAE in the latent FM model (Fig. 9), based on the [7] architecture with a size of 60k, ensures high reconstruction quality and stable training via a loss function with gradient normalization wrt. the input:

$$\mathcal{L}(\theta; x) = \frac{\mathcal{L}_{AE}}{|\nabla_x \mathcal{L}_{AE}|} + \frac{\mathcal{L}_{perc}}{|\nabla_x \mathcal{L}_{perc}|} + \frac{\mathcal{L}_{adv}}{|\nabla_x \mathcal{L}_{adv}|},$$

where $\mathcal{L}_{AE}$ is the autoencoder loss, combining l_2 reconstruction loss and a regularization term, the perceptual loss $\mathcal{L}_{perc}$ is based on LPIPS, while the adversarial loss $\mathcal{L}_{adv}$ employs a pre-trained VGG16 network with linear feature extraction and hinge loss. This VAE achieves a Wasserstein metric score of 0.95.

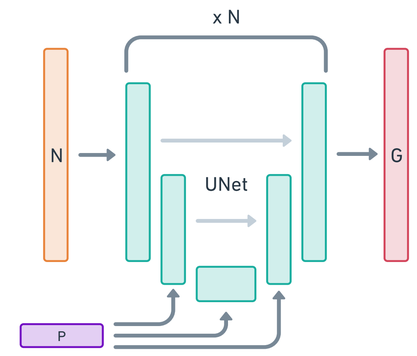

Figure 8. UNet-based FM design.

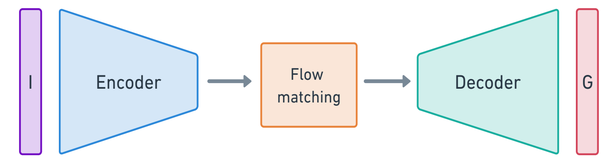

Figure 9. Latent FM design.

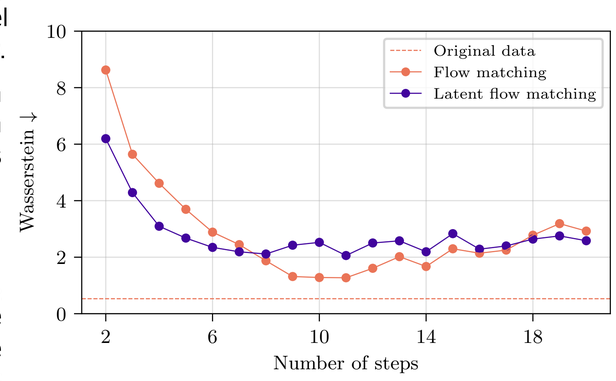

Figure 10. FM and latent FM performance depending on the number of steps.

Table 3. Improvements applied to speed up FM.

Improvement	Time ↓ [ms]	Relative change
Previous model	11.80	–
Fewer num of steps (11)	2.60	-78%
Smaller model (77k)	0.62	-76%
Mixed precision (F16)	0.46	-27%
Bigger batch size (8k)	**0.37**	-18%
Latent space model	0.026	-93%
Bigger batch size (16k)	**0.008**	-63%

References

[1] R. Schicker. (2017). Overview of ALICE results in pp, pA and AA collisions. *EPJ Web Conf*.
[2] ALICE Zero Degree Calorimeter (ZDC), General Pictures. (2003). *ALICE Collection*. https://cds.cern.ch/record/630193.
[3] M. Wojnar. (2024). Applying generative neural networks for fast simulations of the ALICE (CERN) experiment. *arXiv [Physics.Ins-Det]*. https://arxiv.org/abs/2407.16704.
[4] P. Esser, R. Rombach, and B. Ommer. (2021). Taming Transformers for High-Resolution Image Synthesis. In: *2021IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR46437.2021.01268.
[5] P. Sun et al. (2024). Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *arXiv [cs.CV]*. https://arxiv.org/abs/2406.06525.
[6] Y. Lipman et al. (2023). Flow Matching for Generative Modeling. *arXiv [cs.LG]*. https://arxiv.org/abs/2210.02747.
[7] R. Rombach et al. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv [cs.CV]*. https://arxiv.org/abs/2112.10752

We would like to thank Emilia Majerz and Professor Witold Dzwinel from AGH University of Krakow. This work is in part supported by the Ministry of Science and Higher Education (Agreement Nr 2023/WK/07) by the program entitled "PMW" and by the Ministry funds assigned to AGH University in Krakow. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017264.