Denoising Diffusion to Accelerate Detector Simulation

Project Close Out

Principal Investigator: Kevin Pedro (FNAL)
Postdoc : Oz Amram (FNAL)

The CMS detector simulation, which uses Geant4, consumed 40% of grid CPU at the beginning of Run 2. In the HL-LHC era, the CPU time to simulate an event will increase by a factor of 3 or more, because of the more complex detector geometry and the more detailed physics models needed to reproduce the precise measurements of the upgraded detector. Computing resources available for simulation are also expected to be constrained in the HL-LHC era, due to the increased needs of reconstruction. CMS therefore needs a fast and accurate simulation to achieve its physics goals in the HL-LHC era.

Recent years have seen the rise of generative AI models, which have demonstrated impressive capabilities at image and text generation. In particular, a new class of models, based on diffusion, have become popular in the last few years because of their impressive output quality and stable training.

The goal of the project was to develop AI models based on a denoising diffusion approach to simulate complex calorimeter showers, which are the most costly step of the simulation chain. The start of the project coincided with a larger community effort to organize and compare different approaches to fast calorimeter simulation. This took the form of a community data-challenge, CaloChallenge, with public datasets. These datasets allowed easy R&D on different generative AI techniques and comparisons to other approaches.

The first year of this project therefore focused on developing our denoising diffusion model, called CaloDiffusion, on these datasets. At the conclusion of the first year of the project, we published a paper in PRD [1] detailing our model and its performance, and submitted our results to the CaloChallenge [2]. Comparisons to other approaches showed that CaloDiffusion was able to generate the highest quality showers for all datasets of the CaloChallenge, and were found to be nearly indistinguishable from Geant4 using several different metrics (log posterior probabilities, distances via integral probability metrics, and neural network classifier scores), which are becoming standardized in the community. However, it was found that, while CaloDiffusion inference running on a GPU with large batches could generate showers up to 1000x faster than Geant, it was not as fast as other generative models.

The second year of the project was focused on speeding up the model and applying to simulations of CMS High Granularity Calorimeter (HGCal). Using various technical improvements to the diffusion model, as well as separating key shower properties into a second, light-weight model, we were able to achieve an ~8x speedup of the CaloDiffusion model while preserving or in some cases improving the modeling quality. Various machine learning techniques to 'distill' the multi-step diffusion process into fewer steps were also tried. Some of these methods additionally incorporated an approximate starting point for the shower (a proxy for classical GFlash-based FastSim) were also experimented. These approaches showed promising results but further exploration of the model space would be needed to settle on a final approach.

The other major effort in the 2nd year was to apply the model to simulate the CMS HGCal. This required handling the complex geometry of the HGCal, which features both highly granular hexagonal cells and has significant irregularity. In anticipation of these challenges, we had developed the Geometry Latent Mapping (GLaM) approach to handle the irregular geometries in the CaloChallenge datasets, which had significantly reduced granularity compared to HGCal. To apply the method to HGCal, it had to be significantly adapted and improved to cope with the increased complexity. Preliminary results of applying CaloDiffusion to the HGCal were shown at Computing in High Energy Physics (CHEP) in Fall 2024 [3]. We were able to show that the model can successfully match Geant4 on many important features of the shower, including the total energy, the energy distribution among the layers, and the angular shape. However, features related to the sparsity and radial distribution of the shower, both of which are sensitive to GLaM choices, still need to be improved.

Overall, the project was able to successfully demonstrate that high quality calorimeter simulations can be produced with an AI generative model, significantly raising the bar of generative quality as compared to prior works. The project also made significant steps towards applying these methods to the CMS HGCal. Though the complex geometry of the HGCal remains a challenge, and final results have not yet been achieved, no show-stoppers have been encountered so far. The results from this project have been presented several times in the broader HEP-ML community [3-7] and regular updates were given within the CMS ML4Sim group [7-9]. Work is ongoing in CMS to finalize and publish several HGCal datasets that can be used for training and evaluation both within the collaboration and in the broader community, following the CaloChallenge example.

Work on the project will continue under new funding sources. We hope to fix the mismodeled features in the HGCal simulation with improvements to the GLaM approach and then begin to integrate the model into CMSSW so it can be tested in fast simulation workflows. We plan to publish a paper with those results under the CMS MLG group.

References:
[1] O. Amram and K. Pedro. “*Denoising diffusion models with geometry adaptation for high fidelity calorimeter simulation*”. Phys.Rev.D 108 (2023) 7, arXiv:2308.03876

[2] Krause et al. “*CaloChallenge 2022: A Community Challenge for Fast Calorimeter Simulation*”. Submitted to ROPP. arXiv:2410.21611

[3] K. Pedro “Simulating the CMS High Granularity Calorimeter with ML” CHEP, Oct. 2024 https://indico.cern.ch/event/1338689/contributions/6015962/

[4] K. Pedro “CaloDiffusion with GLaM for High Fidelity Calorimeter Simulation” ML4Jets Nov. 2023 https://indico.cern.ch/event/1253794/contributions/5588571/

[5] O. Amram "Fast and Accurate Calorimeter Simulation with Diffusion Models", CHEP 2023 https://indico.jlab.org/event/459/contributions/11736/

[6] O. Amram “Denoising Diffusion Models for High Fidelity Calorimeter Simulation", CaloChallenge Workshop 2023 https://agenda.infn.it/event/34036/contributions/200894/

[7] O. Amram “Fast Simulation of Particle Physics Calorimeters” FastML 2024 https://indico.cern.ch/event/1387540/contributions/6153612/

[8] O. Amram “CaloDiffusion for HGCal” CMS ML4Sim Meeting (Oct. 2nd, 2024) https://indico.cern.ch/event/1462757/#11-hgcalodiffusion

[9] O. Amram “CaloDiffusion” CMS ML4Sim Meeting (Jan. 24, 2024) https://indico.cern.ch/event/1373532/