CaloHadronicVAE: Fast Point Cloud Generation via Latent Diffusion

Martina Mozzanica
Universität Hamburg
martina.mozzanica@uni-hamburg.de

September 1, 2026

Introduction

Simulation is the bridge between theoretical predictions and experimental measurements: to exploit the data collected at current and future colliders, the simulated samples must match the recorded data in statistics while numbering on the order of ten times as many. The most expensive step of the simulation chain is the first-principles modeling of the detector response with Geant4, whose projected cost for high-granularity detectors grows faster than the available resources.

This motivates fast simulation: surrogate models that reproduce the detector response at a fraction of the cost. More recently, deep generative models have been used to learn the full, high-dimensional shower distribution from Geant4 data: GANs, variational autoencoders, normalizing flows, and diffusion models have all been explored. Calorimeter showers are the natural target being the dominant simulation cost. My project develops such a model.

Problem setup

A calorimeter absorbs incident particles and produces signals proportional to their energy. I focus on the highly granular sampling calorimeters of the International Large Detector (ILD) [1] concept for the International Linear Collider. I target *hadronic* showers, the harder case: unlike electromagnetic showers (dominated by bremsstrahlung and pair production), they involve complex, non-perturbative physics and a mixed electromagnetic/hadronic structure.

A shower can be represented as a fixed 3D voxel grid or as a *point cloud*—a variable-length, permutation-invariant set of energy deposits. Since only about 4% of the cells carry signal, the point-cloud representation is far more economical: it stores only the non-zero deposits, is independent of the cell geometry, and allows several points per cell.

The generator produces 5D point clouds with coordinates $[x, y, z, \text{energy}, \text{time}]$, conditioned on the incident particle energy, and is trained on 10–90 GeV π^+ showers. The inclusion of a per-hit *time* feature is, to our knowledge, a first for highly granular calorimeters applied to hadronic showers.

The model develops CaloHadronic [2] further and combines three components:

- a *continuous normalizing flow* that generates the layer-wise quantities—the number of points per layer and the minimum hit time per layer;
- a *variational autoencoder (VAE)* that compresses each shower into a compact latent vector and decodes it back to a point cloud;
- a *diffusion model* that learns to generate showers directly in this latent space.

The layer-wise quantities from the flow, together with the incident energy, condition both the diffusion model and the VAE decoder.

Results

Running a diffusion process directly on a full point cloud of $\mathcal{O}(10^3)$ points is expensive.

Instead, each shower is encoded into a much smaller latent vector, the diffusion runs there, and the decoder decompresses back to the full cloud. To keep every stage cheap while preserving accuracy, attention is restricted by a *sparse mask* that lets each element attend only to neighbors within a few layers. This latent-diffusion design is what makes the approach significantly cheaper than direct point-cloud generation: at matched NFE it is consistently faster than baselines like AllShowers [3] (fig. 1), and the gap widens with both the number of sampling steps and the incident energy.

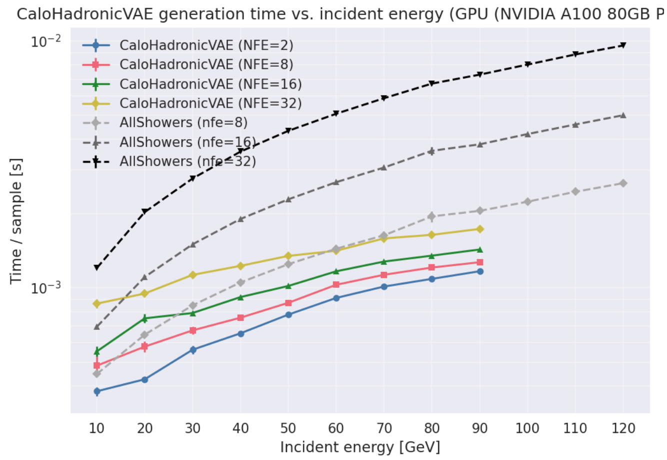

Figure 1: Per-sample generation time versus incident energy on a single NVIDIA A100 80 GB GPU at batch size 128. Solid lines: the latent-diffusion CaloHadronicVAE for NFE $\in \{2, 8, 16, 32\}$; dashed lines: the direct point-cloud baseline AllShowers at nfe $\in \{8, 16, 32\}$.

Bonus: the Step2Point library.
To avoid grid artifacts, point-cloud models are best trained at a resolution finer than the calorimeter cells. Step2Point [5] is a companion library that turns detailed Geant4 step-level deposits into compact point clouds for ML-based fast simulation, designed to work across different detectors. Clustering the steps with HDBSCAN is effective and independent of the cell geometry, so it applies to arbitrary cell shapes. This was validated on photon showers in ILD together with CaloClouds3 [4].

Conclusion
Point-cloud models are currently the fastest option for ML-based calorimeter simulation, and latent diffusion adds a compact, scalable shower representation on top. CaloHadronicVAE reproduces hadronic π^+ showers—including timing information—while scaling favorably with energy and point multiplicity, and Step2Point provides the detector-independent infrastructure to feed it. Performance can still be improved, and two publications are in preparation.

References

[1] ILD Collaboration, *International Large Detector: Interim Design Report*, arXiv:2003.01116 (2020).

[2] T. Buss, F. Gaede, G. Kasieczka, A. Korol, K. Krüger, P. McKeown and M. Mozzanica, *Calo-Hadronic: a diffusion model for the generation of hadronic showers*, JINST **21** (2026) P01042, arXiv:2506.21720.

[3] T. Buss, H. Day-Hall, F. Gaede, G. Kasieczka and K. Krüger, *AllShowers: One model for all calorimeter showers*, arXiv:2601.11716 (2026).

[4] T. Buss, H. Day-Hall, F. Gaede, G. Kasieczka, K. Krüger, A. Korol, T. Madlener, P. McKeown, M. Mozzanica and L. Valente, *CaloClouds3: Ultra-Fast Geometry-Independent Highly-Granular Calorimeter Simulation*, JINST **21** (2026) P03018, arXiv:2511.01460.

[5] A. Zaborowska and P. McKeown, *The step2point dataset*, arXiv:2509.22340 (2025).