GPU-Accelerated Machine Learning

In Practice & Under the Hood

Spring 2027 · UW ECE Certificate in GPU-Accelerated Computing & Visualization

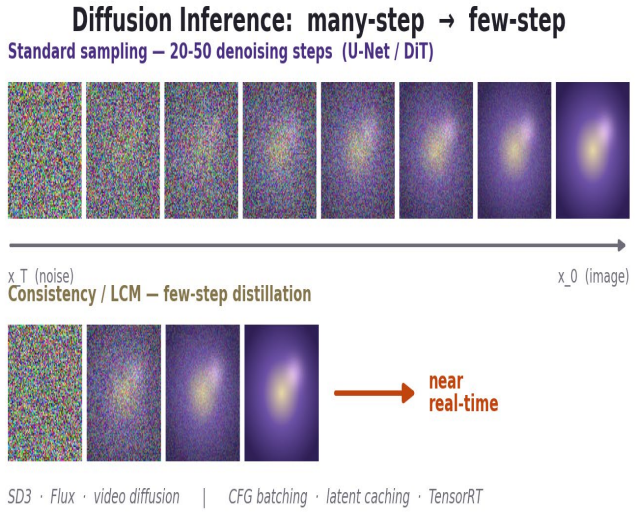
- **Python-centric GPU programming —** block-level kernels in Triton and the CuTe DSL (CUTLASS 4), autotuning
- **PyTorch integration & profiling —** custom operators (torch.library), torch.compile / TorchInductor fusion, CUDA Graphs, and trace-driven profiling (Nsight, Kineto / HTA).
- **Advanced attention & LLM serving —** FlashAttention v3/v4 (warp specialization, FP8 / FP4), TensorRT-LLM & Dynamo, speculative decoding, prefix caching.
- **Quantization —** FP8 training (Transformer Engine), GPTQ / AWQ / Marlin W4A16;
- **Diffusion —**diffusion for image & video synthesis (U-Net vs. DiT, consistency models, SD3 / Flux).
- **3D vision/rendering/understanding —** point-cloud learning (PointNet++ → Point Transformer V3), neural fields (Instant-NGP, 3D Gaussian Splatting), and tracing production CUDA from Python → PTX → SASS.

Toolchain: Triton · CuTe DSL · PyTorch / torch.compile · TensorRT-LLM · Nsight · Hopper & Blackwell

Python + CUDA C++ · kernel-optimization intensive

Authoring high-performance kernels in Python — compiled down to SASS

authored from PyTorch / torch.compile



Fewer denoising steps → faster image & video generation