

# Shlok Limbhare

Machine Learning — GPU performance

[Shlokvfx2003@gmail.com](mailto:Shlokvfx2003@gmail.com) — +91-86-988-20-100

Nashik, MH — [GitHub: ShlokVFX](#) — [LinkedIn: ShlokLimbhare](#)

## Projects

### 100 Days of CUDA: Optimized GPU Kernel Programming

[GitHub](#)

- Documented a structured 100-day learning journey in CUDA, covering core concepts, memory hierarchy, and kernel optimizations.
- Developed and profiled high-performance CUDA kernels with Streams, Shared Memory Tiling, Unified Memory, and Loop Unrolling to enhance execution efficiency.
- Implemented Tensor Core acceleration with WMMA, explored fused operations in Triton, and optimized cuBLAS-based linear algebra routines.
- Designed custom GPU-accelerated implementations for activation functions (Softmax, ReLU) and performance-optimized numerical kernels.
- Leveraged CUDA Graphs and `torch.compile` for reduced computational overhead in deep learning pipelines.
- Achieved **3x–5x speed improvements** in various CUDA-based workloads, demonstrating expertise in multi-GPU computing and performance optimization.

### AMD MI300X GPU Kernels: FP8 MatMul, MoE, MLA Decode

[GitHub](#)

Contributor — *HIP, FP8 Kernels, MoE Inference, Popcorn Eval*

- Developed HIP-based double-buffered matrix multiplication kernels supporting FP8 blockwise computation using MFMA intrinsics on AMD MI300X.
- Built optimized inference modules for Mixture-of-Experts (MoE) and MLA (Multi-head Latent Attention) decoding targeting large model inference.
- Benchmarked performance using Popcorn's eval harness on MI300X hardware to validate throughput, latency, and scaling behavior.
- Tuned shared memory usage, vectorized tile access, and kernel launch configurations to match AMD architecture design.
- Demonstrated expert-level understanding of ROCm, FP8 formats, shared memory pipelining, and GPU-specific compiler flags for HPC workloads.

## Technical Skills

**GPU Programming:** CUDA, HIP, Triton, Nsight, WMMA, MFMA      **Frameworks:** PyTorch, TensorFlow, Triton, TorchScript

**Performance Tools:** cuBLAS, cuDNN, CUDA Graphs, Popcorn, Perfetto      **Languages:** Python, C++, Docker, Git, CMake

**Model Types:** MoE, LLM Inference, MLA Decode, Transformer Ops      **Platforms:** AMD ROCm (MI300X), NVIDIA RTX, Linux HPC Clusters

## Professional Experience

### Houdini FX Artist

2021 – 2024

*DNEG, Mumbai, MH*

[Showreel](#)

- Created high-quality dynamic VFX simulations including **smoke, water, debris, and sand** using Houdini.
- Collaborated with the lighting and compositing teams to ensure seamless integration of FX into final shots.
- Optimized simulation setups for performance and memory usage while maintaining visual fidelity.
- Delivered procedural FX for both feature films and episodic content under tight deadlines.
- Used VEX and Python scripting to build reusable digital assets and tools for the FX pipeline.

## Education

**B.Sc. Computer Science**  
2023 – 2027

*BITS Pilani, Pilani, RJ*

- Specialized in Artificial Intelligence, Machine Learning, and High-Performance Computing.
- Completed coursework in Data Structures, Algorithms, Operating Systems, and Distributed Systems.
- Led academic projects involving GPU-accelerated deep learning pipelines and compiler-aware optimizations.

## About Me

Prev Hollywood Artist Professional turned ML developer. Please give out