Roadmap

Features considered through Feb 2025

Features

- Support multi-gpu execution
- Support determinism for cross batch
- Support input volume >2^31 and tensor dimensions >32 bit
- Support more flexible ControlNet
- Support fine-grained kernel selection for better debuggability
- Provide load-time memory size adjustment
- HFC feature upgrade

Performance

- FP8 and FP4 quantization for ControlNet and Lora
- Multi-Lora optimizations
- MHA optimizations

User + Dev Experience

- Expose more debug information in TensorRT engines
- Improve documentation on framework -> TensorRT workflows.

Feedback? Feature requests? Raise an issue!