Luma

Luma AI's Systems Challenges

To build multimodal general intelligence that can understand, generate, and operate in the physical world

Challenge #1

Training

- Mixture of Experts
  - rapid iteration on experiments
  - novel architectures
  - custom routing kernels
  - efficient comms & compute overlap
- Long context training
  - memory spikes for vae & loss
  - sparse video attention
  - ring attention w/ symmetric memory
- Reinforcement learning
  - custom stack from scratch
  - heterogeneous workloads
    - VAE, denoisers, loras, controlnets, text

Challenge #2

Inference

- Hardware heterogeneity
  - AMD & NV
- Multiple clusters
  - >10k across different clusters
- auto-scaling / load balancing
- jit compilation / warmup
  - torch.compile
  - custom (cuda|hip)graph capture
- checkpoint loading
- custom kernels
  - CUDA/PTX, Triton, CuTeDSL, FlyDSL, HipKittens
- Symmetric memory, custom FA4