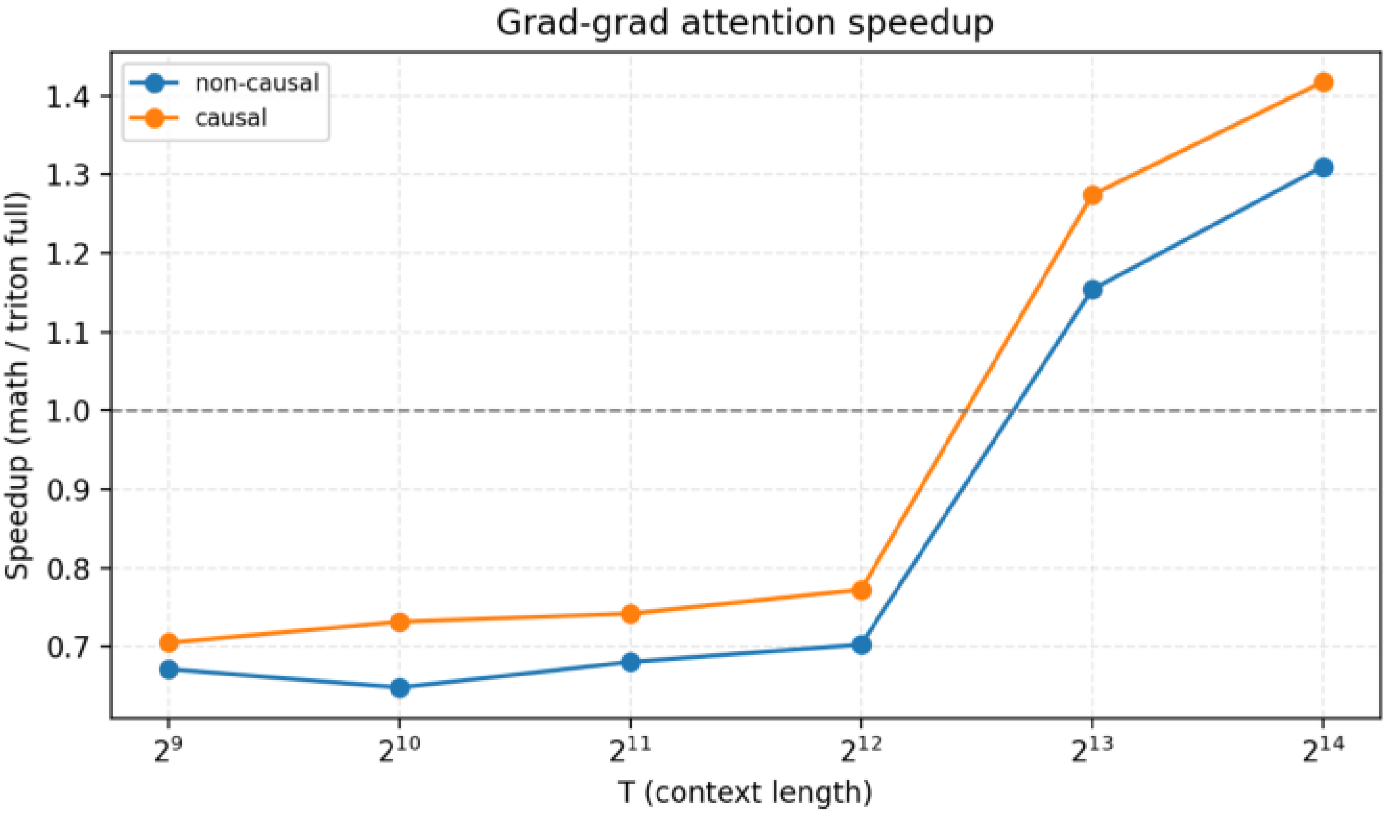
Grad-grad-safe Triton Attention for TTT-E2E

H100 benchmark: fused softmax-backward + block-sparse causal + custom dV/dQ/dK

Repo: https://github.com/ry2009/TestTimeIdeas



Key results (grad-grad timing):

- Non-causal T=16k: 11.466 ms vs 15.023 ms (1.31x)
- Causal T=16k: 13.649 ms vs 19.350 ms (1.42x)
- Crossover ~8k (launch overhead dominates below).

Correctness spot-check (b=1,h=1,t=64,d=64):

- forward max abs err: 0.0
- grad max abs err: ~6e-7
- grad-grad max abs err: 0.0

Repro:

```
python tests/bench_gradgrad_sweep.py --b 1 --h 1 --d 64 --dtype fp16 --iters 8 --warmup 2 --repeats 4
  --bwd_mode save_p_triton_full --ts 512,1024,2048,4096,8192,16384 --out artifacts/gradgrad_sweep.cs
```