? Gemma4 MTP Performance

> **SYSTEM**
> **CPU:** Ryzen 5 5500.
> **RAM:** 32G DDR4 3200.
> **GPU:** RTX 3060 12G VRAM.

With MTP: Sometime the tps hit 147 token/s.
Without MTP: Hit 32 token/s.

But MTP is only hit 147 token/s at start. After few seconds, the tps is only 29 token/s.
I think the hardware have no enough VRAM. If all the layers is fully loaded, the TPS can be keep at 147 token/s.

Revision #3
Created 2026-05-28 18:25:04 UTC by sandichhuu
Updated 2026-05-31 13:39:59 UTC by sandichhuu