

? 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.



```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
- Ability to add more LoRA items.
- Each LoRA has: name, weight, remove button.
- On generate: merge <name:weight> into the prompt (but don't show it in the prompt field).

Code Rules:
1. No comments in code.
2. Do not run any python files (owner will review and run manually).
3. Use mcp-workspace if needed.

Next Step: I need to read implement_plan.md to get more details.

read implement

" Working...

W:\repos\contribute\ComfyUI-Gradio (main)
14.2k 149 1.7%/262k (auto)
MCP: 0/4 servers TPS: 109.9 tok/s

(chenjia.org) think-code • high
sandichhuu (6 hours ago) Ln 19, Col 24 Spaces: 4 UTF-8 CRLF Markdown Pi No Environment Go Live Continue (NE)
```

Revision #3

Created 2026-05-28 18:25:04 UTC by sandichhuu

Updated 2026-05-31 13:39:59 UTC by sandichhuu