□□ Google Turboquant + llama.cpp

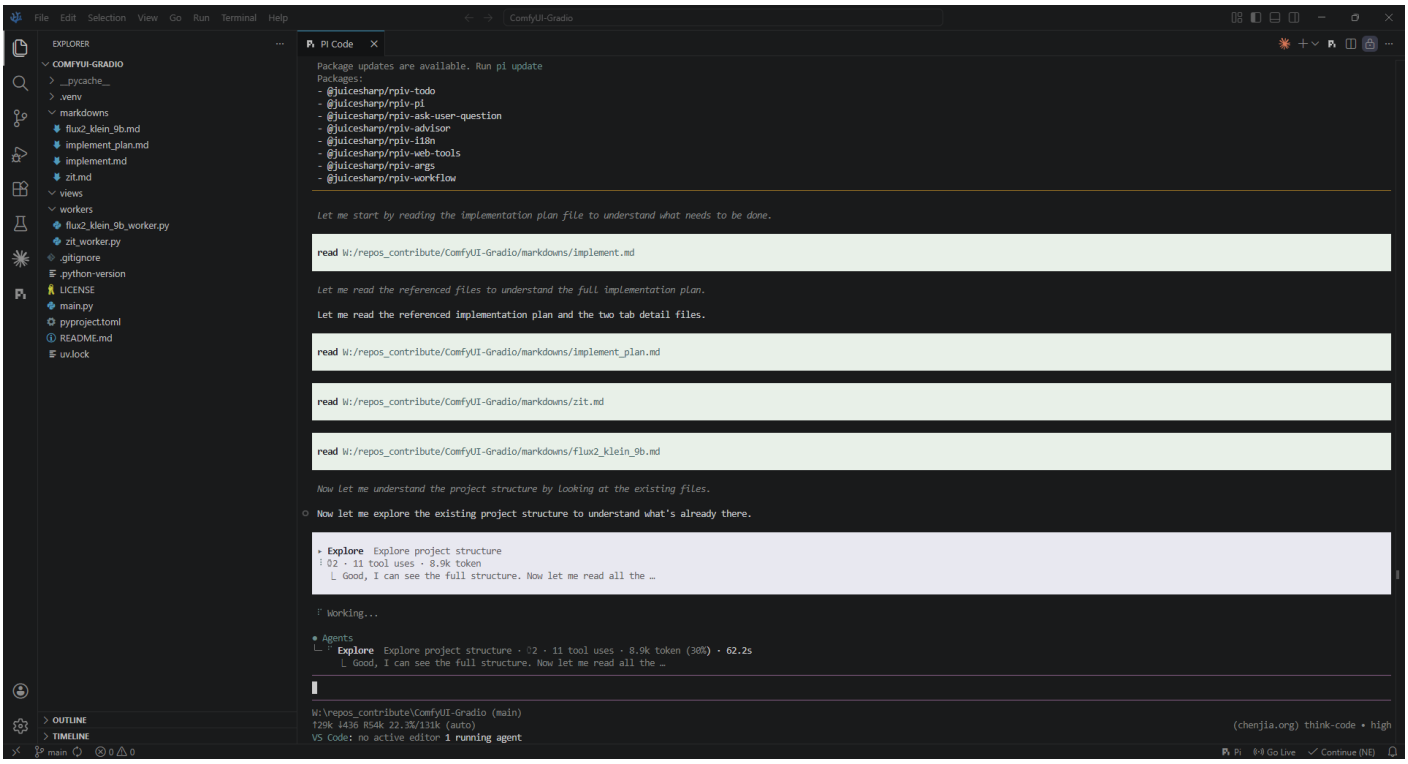
□□ Helloworld

Google public **Turboquant**, which is KV-Cache smaller and fast enought.
It reduce 16 bit data into 3-4bit only and frees up to 83% of the memory typically consumed by long prompts.

I self-build this fork llama-cpp-turboquant and have a first try.

docker-compose.yaml

```
services:
  llama-cpp:
    image: llama-cpp:turboquant
    container_name: llama-cpp
    ports:
      - 8080:8080
    privileged: true
    ipc: host
    volumes:
      - /mnt/data/files/models:/app/models
    command: >
      -m /app/models/unsloth/Qwen3.6-35B-A3B-MTP-GGUF/Qwen3.6-35B-A3B-UD-IQ4_NL.gguf
      --port 8080 --host 0.0.0.0
      -t 12 -c 262144
      --parallel 1 --no-context-shift --no-mmap --jinja
      -b 4096 -ub 4096
      -ngl 999 --cpu-moe
      -ctk turbo3 -ctv turbo2
      --spec-type draft-mtp --spec-draft-n-max 3
      --kv-unified --cache-ram 0
    deploy:
      resources:
        reservations:
          devices:
            - driver: nvidia
              count: all
              capabilities:
                - gpu
    restart: unless-stopped
```



□□ Result

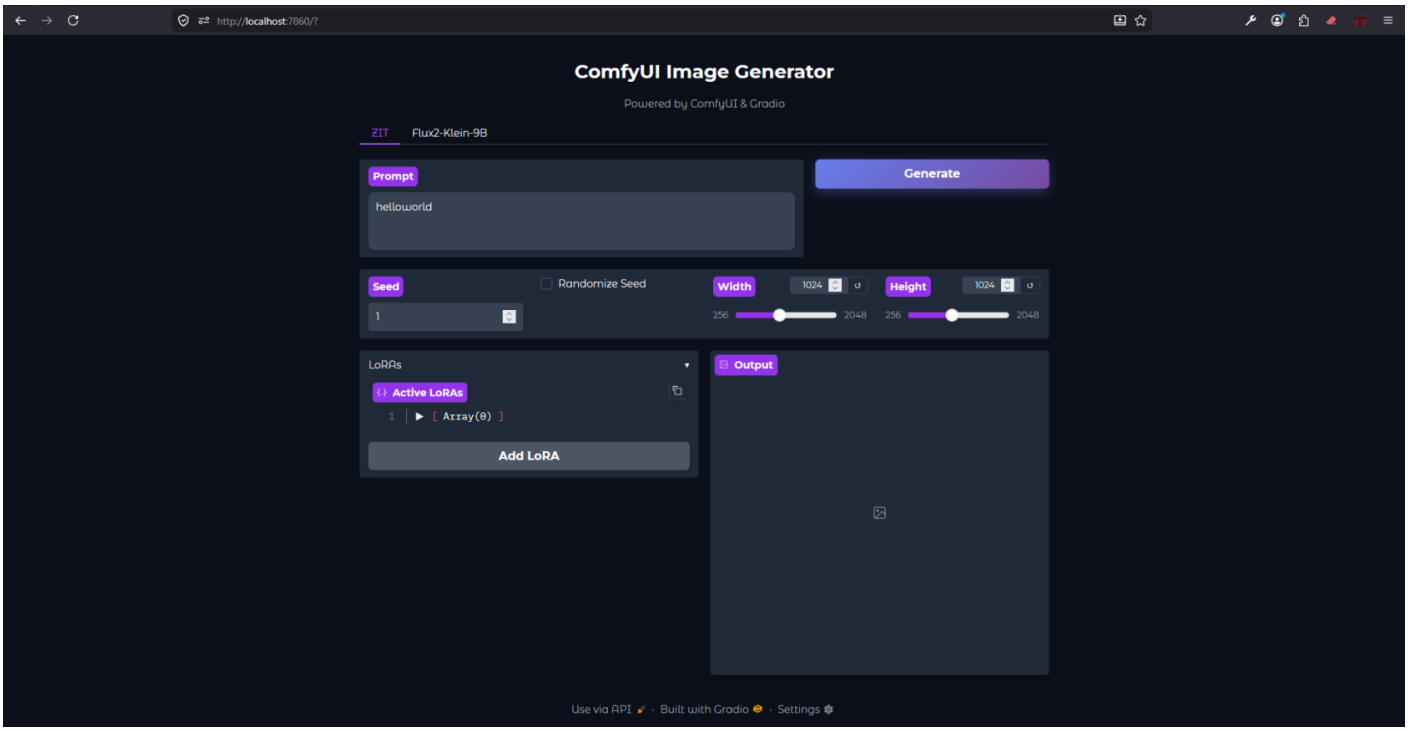
	Original	TurboQuant
token/s	29	32
RAM	26 GB	20 GB
vRAM	11 GB	10 GB

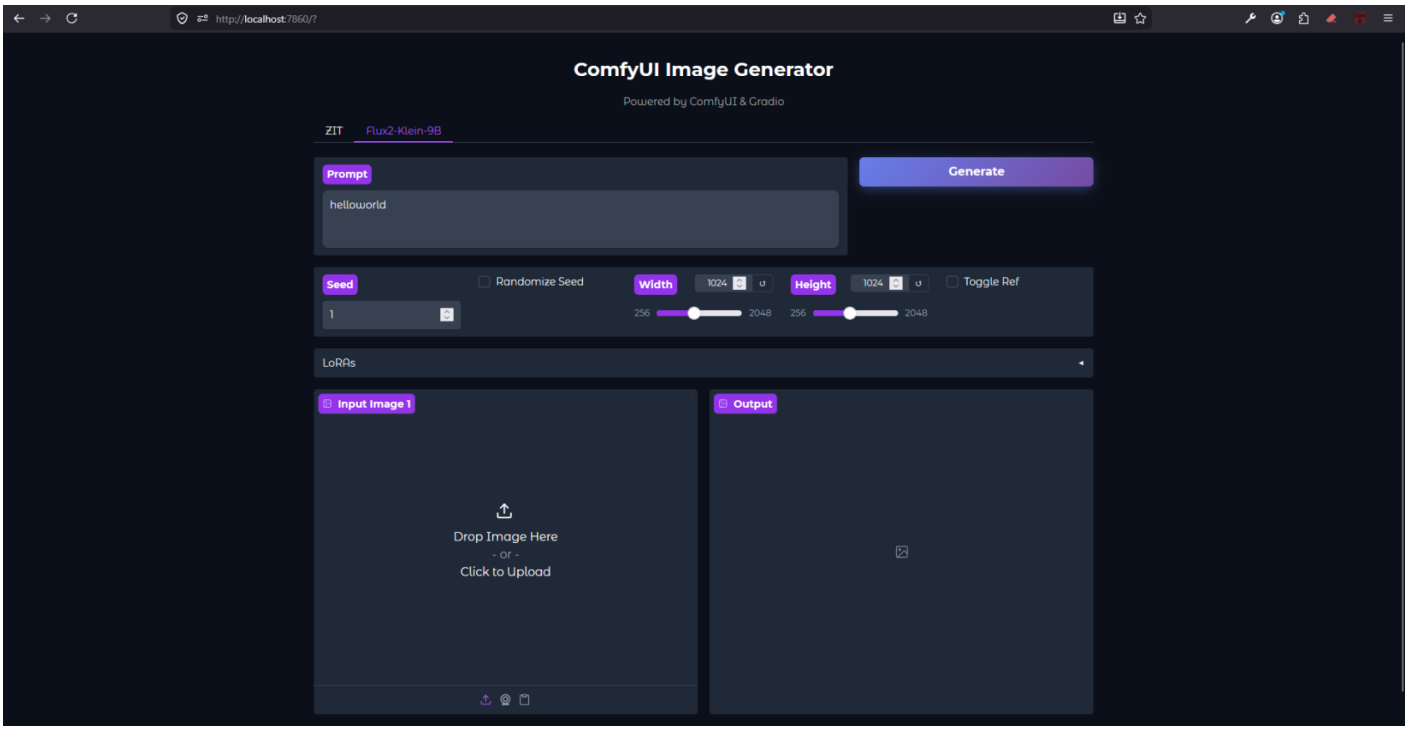
Speed loose **~10%**, but RAM consumption lower.

□□ Precision

Ok, now we known Turboquant help memory usage lower.
But how about precision ?

> “ The result is good, better UI/UX than Qwen Coder web.





Revision #10
Created 2026-05-28 12:23:24 UTC by sandichhuu
Updated 2026-05-28 13:33:17 UTC by sandichhuu