Character Consistency:

Using chain of edit with the Qwen edit AIO model.
The edit model is plugged with NSFW Lora and Lora for reducing denoising steps.
The Lora used:
NSFW(MCNL (Multi Concept NSFW Lora) [Qwen Image] - v1.0 | Qwen LoRA | Civitai)
PENISLORA(Penis Lora (+Blowjob, +Cumshot) [Taz] - WAN 2.2 14b / 5B / 1.3b T2V & I2V (Wan 2.1 & 2.2) + Qwen + Zimage Turbo - Qwen V2 | Qwen LoRA | Civitai)
The image generation goes through several small editing steps to reach the editing goal
Grok will decompose User's desire into several editing steps and organize the editing order.
There are totally 4 editing step:

- Background
- Pose
- Outfit
- Camera view

After each generation step, workflow will check face similarity against the original image, if the face sim is below 40 (tuned after several running), workflow will rerun the image, max is 3 times.

The face sim check algorithm is used as a custom node in comfyui, including a small CNN model to detect face and using face landmark detection model and similarity is computed using cosine sim.

Technical Stack:
ComfyUI is used to serve the model and the workflow.
A backend service is used to handle user requests and queueing requests to ComfyUI url localhost.

Grok is used as a chat bot character and orchestrator, it also generates a prompt of each edit step.
Grok API usage is also logged and the output and input of Grok are texts.

Grok not only orchestrate, it also choose which Lora is suitable for each generation step
Eachtime, backend will override positive prompt and image size and Lora switch and receive generated image and face sim from ComfyUI.

=> Why not using InstantID or IP-Adapter-FaceID => these are for older Image generation models like SDXL,... and not worked with Qwen family.

Qwen Image edit has several Lora support and is optimized for VRAM limited GPUs.

We have tried Inpainting for background editing and Controlnet for pose generation, however, separating these step increase the generation time, more over, while working with inpainting, model often lose the space and geometry consistency as it can not understand the whole image context.

The Controlnet works for pose generation but it limit the choices of pose, which is contrary to the intent of user freely prompt poses .