Rx 6750 XT / Testing parameters Stable Diffusion

PYTORCH_CUDA_ALLOC_CONF=

The variations of the threshold in this parameter, dont have mayor effects, could looks a placebo

garbage_collection_threshold:0.5, max_split_size_mb:256 ⭐

The variations of the threshold in this parameter, dont have mayor effects, could looks a placebo, i test it with 0.3 - 1 Values.

max_split_size_mb, The VRAM problems occurs when try use 512 in this option.
-512 ❌
- 256 ✅ ⭐ (I cant see differences)
-128 ✅ (I cant see differences)
- Less ✅ (I cant see differences)

Tema flotante

--opt-split-attention

The "opt-split" It very important or the VRAM problems keep always, sure to add this to you command line ✅ ⭐
If you remove this flag, your get a VRAM problem all time ❌

--medvram

contrary to what I thought, the performance has a slight improvement, which I consider very important

Performance comparation

	❌ Disabled	✅ ⭐ Enabled
	1.21s/it 1.25s/it 1.19s/it	1.07s/it 1.12s/it 1.06s/it

quad Family

This particular family of arguments capte my attention.

They are 3 parameters

--opt-sub-quad-attention

By default is false
Enable memory efficient sub-quadratic cross-attention layer optimization.

I am not sure about of this. Could said that in false and true get the same perfomance, maybe maybe in false as default have a minimum impact, with a very small best performance. Test by your self

For now, I prefer keep it in true

--sub-quad-q-chunk-size

	1024 ✅ ⭐	512	256	128
Note: Now, when i try use higher value 1024 like 2048 ↘ (Decrese perfomance too { in my Rx6750XT}) but, its posible that using max_split_size_mb:512 and --sub-quad-kv-chunk-size 256, you could get better performance, you should try it if you have a gpu with more of 12VRAM. I am not sure because i only have 12VRAM.	It is the default value And in my test, when I use default value as higher value and have good the other settings, it increse the perfomance.	↘ (Decrese perfomance)	↘ (Decrese perfomance)	↘ (Decrese perfomance)

--sub-quad-kv-chunk-size

--sub-quad-chunk-threshold

They have a cute relationship with the VRAM problems, and working together, this values are very very important !!!

kv-chunk-size by default is none

And according to value, you must change the --sub-quad-chunk-threshold

64

If you set --sub-quad-kv-chunk-size 64

You must test the values on, --sub-quad-chunk-threshold

100 is the max aprox or you get VRAM problem, If you try keep higher values could get more performance and less stability.

128 ✅ ⭐

If you set --sub-quad-kv-chunk-size 128

You must test the values on, --sub-quad-chunk-threshold

80 - 85 is the max aprox or you get VRAM problem, If you try keep higher values could get more performance and less stability.

256

If you set --sub-quad-kv-chunk-size 256

You must test the values on, --sub-quad-chunk-threshold

60 - 65 is the max aprox or you get VRAM problem, If you try keep higher values could get more performance and less stability.

512

If you set --sub-quad-kv-chunk-size 512

You must test the values on, --sub-quad-chunk-threshold

50 - 55 is the max aprox or you get VRAM problem, If you try keep higher values could get more performance and less stability.