Rx 6750 XT / 测试参数稳定扩散

pytorch_cuda_alloc_conf=

这个参数中的阈值变化，没有市长效应，可能看起来是安慰剂。

garbage_collection_threshold:0.5, max_split_size_mb:256 ☆

这个参数中的阈值的变化，没有可能的影响，可能看起来是安慰剂，我用**0.3-1**值进行测试。

max_split_size_mb, 当尝试在这个选项中使用**512**时，就会发生**VRAM**问题。
-512 □
- 256 □☆ **(I cant see differences)**
-128 □ **(**我看不出有什么不同**)**
- 更少□ **(**我看不出有什么不同**)**

漂浮主题

--opt-split-attention

opt-split "非常重要，否则**VRAM**问题会一直存在，一定要把它添加到你的命令行中 □ ☆
如果你删除这个标志，你会得到一个**VRAM**问题所有时间的□

--medvram

与我所想的相反，性能有轻微的改善，我认为这非常重要。

业绩比较		
	□已禁用	□☆已启用
	1.21s/it **1.25s/it** **1.19s/it**	**1.07s/it** **1.12s/it** **1.06s/it**

四个家庭

这个特殊的论证系列引起了我的注意。

它们是**3**个参数

--opt-sub-quad-attention

默认为假
启用内存效率高的次四元交叉注意层优化。

我不确定这一点。可以说，在假的和真的情况下得到同样的性能，也许在假的情况下，作为默认情况，影响最小，性能最好的是非常小。你可以自己测试一下

现在，我更愿意保持它的真实性

--次四边形-q-chunk-size				
	1024□☆	5 1 2	256 📉	128
注意：现在，当我尝试使用更高的**1024**值，如**2048**↘（也降低了性能**{**在我的**Rx6750XT}**），但是，它可能使用**max_split_size_mb:512**和**--sub-quad-kv- chunk-size 256**，你可以得到更好的性能，如果你有一个**12VRAM**以上的**gpu**，你应该试试。我不确定，因为我只有**12VRAM**。	它是默认值 在我的测试中，当我使用默认值作为较高的值，并有良好的其他设置，它增加了性能。	↘ **(**减弱性能**)**	↘ **(**减弱性能**)**	↘ **(**减弱性能**)**

--sub-quad-kv-chunk-size

--sub-quad-chunk-threshold

他们与**VRAM**问题有着可爱的关系，并且一起工作，这种价值观非常非常重要！！！！

kv-chunk-size默认为无

而根据数值，你必须改变**--**次四分体**--**阈值

64

如果你设置**--sub-quad-kv-chunk-size 64**

你必须在**--**次四边形**--**块状阈值上测试数值

100是最大的，否则你就会出现**VRAM**问题，如果你试图保持更高的值，可能会得到更多的性能和更少的稳定性。

128 □☆

如果你设置**--sub-quad-kv-chunk-size 128**

你必须在**--**次四边形**--**块状阈值上测试数值

80-85是最高值，否则你会遇到**VRAM**问题，如果你试图保持更高的值，可能会得到更多的性能和更少的稳定性。

256

如果你设置**--sub-quad-kv-chunk-size 256**

你必须在**--**次四边形**--**块状阈值上测试数值

60-65是最高值，否则你会遇到**VRAM**问题，如果你试图保持更高的值，可能会得到更多的性能和更少的稳定性。

512

如果你设置**--sub-quad-kv-chunk-size 512**

你必须在**--**次四边形**--**块状阈值上测试数值

50-55是最大的值，否则你就会出现**VRAM**问题，如果你试图保持更高的值，可能会得到更多的性能和更少的稳定性。