

5.2.6 Finale Umsetzung, InterStableLLMRLLine

5.2.6.1 R-Line

Die *InterStableLLM Pipeline* erfüllt bis auf das eher monotone Ergebnis, beschrieben in Absatz 5.2.5.3 alle Kriterien. *InterStableLLMRLLine* versucht die menschliche Kreativität nachzuahmen, indem er generative Künstliche Intelligenz mit zufällig in das Eingangsbild generierten Inhalten kombiniert. Die Innovation liegt in dieser Interpretation menschlicher Kreativität als genau dieser Kombination von zufälligen Impulsen in der menschlichen Wahrnehmung, abgebildet durch die Zufallsgeneratoren, und der Erkennung von Mustern, abgebildet durch KI.

5.2.6.2 InterStableLLMRLLine Pipeline

Die *InterStableLLMRLLine Pipeline* basiert auf dem gleichen Prinzip wie die *InterStableLLM Pipeline* beschrieben in Absatz 5.2.5.3 nur wird hierbei das Originalbild durch den *R-Line Algorithmus* beschrieben im Absatz 5.2.7.1 verändert. Zu sehen ist die Pipeline in der Abbildung 5.2.7.2. Sie kann noch durch *Hyperparameter Tuning* und *Prompt-Engineering* verbessert werden, jedoch liefert die derzeitige Version bereits angemessene Resultate.

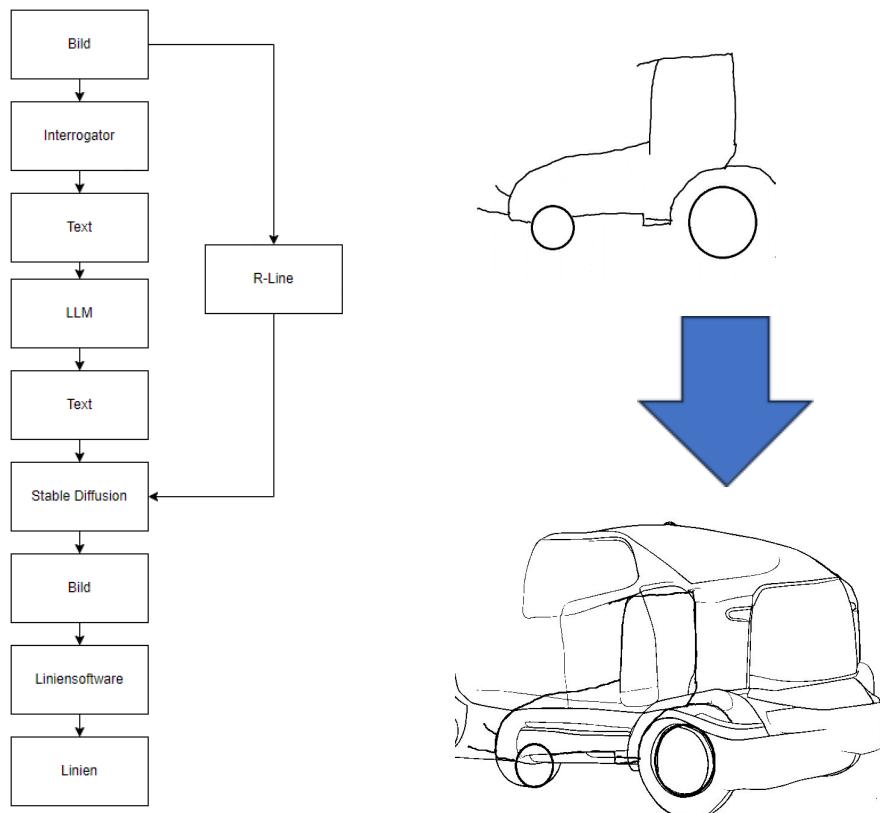


Abbildung 5.2.7.2 InterStableLLMRLLine Pipeline

5.4.3 Time / Memory Management

Um die Ausführzeit von neuronalen Netzwerken zu vermindern, werden die Rechenprozesse auf die GPU des Computers verlagert. Die neuronalen Netzwerke der Pipeline sind:

- der Interrogator
- das LLM (Mistral 7b)
- Stable Diffusion

Der Rechner, auf dem die Experimente ausgeführt wurden, hat eine Ryzen 5900 CPU, 32 GB Ram, eine RTX3070Ti mit 8 GB VRAM.

Je nach Bildgröße, in unserem Fall 512x768 Pixel, wird mehr RAM benötigt, da es mehr Rechenleistung erfordert. Stable Diffusion ist ein iterativer Prozess, und somit herrscht ein linearer Zusammenhang zwischen Zeit und Anzahl an Diffusionsprozessen. In der Regel sind 20 bis 50 Steps notwendig, um ein qualitatives Bild zu erzeugen, wobei eine geringere Diffusion Step Anzahl einer schlechteren Bildqualität entspricht.

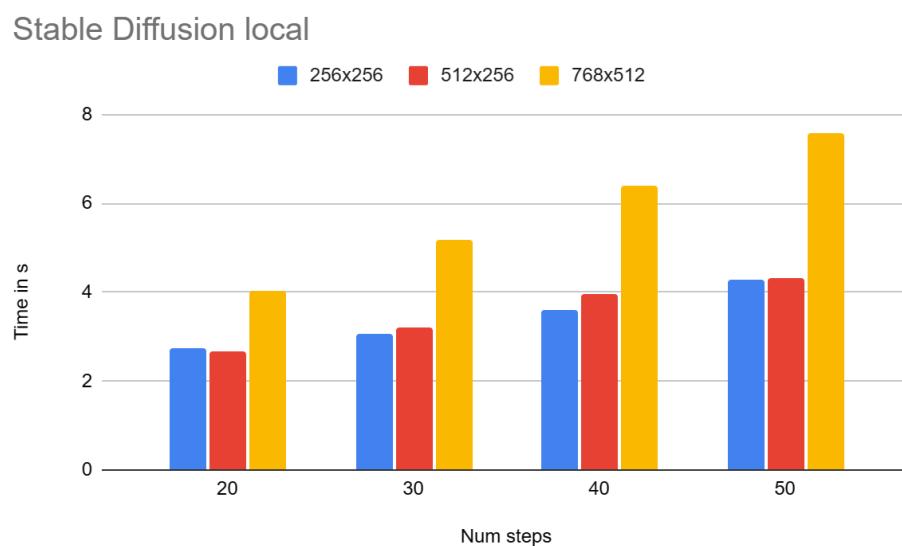


Abbildung 5.4.3.1 Stable Diffusion Time Management

Wie bei dem Stable Diffusion Algorithmus hängt auch die Zeit, die der Interrogator benötigt, von der Bildgröße ab. Da dieser Prozess vergleichsweise mit den anderen Algorithmen wenig Rechenleistung benötigt, ist es möglich, ihn auf der CPU laufen zu lassen, jedoch wird dadurch die Laufzeit verdoppelt.

Wie schon in Absatz 5.3.4.2 erwähnt, ist es möglich, einen Teil des LLMs auf die GPU zu verlagern. Auch ist es möglich durch verschiedene quantisierungs Methoden (Gewichte durch weniger Bits zu repräsentieren) die Inferenzdauer zu minimieren, jedoch wird hierbei die Qualität und Performance der Texte schlechter.

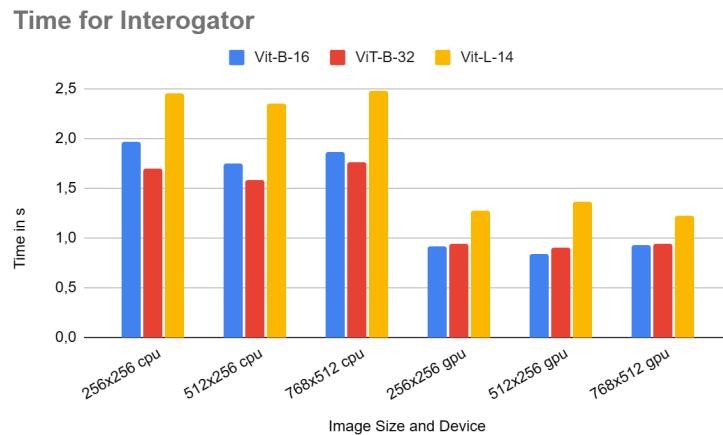


Abbildung 5.4.3.2 Interrogator Time Management

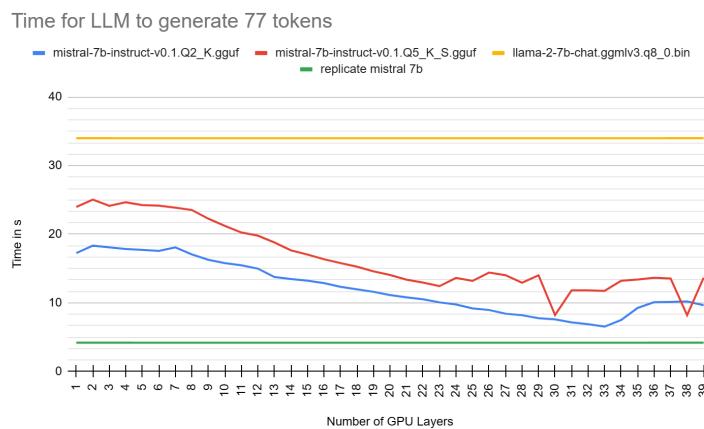


Abbildung 5.4.3.3 LLM Time Management