Limites GPU et impact sur la scalabilité

Contexte

On utilise deux modèles de diffusion: Stable Diffusion 2.1 & Stable Diffusion XL

Objectif : tester leur utilisation dans notre pipeline et voir les limites.

Notre machine actuelle est limitée à **16 GB de VRAM**.

C'est OK pour des tests simples.

Mais ça devient vite bloquant dès qu'on veut augmenter la production et avoir des très bonnes textures.

Observation

Stable Diffusion 2.1 reste utilisable.

On peut faire des générations de canard sans trop de problème.

Le modèle est plus gros et fonctionne souvent en deux étapes.

Ça augmente le temps de calcul et la complexité

Même si ça passe en 16 GB, ce n'est pas confortable.

Comparaison

This is generated with LLM.

Critère	**SD 2.1**	**SDXL**
Complexité	Moyenne	Élevée
Résolution	Moyenne	Élevée
Coût GPU	Gérable	Très élevé
16 GB VRAM	OK	Limité

Impact

Avec 16 GB :

- batchs très petits ⟶ lent
- résolution limitée ⟶ perte de qualité
- difficile de générer beaucoup de données
- pas de marge pour ajouter des modules

Résultat : le pipeline ne scale pas.

Besoin

On a besoin de plus de GPU.

Objectifs :

- stabiliser SDXL
- accélérer les tests
- produire plus de données

La machine actuelle suffit pour prototyper.

Conclusion

SD 2.1 est utilisable.

SDXL montre clairement les limites.

Le problème principal est la VRAM (16 GB).

Ça bloque :
- la vitesse
- la qualité
- la scalabilité

Il faut plus de ressources pour passer à l'échelle.

Synthèse

J'ai identifié que la limite principale du pipeline est le coût GPU des modèles de diffusion, en particulier SDXL qui ne se lance souvent pas. Avec 16 GB de VRAM, on peut tester mais pas scaler. Pour aller vers une production stable et rapide, il faut plus de puissance.