IA Générative : Apprenez à contrôler la génération d'images avec Stable Diffusion

Culture Sciences de l'Ingénieur

La Revue 3E.I

école normale supérieure paris-saclay

Simon Playe[1]

Édité le 11/03/2025

[1] *Data Scientist chez Sicara (Theodo Data & AI)*

Cette ressource fait partie du N°115 de La Revue 3EI du deuxième trimestre 2025.

1 - Introduction

Avez-vous déjà utilisé ChatGPT pour créer des images et avez-vous été déçu par les résultats ? Si vous avez ressenti cette frustration ou si vous êtes simplement intéressé par des outils de génération d'images plus efficaces, Stable Diffusion pourrait être la solution idéale.

Avec la popularité de ChatGPT, nombreux sont ceux qui sont curieux de découvrir ce que l'IA générative peut accomplir. Elle ne se limite pas à la création de texte ; cette technologie peut également générer des images, des vidéos et même de la musique. Pourtant, il est parfois difficile d'obtenir les images que vous souhaitez simplement à partir d'une invite textuelle.

Cet article se penche sur Stable Diffusion, un outil conçu spécifiquement pour générer des images. Ici, je vous montrerai comment utiliser une API pour contrôler la création d'images avec Stable Diffusion, en couvrant des options tant pour les non-développeurs que pour les développeurs. Les méthodes abordées dans cet article incluent :

- Diriger le processus de diffusion vers une image spécifique (IP Adapter).
- Maintenir certaines caractéristiques dans une image via le processus de diffusion (ControlNet).
- Extraire des caractéristiques à partir d'une image initiale (Image-to-Image).
- Former un modèle spécifique avec un ensemble limité d'images pour des ajustements ciblés (LORA, Model Fine-Tuning).
- Ajouter de nouveaux poids aux couches de croisement d'attention pour modifier les caractéristiques de l'image (LORA).

2 - Qu'est-ce que Stable Diffusion ?

Stable Diffusion est un modèle de diffusion spécifiquement conçu pour la génération d'images. Il commence avec un motif initial de bruit aléatoire et affine systématiquement ce bruit ou le "débruite" pour produire des images qui ressemblent de près à des photos réelles. Le modèle guide cette transformation en appliquant certaines conditions (par exemple une invite textuelle) durant le processus de débruitage.

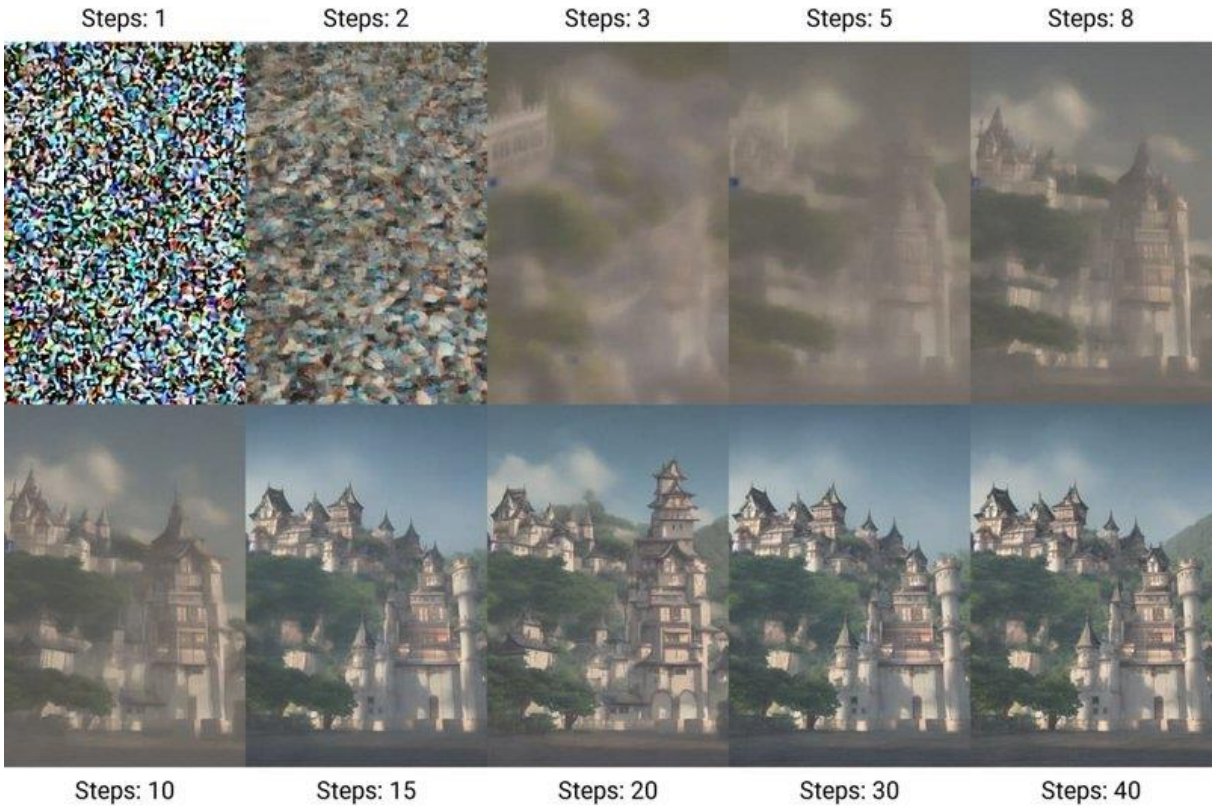

Figure 1 : Les étapes de débruitage
source: https://en.wikipedia.org/wiki/Diffusion_model#/media/File:X-Y_plot_of_algorithmically-generated_AI_art_of_European-style_castle_in_Japan_demonstrating_DDIM_diffusion_steps.png

Contrairement aux méthodes traditionnelles qui se contentent de prompts textuels, Stable Diffusion offre des contrôles plus flexibles et sophistiqués :

- Conditionnement basé sur l'image : Grâce à des fonctionnalités comme l'IP-Adapter et ControlNet, Stable Diffusion peut utiliser une image existante pour orienter la génération, permettant des modifications ou des améliorations basées sur cette image.
- Transformation Image-à-Image : Cette fonctionnalité permet au modèle de partir non pas de zéro mais d'une image existante, qu'il transforme ensuite en une nouvelle création.
- Contrôle de style : Les utilisateurs peuvent choisir des modèles générant des images dans des styles artistiques spécifiques, ou même entraîner leurs modèles pour produire des styles personnalisés.
- Conditionnement hybride : Il permet de combiner des invites textuelles, des images et différents modèles pour conditionner la génération, offrant un contrôle sans précédent sur la sortie.

Pour expérimenter avec ces diverses techniques de contrôle de la génération d'images, ModelsLab propose une API Stable Diffusion pratique. Elle est disponible sous deux formats :

- **Version Playground** : Une interface conviviale conçue pour l'expérimentation et l'exploration sans nécessiter de connaissances techniques étendues.
- **API pour Développeurs** : Offre un contrôle et une personnalisation plus détaillés, adaptés aux développeurs souhaitant intégrer ces capacités dans leurs applications.

Dans les sections suivantes, je vais approfondir chaque technique, en montrant comment utiliser la plateforme ModelsLab pour réaliser votre vision créative.

2.1 - Stable Diffusion text-to-image

Avant de plonger dans les divers outils permettant de contrôler la génération d'images, commençons par une introduction à l'API Stable Diffusion text-to-image. Cette API, semblable à Dall-E, facilite la génération d'images à partir de prompts textuels.

Pour les développeurs, il existe deux points de terminaison principaux : **text2img** et **realtime-stable-diffusion**. Le point de terminaison **text2img** est le choix principal pour la génération d'images, accessible via le playground et l'API pour développeurs. Il permet aux utilisateurs de créer des images avec des modèles de diffusion entraînés par la communauté, disponibles sous les versions **Stable Diffusion** et **Stable Diffusion XL**. La version XL offre des images plus précises mais nécessite plus de temps pour la génération.

Dans le playground, les utilisateurs peuvent personnaliser leur génération d'images en définissant différents paramètres tels que :

- **Negative Prompt** : Spécifiez ce que vous ne voulez pas voir apparaître dans l'image.
- **Guidance Scale** : Définissez l'importance de l'influence du prompt sur le débruitage.
- **Steps** : Déterminez le nombre d'étapes de génération, ce qui affecte le détail de l'image.

L'API développeur offre également des options supplémentaires, telles que :

- **enhance_style** : Choisir un style spécifique pour l'image.
- **highres_fix** : Créer des images en haute résolution.

De plus, le point de terminaison **realtime-stable-diffusion** offre moins d'options de personnalisation et ne permet pas de choisir un modèle de diffusion, mais il est plus rapide pour générer des images.

Voici un exemple rapide de fonctionnement de **text2img**.

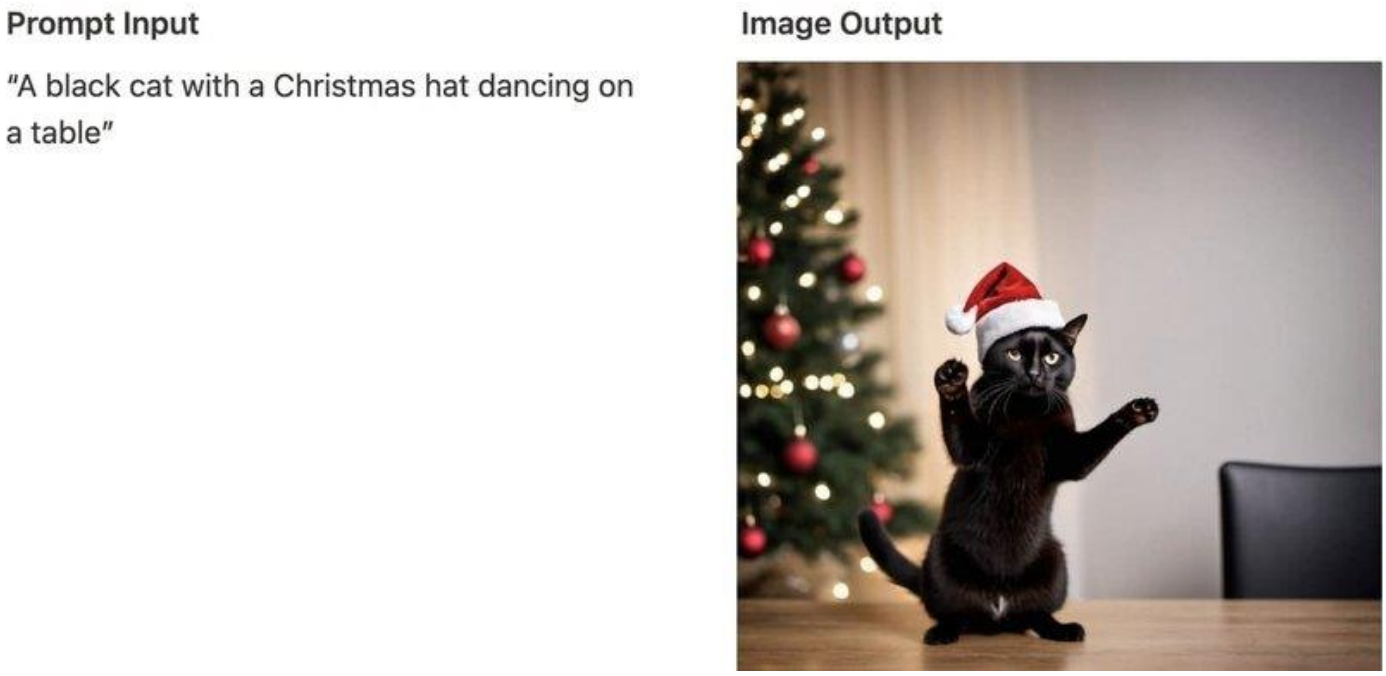

Figure 2 : Un chat dansant sur une table avec un chapeau de noël

3 - Contrôle de la génération d'images avec Stable Diffusion

Dans cette section, nous explorerons trois outils avancés fournis par Stable Diffusion qui offrent des façons alternatives de guider la génération d'images, au-delà du prompt textuel conventionnel :

- **IP-Adapter**
- **ControlNet**
- **Image-to-Image**

3.1 - IP-Adapter

Imaginez que vous puissiez guider la génération d'images non pas en donnant des instructions textuelles, mais en utilisant une image cible à la place. C'est exactement ce que fait **IP-Adapter**. Il permet de diriger le processus de génération avec une image, tout comme un prompt textuel. Cependant, l'IP-Adapter doit être utilisé avec un prompt textuel (ou une image, comme nous le verrons ci-dessous).

Cependant, l'IP-Adapter est uniquement disponible via l'API développeur sur le point de terminaison **img2img**. Dans ce paramètre, vous pouvez ajuster des paramètres tels que :

- **ip_adapter_id**, qui définit comment l'image est encodée,
- **ip_adapter_scale**, qui affecte l'impact de l'image IP-Adapter sur la réduction du bruit,
- **ip_adapter_image**, qui est l'URL de l'image IP-Adapter.

Adapter avec un prompt textuel

Figure 3 : Create a rose and cute castle with IP-Adapter (Créez un château rose et mignon avec IP-Adapter)

3.2 - ControlNet

ControlNet fonctionne de manière similaire à l'IP-Adapter, mais avec une approche plus nuancée. Contrairement à l'IP-Adapter, qui influence la génération d'images en fonction de l'image entière, ControlNet cible des caractéristiques spécifiques dans l'image pour guider le processus de génération vers ces détails.

Il existe plusieurs types de ControlNet qui permettent d'extraire des caractéristiques spécifiques d'une image :

- **softedge :** trouve les contours dans les images
- **canny :** délimite avec précision les frontières dans des environnements contrôlés
- **openpose :** détecte les points clés du corps humain, des mains, du visage et des pieds
- etc.

Vous pouvez également intégrer plusieurs modèles ControlNet en énumérant plusieurs paramètres controlnet_model, séparés par des virgules, comme "canny, softedge".

L'adaptabilité de ControlNet signifie qu'il peut fonctionner en parallèle avec des prompts textuels et des images IP-Adapter. Il est facile à utiliser pour les non-développeurs dans l'interface, tandis que les développeurs peuvent ajuster des paramètres supplémentaires, tels que :

- **controlnet_type :** Cela définit l'un des différents types acceptés de modèles ControlNet, comme canny, depth, hed, etc. Vous pouvez trouver la liste complète des types de modèles disponibles ici.
- **controlnet_model :** Cela spécifie le modèle ControlNet spécifique utilisé, qui peut être un modèle par défaut ou un modèle communautaire. Dans le cas d'un modèle par défaut, controlnet_model correspond directement à controlnet_type.
- **controlnet_conditioning_scale :** Cela détermine dans quelle mesure ControlNet influence le processus de réduction du bruit.
- **control_image (optionnel) :** Il s'agit de l'image à partir de laquelle les caractéristiques seront extraites. Si ce paramètre n'est pas spécifié et qu'une image initiale est fournie, cette image sera utilisée.

ControlNet sans prompt textuel et sans image IP-Adapter.

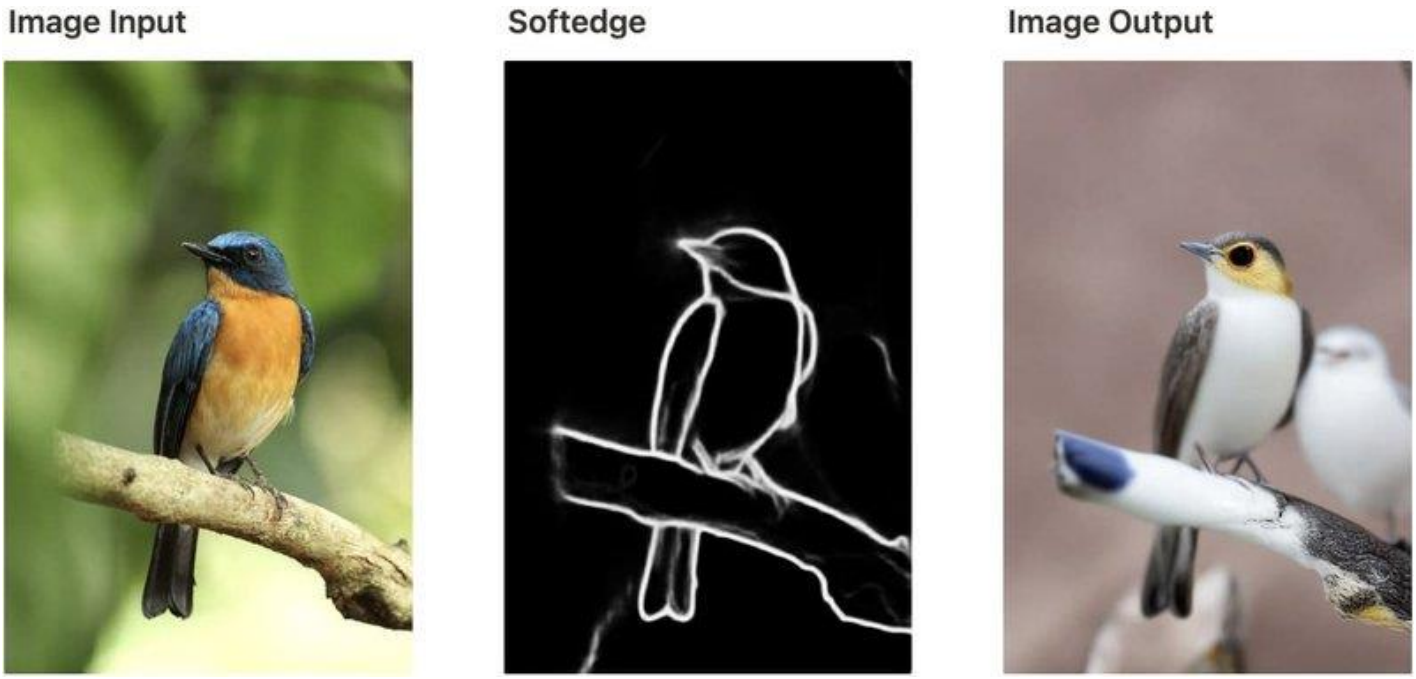

Figure 4 : Transforming a bird picture with a softedge ControlNet (Transformation d'une image d'oiseau avec un softedge ControlNet)

ControlNet avec prompt textuel et sans image IP-Adapter

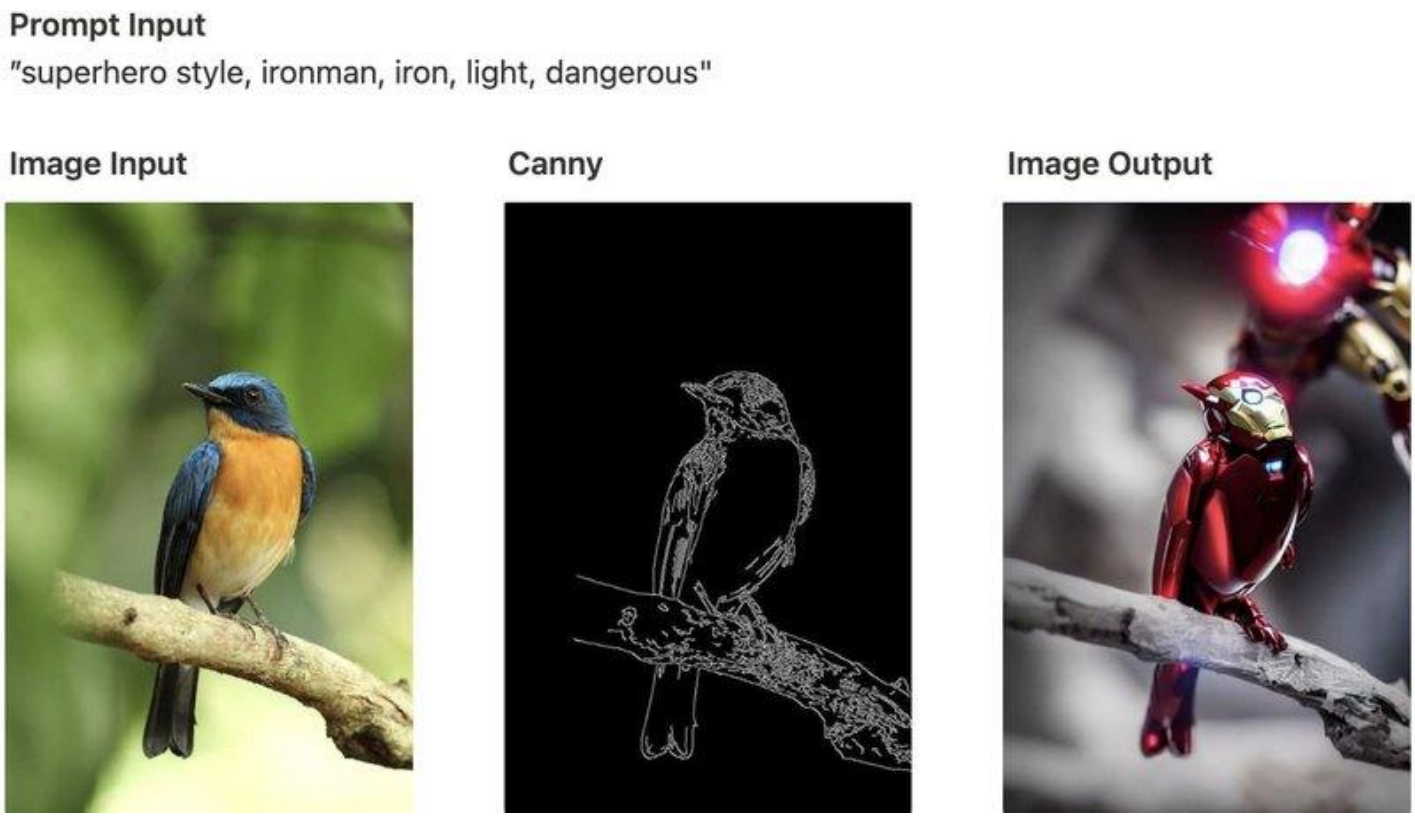

Figure 5 : Transforming a bird picture with a prompt and a canny ControlNet (Transformer une photo d'oiseau grâce à un prompt et un canny ControlNet)

3.3 - Génération d'Image-à-Image

Enfin, au lieu d'orienter la génération d'image en conditionnant la sortie, pourquoi ne pas commencer la génération à partir d'une image qui partage des caractéristiques avec votre image finale ? C'est ainsi que fonctionne l'API Stable Diffusion img2img. Elle ne part pas du bruit aléatoire complet, mais ajoute du bruit à l'image initiale. Elle est conçue pour capturer les caractéristiques générales de l'image initiale, comme sa couleur et sa composition.

La génération Image-à-Image peut être combinée avec un prompt textuel, ControlNet et/ou IP-Adapter.

Parce que la génération img2img fonctionne de manière similaire à la génération texte-à-image, les deux API partagent presque les mêmes fonctionnalités. Les modèles communautaires et realtime-stable-diffusion peuvent être appliqués à l'un ou l'autre type de génération.

Génération d'image avec un prompt textuel et sans IP-Adapter

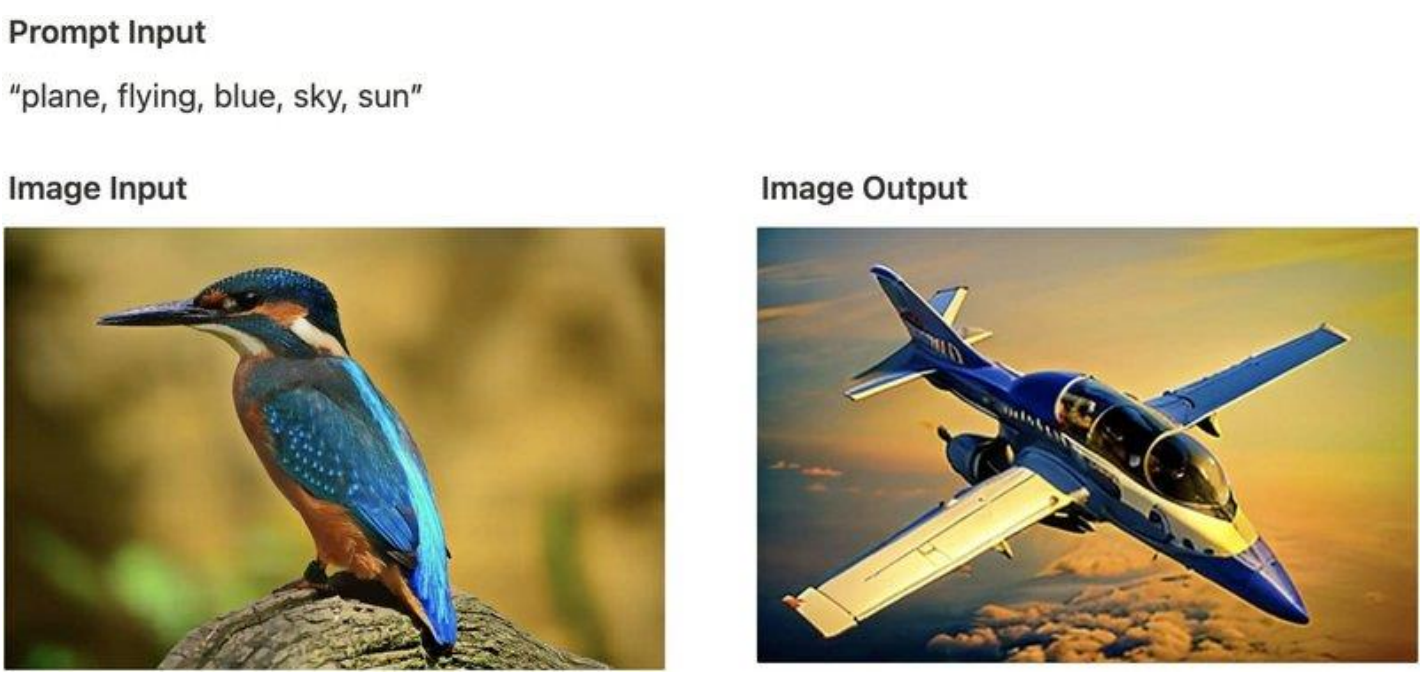

Figure 6 : A plane inspired from a bird picture (Un avion inspiré d'une photo d'oiseau)

Génération d'image sans commande textuelle et sans IP-Adapter

Figure 7 : Image d'un paysage de « fantasy » inspiré d'un dessin, source : https://tinyurl.com/2eyf5ky5

3.4 - Conclusion sur le Contrôle de la Génération d'Images

En conclusion, les générateurs d'images traditionnels comme Dall-E s'appuient uniquement sur des prompts textuels pour le contrôle, mais **Stable Diffusion** introduit des outils puissants tels que **IP-Adapter**, **ControlNet** et **Image-à-Image** pour diversifier et améliorer la génération d'images. Les utilisateurs peuvent utiliser ces outils indépendamment ou en combinaison, offrant une grande flexibilité et créativité pour générer des images, avec ou sans prompts textuels.

Sélection et entraînement de modèles spécifiques

Jusqu'à présent, nous avons parlé de l'ajout de conditions pour contrôler la génération d'images. Mais pourquoi ne pas modifier totalement le processus de génération ? Derrière la génération d'images se cache un modèle de diffusion. Cependant, ces modèles peuvent être modifiés pour générer des images plus spécifiques. Ici, je vais présenter deux façons d'améliorer le processus de génération : en sélectionnant ou en affinant un modèle de diffusion ou un **LoRA**. Ces deux approches peuvent être combinées avec les outils présentés ci-dessus.

Modèles de Diffusion

Comme mentionné précédemment, **Stable Diffusion** fonctionne avec un modèle de diffusion. Un modèle de diffusion est un modèle formé pour générer des images. Pour ce faire, ce modèle est entraîné sur un ensemble d'images qu'il cherche à reproduire. Par exemple, si vous voulez un modèle qui génère des images de chiens, vous devrez entraîner un modèle de diffusion en utilisant de nombreuses images de chiens.

La puissance de **Stable Diffusion** réside dans le fait qu'il permet un accès facile aux modèles affinés par la communauté. L'**affinage** est un processus de réentraînement partiel du modèle standard de **Stable Diffusion** sur votre propre ensemble d'images. Au lieu d'utiliser le modèle standard qui génère des images classiques, vous pouvez choisir un modèle qui génère des images de pixels, par exemple.

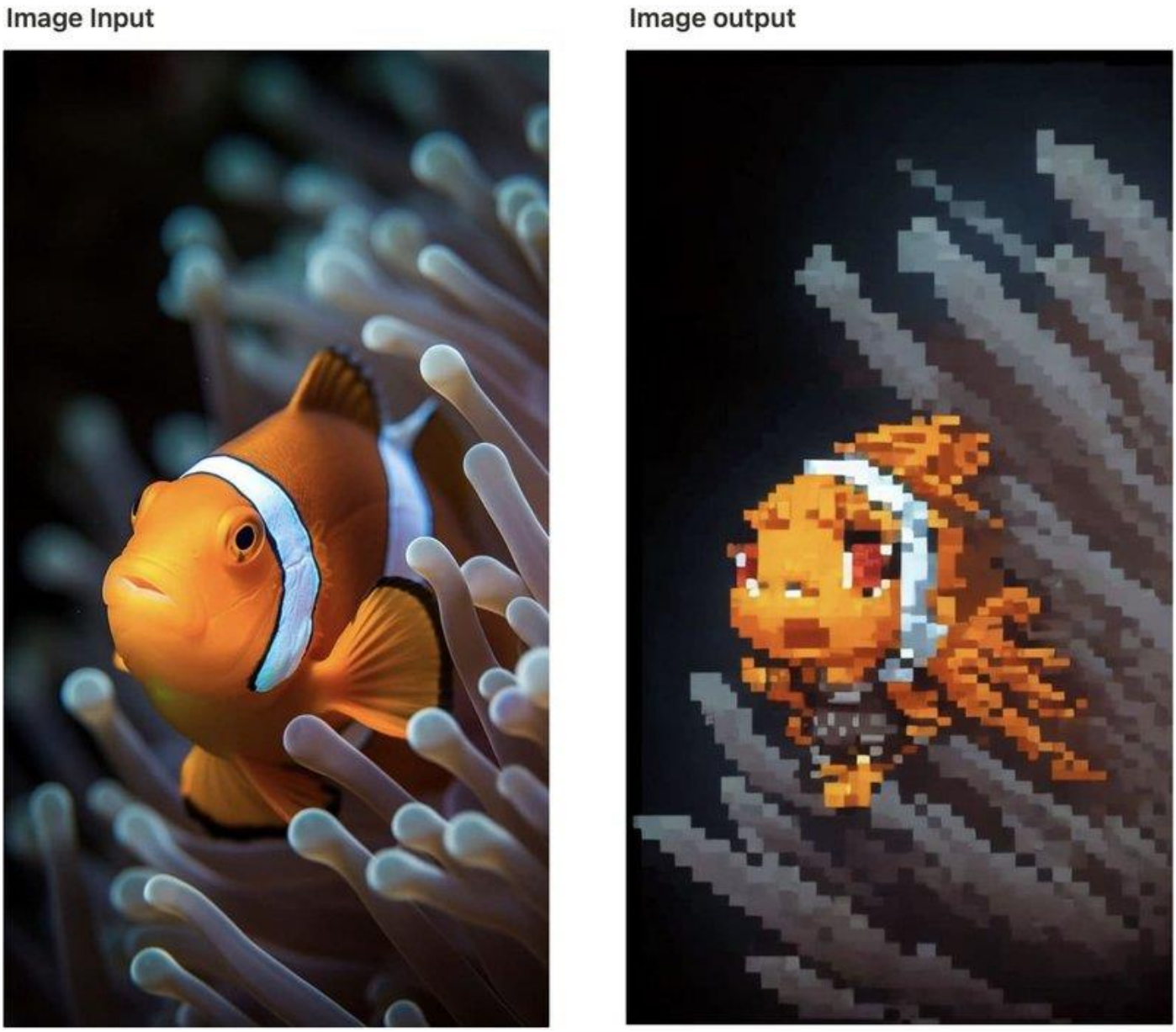

Figure 8 : A pixelated fish using model Chibi Pixel Art Style (Un poisson pixelisé utilisant le modèle Chibi Pixel Art Style)

Ou pour générer des images de cartoons :

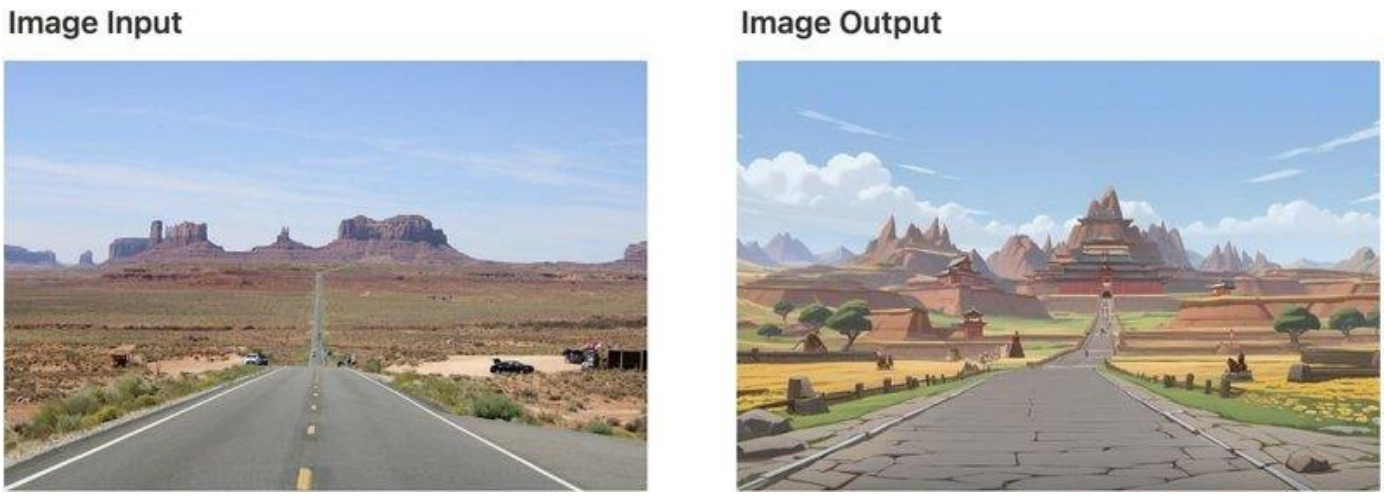

Figure 9 : Cartoon Image using model Cartoon Backgrounds (Image de bande dessinée utilisant un modèle Arrière-plans de bande dessinée)

La sélection de modèles est disponible à la fois pour les utilisateurs non-développeurs et via l'API développeur.

Enfin, si aucun modèle ne correspond à vos attentes, vous pouvez également affiner votre propre modèle. Selon la documentation, l'affinage nécessite seulement 7 à 8 images. Cependant, cela n'est disponible que via l'API développeur.

4 - LoRA

Enfin, **LoRA** est le dernier outil que je vais vous présenter pour avoir plus de contrôle sur la génération d'images. Les **LoRA** sont des versions compactes des modèles **Stable Diffusion**, généralement 10 à 100 fois plus petites. Les LoRAs ajoutent des modifications par-dessus le modèle de diffusion. Elles affinent les modèles standards, modifiant subtilement des styles spécifiques ou l'apparence générale des images générées. L'avantage des LoRAs réside dans leurs faibles besoins en calcul et leurs temps d'entraînement rapides.

Tout comme pour les modèles de diffusion, vous pouvez facilement utiliser des LoRAs affinées par la communauté. Mais, contrairement aux modèles de diffusion et de manière similaire à **ControlNet**, vous pouvez utiliser plusieurs LoRAs à la fois en les séparant par des virgules. La sélection de LoRAs est disponible tant pour les utilisateurs non-développeurs que via l'API développeur.

De plus, entraîner votre propre LoRA est simple à l'aide de l'API développeur, et cela nécessite typiquement seulement 7 à 8 images.

Vous trouverez deux nouveaux paramètres pour les LoRAs dans l'interface utilisateur et l'API développeur :

- **lora_model** : Spécifie quel modèle LoRA utiliser.
- **lora_strength** : Détermine l'ampleur de l'impact du LoRA pendant le processus de réduction du bruit.

Voici un exemple d'un LoRA intitulé "Princess" :

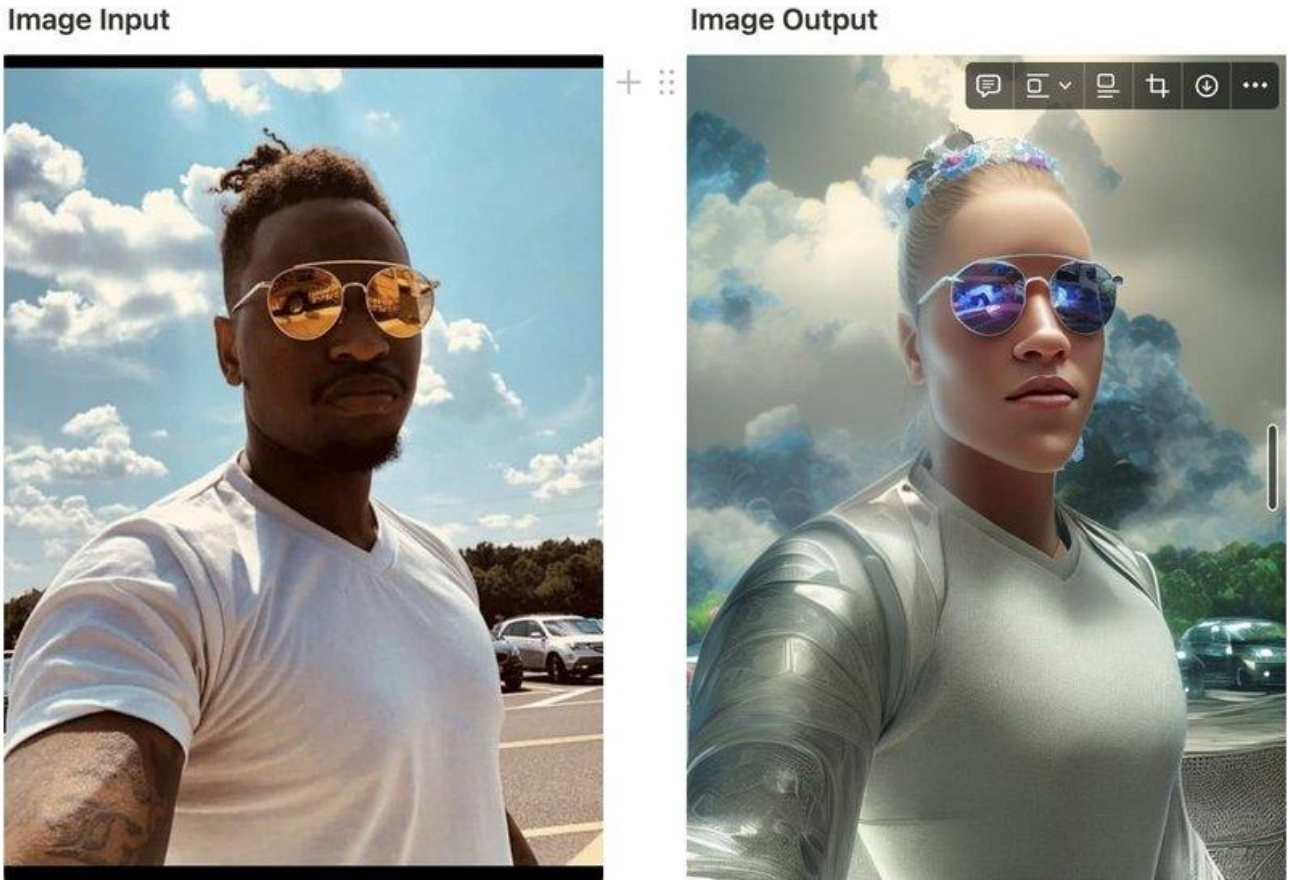

Figure 10 : LoRA : Princess

Et ici une illustration pour enfant :

Figure 11 : LoRA : Illustration pour enfant

5 - Conclusion

En résumé, Stable Diffusion offre plusieurs façons de contrôler la génération d'images au-delà de l'utilisation simple de prompts.

En guise de conclusion, l'intégration des différents composants de Stable Diffusion dans un flux de travail cohérent peut être complexe. Pour ceux qui cherchent à créer des images détaillées avec un meilleur contrôle dans un format facile à utiliser, ComfyUI offre une solution idéale.

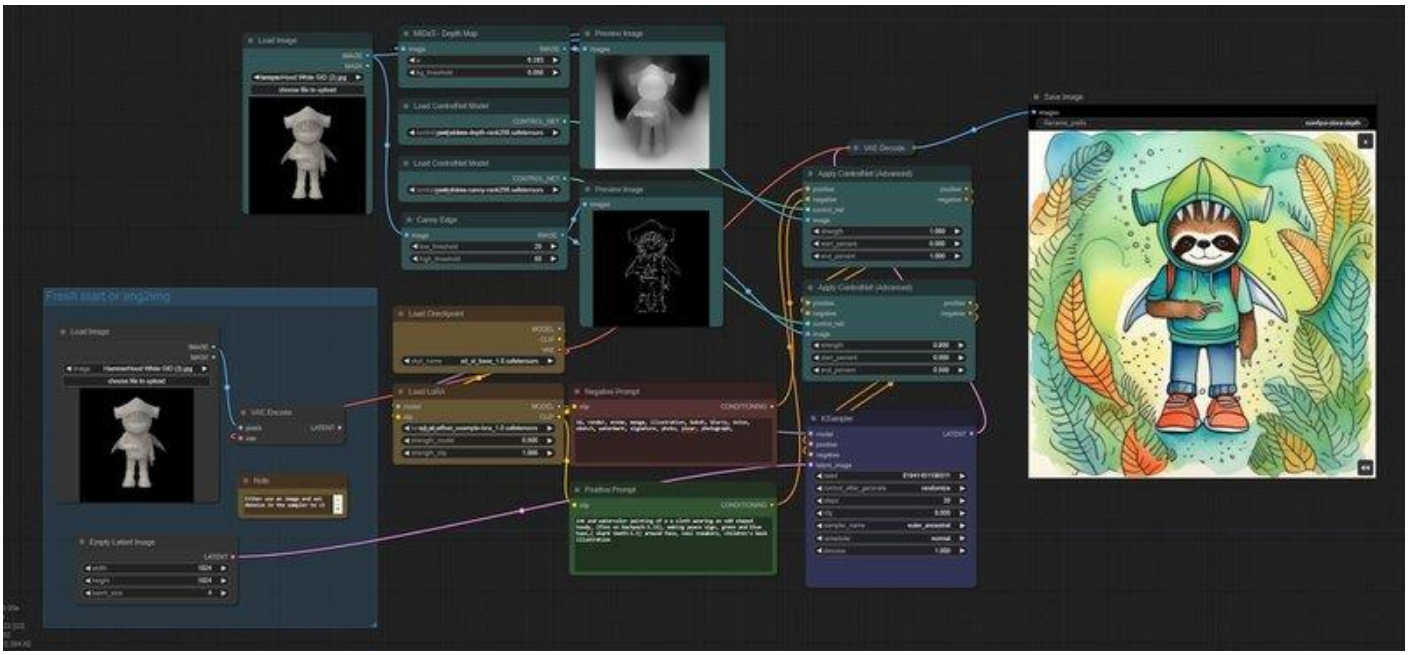

Figure 12 : Création d'une image détaillée à l'aide de ComfyUI et satble Diffsuion
source : https://tinyurl.com/2wuudfat

ComfyUI vous permet de concevoir des pipelines pour la génération d'images en utilisant **Stable Diffusion**. Pour plus de détails sur ses fonctionnalités, vous pouvez consulter la documentation. Avec les outils appropriés et les connaissances nécessaires, le potentiel de l'IA générative est illimité — profitez de votre créativité !

6 - Références :

[1]: https://data-ai.theodo.com/en/technical-blog/generative-ai-image-generation-stable-diffusion