VERSATILE DIFFUSION : TEXTE, IMAGES ET VARIATIONS

Mathys ARNAUD, Ugo MERLIER, Lucas COTOT, Marwane Laghzaoui

12 avril 2025

1 Introduction

Dans le cadre de ce projet, nous avons étudié Versatile Diffusion, un modèle de diffusion génératif capable de produire du contenu à partir de différentes modalités d'entrée, notamment du texte et des images. Ce modèle se distingue par sa flexibilité et sa capacité à gérer des tâches variées de génération multimodale, ce qui en fait une solution puissante dans le domaine de l'intelligence artificielle générative. Notre travail a consisté à récupérer, analyser et expérimenter ce modèle afin d'en comprendre le fonctionnement interne, d'évaluer son efficacité, et d'explorer ses possibilités d'implémentation. Ce rapport présente donc le principe de base des modèles de diffusion, les spécificités de Versatile Diffusion, ainsi que les étapes techniques que nous avons suivies pour le faire fonctionner dans notre environnement.

2 Modèles de diffusion

Un modèle de diffusion est un algorithme utilisé en IA générative qui apprend à transformer du bruit en données structurées, comme des images ou des sons. Il fonctionne en inversant progressivement un processus de dégradation, recréant ainsi du contenu réaliste à partir de bruit aléatoire.

2.1 Modèles Denoising Diffusion Implicit Models

Dans le domaine des modèles de diffusion, un DDIM (Denoising Diffusion Implicit Model) est une variante qui permet de générer des données plus rapidement qu'un modèle de diffusion classique. Contrairement au processus standard, qui utilise beaucoup d'étapes pour débruiter progressivement une image, le DDIM propose une méthode de débruitage plus directe (non-stochastique), ce qui réduit le nombre d'étapes nécessaires tout en conservant une bonne qualité. C'est une version optimisée qui accélère la génération d'images sans trop sacrifier la précision.

3 Présentation Versatile Diffusion

Versatile Diffusion (VD) se distingue par sa capacité à traiter plusieurs modalités de manière fluide et cohérente dans un seul modèle. Contrairement aux approches traditionnelles, VD permet de générer des images à partir de descriptions textuelles (Text-to-Image). Par exemple, une description comme "un chat orange tigré dans un jardin" produit une image réaliste correspondant à cette scène. VD permet également de créer des descriptions textuelles à partir d'images (Image-to-Text). Ainsi, une image d'un chien jouant dans un parc peut être décrite par l'énoncé "un chien jouant dans un parc verdoyant".

Enfin, VD offre la possibilité de modifier des images existantes en fonction d'instructions textuelles (Image-to-Image). Par exemple, il est possible de transformer une image d'une voiture rouge en une voiture bleue tout en conservant ses proportions et ses détails.

Cette grande flexibilité repose sur une architecture multi-flux qui permet à VD de gérer simultanément plusieurs types de données. Cette approche unifiée contraste fortement avec les modèles classiques, qui nécessitent souvent des architectures distinctes pour chaque tâche ou modalité.

3.1 Différences Clés entre les Modèles Classiques et Versatile Diffusion

Aspect	Modèles classiques (Stable Diffusion, DALL-E)	Versatile Diffusion (VD)
Modalité	Spécialisés dans une seule modalité (texte ou image).	Multimodal : gère à la fois le texte et l'image dans un seul modèle.
Tâches prises en charge	Text-to-Image ou Image-to-Image uniquement.	Text-to-Image, Image-to-Text, Image-to-Image (et Text-to-Text dans certaines versions).
Architecture	Modèles distincts pour chaque tâche.	Architecture unifiée avec partage de paramètres.
Flexibilité	Limitée à une tâche spécifique.	Capable de passer d'une modalité à une autre de manière fluide.
Guidance	Souvent limité à un guidage textuel.	Guidance conditionnelle multimodale (texte et image).
Applications	Génération d'images uniquement.	Génération, modification et description d'images, ainsi que génération de texte.

FIGURE 1 – Tableau des différences entre les modèles classiques et Versatile Diffusion

3.2 La guidance conditionnel

Le guidage conditionnel est assuré par une stratégie de Classifier-Free Guidance. Cette méthode combine deux trajectoires : une trajectoire conditionnelle, influencée par un input (texte ou image), et une trajectoire non conditionnelle, générée sans contrainte. En ajustant un paramètre de guidance, VD équilibre la fidélité à l'input et la diversité des résultats. Par exemple, dans une tâche Text-to-Image, une description telle que "un chien jouant dans un parc" guidera la génération pour produire une image cohérente avec cette description.

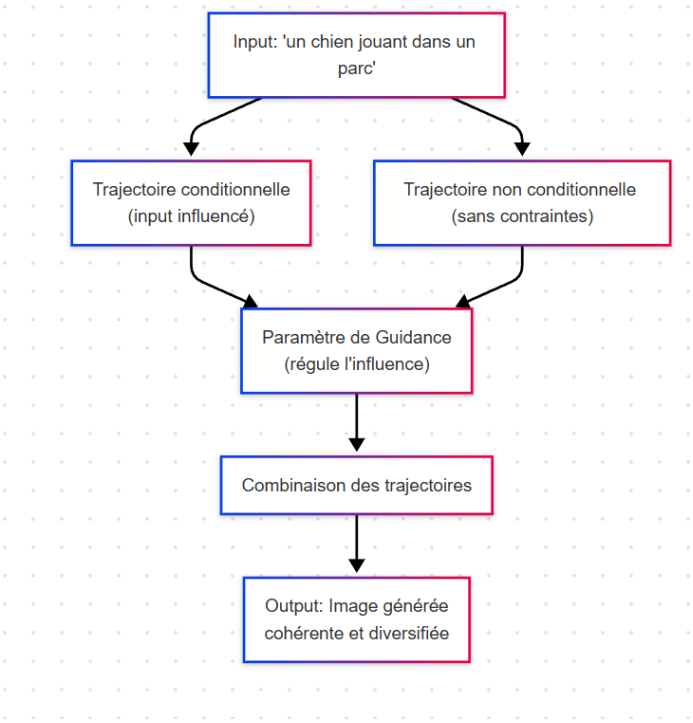

FIGURE 2 – Shema de la guidance conditionnel

3.3 Tâches ciblées et résultats

Le modèle VD peut s'appliquer à plusieurs tâches. Il permet de générer des images à partir de descriptions textuelles (Text-to-Image), de produire des descriptions textuelles à partir d'images (Image-to-Text), de générer des variantes d'une image en modifiant certains aspects tout en conservant sa sémantique (Image-Variation) et enfin de créer des reformulations ou des extensions de textes donnés (Text-Variation). Les résultats obtenus démontrent que VD est compétitif sur ces différentes tâches. Par ailleurs, cette approche ouvre la voie à des extensions innovantes, telles que la séparation du style et de la sémantique ainsi que la fusion contextuelle, offrant ainsi de nouvelles perspectives dans la génération multimodale.

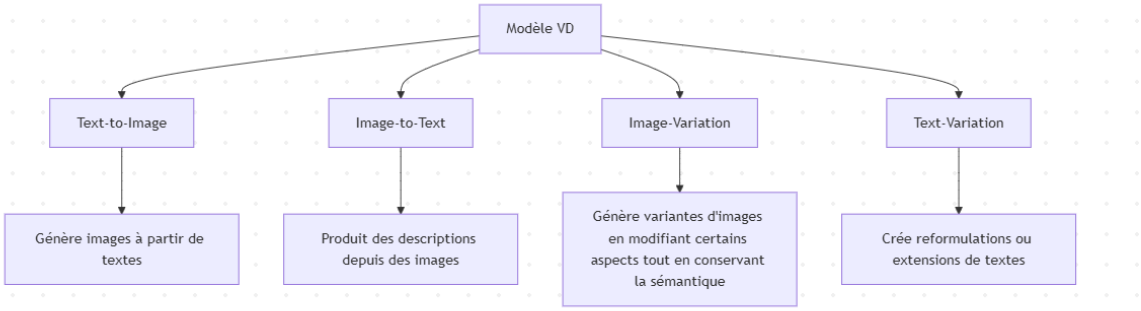

FIGURE 3 – Shema des différentes tâches du modèle VD

3.4 Architecture multi-Flux et partage de paramètres

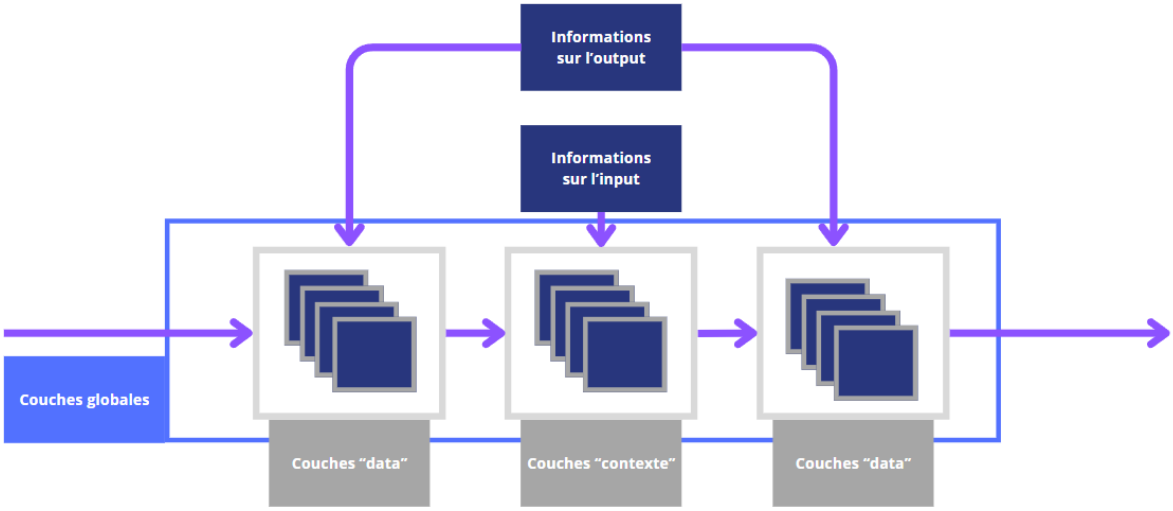

FIGURE 4 – Shema architecture multi-flux

Le schéma ci-dessus illustre le fonctionnement général de l'architecture multi-flux. Tout d'abord, la première couche «data» prépare la forme de ce que l'on souhaite créer. Par exemple, pour une image, cette couche initialise un tenseur rempli de zéros, et pour du texte, elle structure une chaîne de caractères adaptée. Cette préparation définit la base sur laquelle se construit la génération.

Ensuite, les couches ≪context≫ interviennent en apportant l'input, en précisant le type de donnée (texte ou image). Ces couches activent ainsi le parcours requis par le modèle pour traiter correctement le contenu fourni, en orientant le traitement selon la modalité de l'input.

Par la suite, une seconde couche ≪data≫ est dédiée à la préparation de la sortie, en fonction du résultat désiré. Elle ajuste les caractéristiques générées pour correspondre précisément au format voulu (qu'il s'agisse d'une image ou d'un texte), en activant uniquement les modules nécessaires pour cette tâche spécifique.

L'ensemble du processus repose sur un squelette de couches globales qui reste constant quel que soit le type de données utilisé. Ces couches communes réalisent les transformations essentielles – souvent dans un espace latent – permettant d'extraire et de moduler les caractéristiques nécessaires à la génération finale. Ainsi, selon la tâche effectuée, seules certaines parties de l'architecture sont activées tandis que le reste reste invariable, garantissant à la fois flexibilité et partage optimal des paramètres.

4 Implémentation

Le notebook `run_inference_analyse.ipynb` présente les capacités du modèle Versatile Diffusion (VD) à travers diverses tâches d'inférence multimodale. Le point d'entrée est la classe `VdInference`, qui permet de charger le modèle pré-entraîné et de configurer le sampler DDIM. Ce dernier réduit le nombre d'étapes nécessaires en réalisant une inversion du processus de diffusion de manière rapide et stable, notamment grâce à l'utilisation de la précision FP16 pour accélérer les calculs.

Une fois le modèle initialisé, plusieurs fonctions spécifiques sont mises à disposition pour traiter différents types de génération. Par exemple, la fonction `inference_t2i` (Texte vers Image) encode une description textuelle en un vecteur latent, lequel guide ensuite le sampler DDIM dans la génération d'une image correspondante. De même, la fonction `inference_i2i` (Image vers Image) permet de modifier une image existante. Dans ce cas, l'image d'entrée est d'abord transformée en vecteur latent, et en appliquant des ajustements contrôlés par des paramètres (comme les niveaux de fidélité ou de couleur), le modèle génère une nouvelle image tout en conservant la structure sémantique originale.

Le processus de transformation des images en descriptions textuelles est assuré par la fonction inference_i2t (Image vers Texte). Celle-ci encode l'image en un vecteur latent, qui est ensuite converti via le sampler DDIM en une représentation textuelle, puis décodé pour produire une description cohérente de l'image.

Au cœur de tous ces mécanismes se trouve le sampler DDIM. Ce dernier s'appuie sur des fonctions telles que `ddim_sampling` pour générer itérativement les vecteurs latents et `p_sample_ddim` pour réduire progressivement le bruit, assurant ainsi l'efficacité de l'inversion du processus de diffusion. La fonction `apply_model` intervient pour orchestrer l'ordre et l'interaction des différentes couches du modèle, garantissant un conditionnement adéquat des inputs – que ce soit du texte ou des images – avec le contexte nécessaire pour obtenir des sorties finales cohérentes.

Le graphique ci-dessous illustre visuellement l'ensemble du processus décrit, montrant comment les inputs sont encodés, transformés via le processus de diffusion et finalement décodés en images ou en textes, confirmant ainsi la flexibilité et la robustesse de l'approche VD.

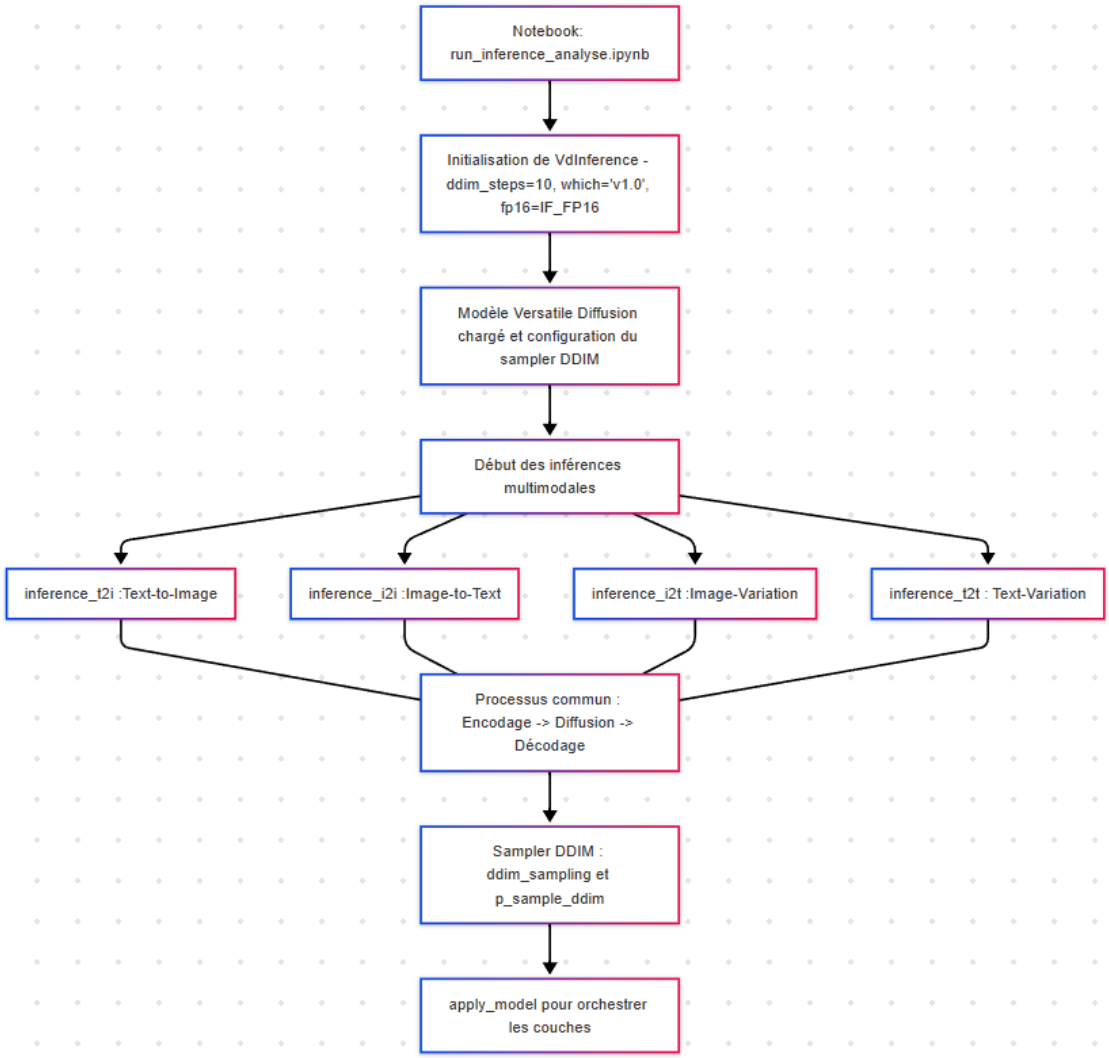

FIGURE 5 – Shema de l'implémentation

5 Conclusion

En résumé, ce rapport a permis de mettre en lumière les atouts et la flexibilité de Versatile Diffusion (VD), un modèle de génération multimodale capable de traiter simultanément plusieurs types de données, telles que le texte et les images. L'étude a démontré que VD offre une approche unifiée, contrastant avec les modèles traditionnels qui nécessitent souvent des architectures distinctes pour chaque modalité et chaque tâche.

Nous avons montré que l'utilisation du sampler DDIM permet d'accélérer le processus d'inférence en réduisant le nombre d'étapes nécessaires, tout en maintenant une qualité élevée des sorties grâce à un débruitage stable et efficace. Le mécanisme de guidage conditionnel, basé sur la stratégie de Classifier-Free Guidance, assure un équilibre judicieux entre la fidélité aux inputs (que ce soit des descriptions textuelles ou des images) et la diversité des résultats générés.

Par ailleurs, l'architecture multi-flux avec partage de paramètres permet d'optimiser la taille et l'efficacité du modèle, en activant dynamiquement des couches spécifiques en fonction des tâches (Text-to-Image, Image-to-Text, et Image-to-Image). Ce fonctionnement modulaire offre une grande flexibilité pour explorer diverses applications, allant de la génération d'images à partir d'un texte à la modification d'images existantes.

En conclusion, Versatile Diffusion se présente comme un outil puissant et innovant pour la génération multimodale, capable de produire des contenus cohérents et de haute qualité grâce à une intégration intelligente des techniques de diffusion, de guidage conditionnel et de conditionnement multimodal. Ces avancées ouvrent la voie à de nouvelles perspectives et applications dans le domaine de l'intelligence artificielle générative.