

DALL-E

La Théorie de la Diffusion Stable

Principes et mécanismes de la diffusion stable



Introduction aux Modèles de Diffusion

Les modèles de diffusion représentent une avancée majeure dans le domaine de la génération d'images par intelligence artificielle.

Contrairement aux GANs (Generative Adversarial Networks) qui ont dominé le domaine pendant des années, les modèles de diffusion utilisent un **processus progressif** qui ajoute puis retire du bruit pour générer des images de haute qualité.

Ces modèles, dont DALL-E est un exemple emblématique, ont révolutionné notre capacité à transformer des descriptions textuelles en représentations visuelles cohérentes et détaillées.



Le processus de diffusion transforme progressivement une distribution de bruit aléatoire en une image structurée, guidée par une description textuelle.

Principe Fondamental: Le Processus de Diffusion

La diffusion stable repose sur un principe fondamental: il est *plus facile d'ajouter du bruit à une image que de l'enlever.*

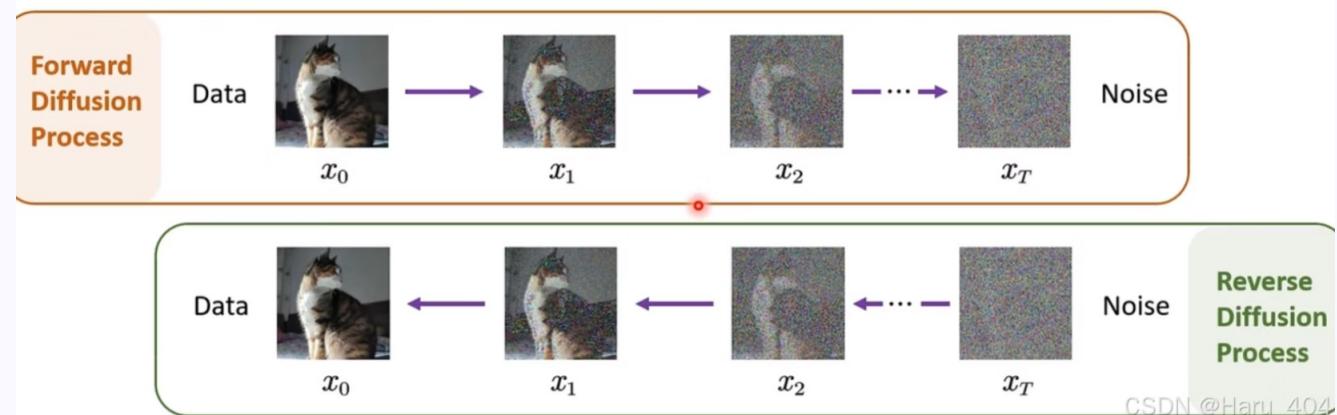
Le modèle exploite cette asymétrie en apprenant à inverser un processus de bruitage progressif.

Durant l'entraînement : le modèle observe des images auxquelles on ajoute progressivement du bruit selon un programme prédéfini. Il apprend alors à prédire le bruit qui a été ajouté à chaque étape, acquérant ainsi la capacité de "débruiter" une image.

Lors de l'inférence : le processus est inversé : partant d'un bruit gaussien pur, le modèle retire progressivement le bruit qu'il prédit, faisant émerger une image structurée guidée par l'embedding textuel.

What is Diffusion Model ?

Denoising Diffusion Probabilistic Models^[1]



Visualisation du processus de diffusion: en haut, l'ajout progressif de bruit (forward process); en bas, le processus de débruitage (reverse process) qui génère l'image finale.

Mathématiques de la Diffusion

Processus Forward (Bruitage)

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$$

Où:

- x_t est l'image bruitée à l'étape t
- β_t est le coefficient de bruitage à l'étape t
- N représente une distribution normale

Processus Reverse (Débruitage)

$$p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

Où μ_θ et Σ_θ sont des fonctions paramétrées par un réseau de neurones.

Objectif d'Entraînement

Le modèle est entraîné à minimiser la divergence entre la distribution réelle et la distribution prédictive:

$$L = E_{t, x_0, \epsilon} [\| \epsilon_\theta(x_t, t, c) \|_2^2]$$

Où:

- ϵ est le bruit réel ajouté
- ϵ_θ est le bruit prédict par le modèle
- c est le conditionnement textuel

Cette formulation permet au modèle d'apprendre à prédire et donc à inverser le processus de bruitage, tout en étant guidé par une description textuelle.

3 éléments majeurs pour de la Diffusion Stable

1

Processeur de Texte (CLIP)

Le modèle CLIP (Contrastive Language-Image Pre-training) développé par OpenAI joue un rôle crucial dans la compréhension des prompts textuels. Il transforme les descriptions en représentations vectorielles qui guideront le processus de génération d'images.

CLIP a été entraîné sur des millions de paires texte-image provenant d'internet, lui permettant de comprendre les relations sémantiques entre le langage et les représentations visuelles.

2

Générateur d'Images (Modèle de Diffusion)

Le cœur du système est un modèle de diffusion qui commence par du bruit aléatoire et le transforme progressivement en une image cohérente. Ce processus inverse une chaîne de Markov qui aurait normalement dégradé une image en bruit.

Le modèle apprend à prédire et inverser les étapes de bruitage, guidé par l'embedding textuel fourni par CLIP.

3

Scheduler (~50 étapes)

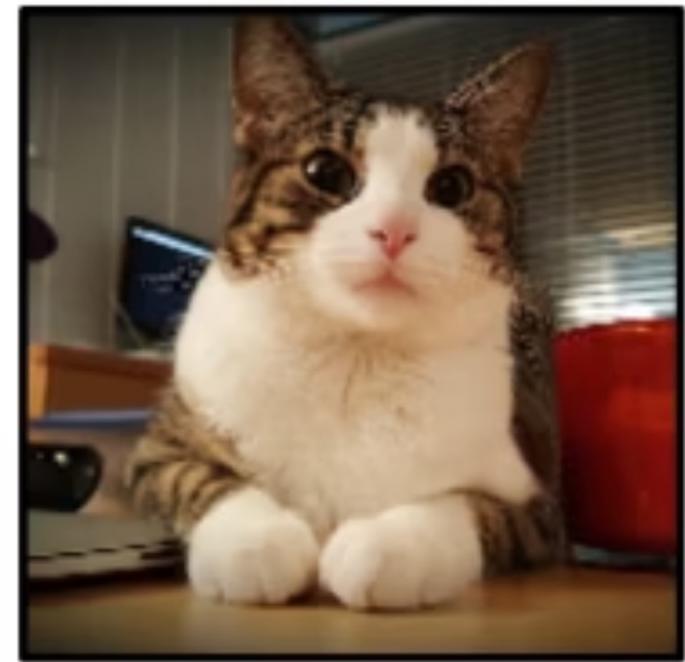
Le scheduler contrôle la progression du processus de débruitage, déterminant comment et à quelle vitesse le bruit est retiré à chaque étape. Typiquement, environ 50 étapes sont nécessaires pour obtenir une image de haute qualité.

Différents schedulers (DDIM, PNDM, etc.) offrent des compromis entre vitesse de génération et qualité d'image.

**Un chat blanc
tigré en gros plan
sur un bureau**

**Processeur
de texte
(CLIP)**

**Générateur d'images
(Modèle de diffusion)**



Architecture CLIP: Le Pont Entre Texte et Image

Principe de Fonctionnement

CLIP (Contrastive Language-Image Pre-training) est un modèle développé par OpenAI qui crée un espace d'embedding unifié pour le texte et les images. Il est entraîné à maximiser la similarité entre les paires texte-image correspondantes tout en minimisant la similarité pour les paires non correspondantes.

Cette approche contrastive permet à CLIP de développer une compréhension sémantique profonde des relations entre le langage et les représentations visuelles, sans nécessiter d'annotations manuelles détaillées.

Rôle dans la Diffusion Stable

Dans le contexte de DALL-E et autres modèles de diffusion, CLIP joue un rôle crucial:

- Il transforme le prompt textuel en un embedding vectoriel dense
- Cet embedding guide le processus de débruitage à chaque étape
- Il permet d'évaluer la correspondance entre l'image générée et le prompt

La richesse sémantique des embeddings CLIP est une des raisons principales de la capacité des modèles de diffusion à générer des images fidèles à des descriptions textuelles complexes et nuancées.

Processeur de texte (CLIP)

Image



Légende

Corgi sur un rocher



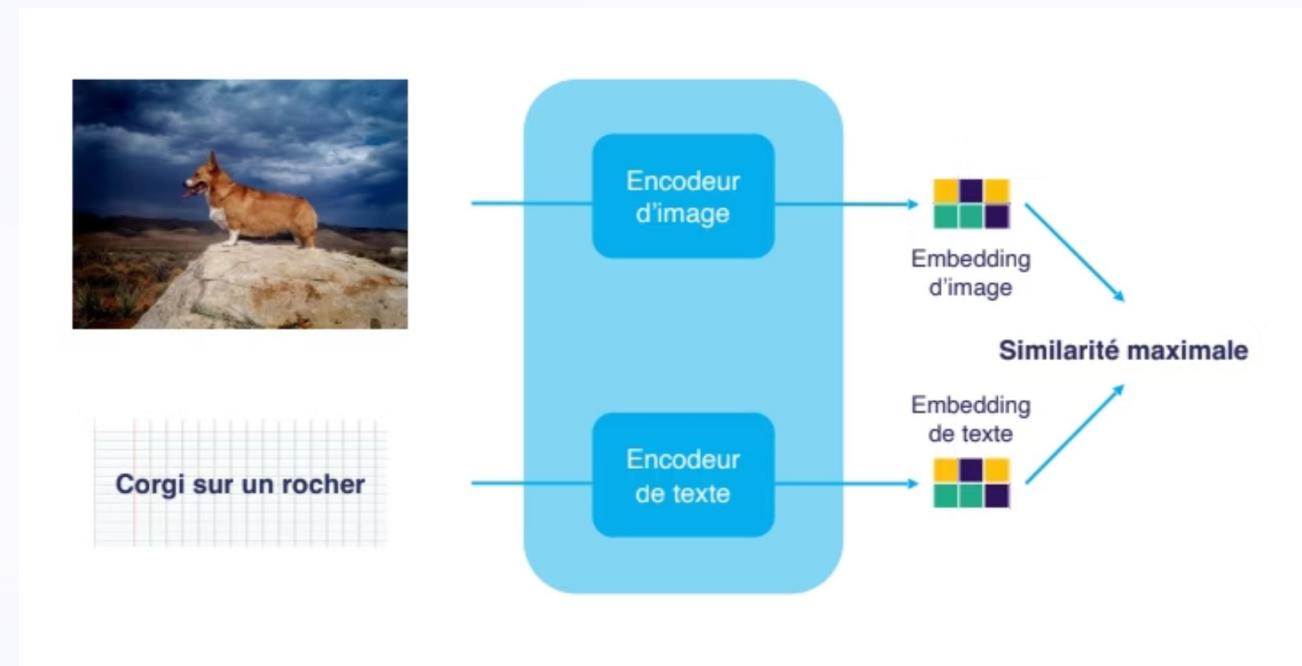
L'aigle royal,
journal de Capestan

Collecte de paires image/légende, qui est une étape essentielle dans le processus de création de modèles de diffusion.

Ces paires servent de base d'apprentissage pour le modèle, lui permettant d'associer visuellement le contenu des images avec les légendes correspondantes.

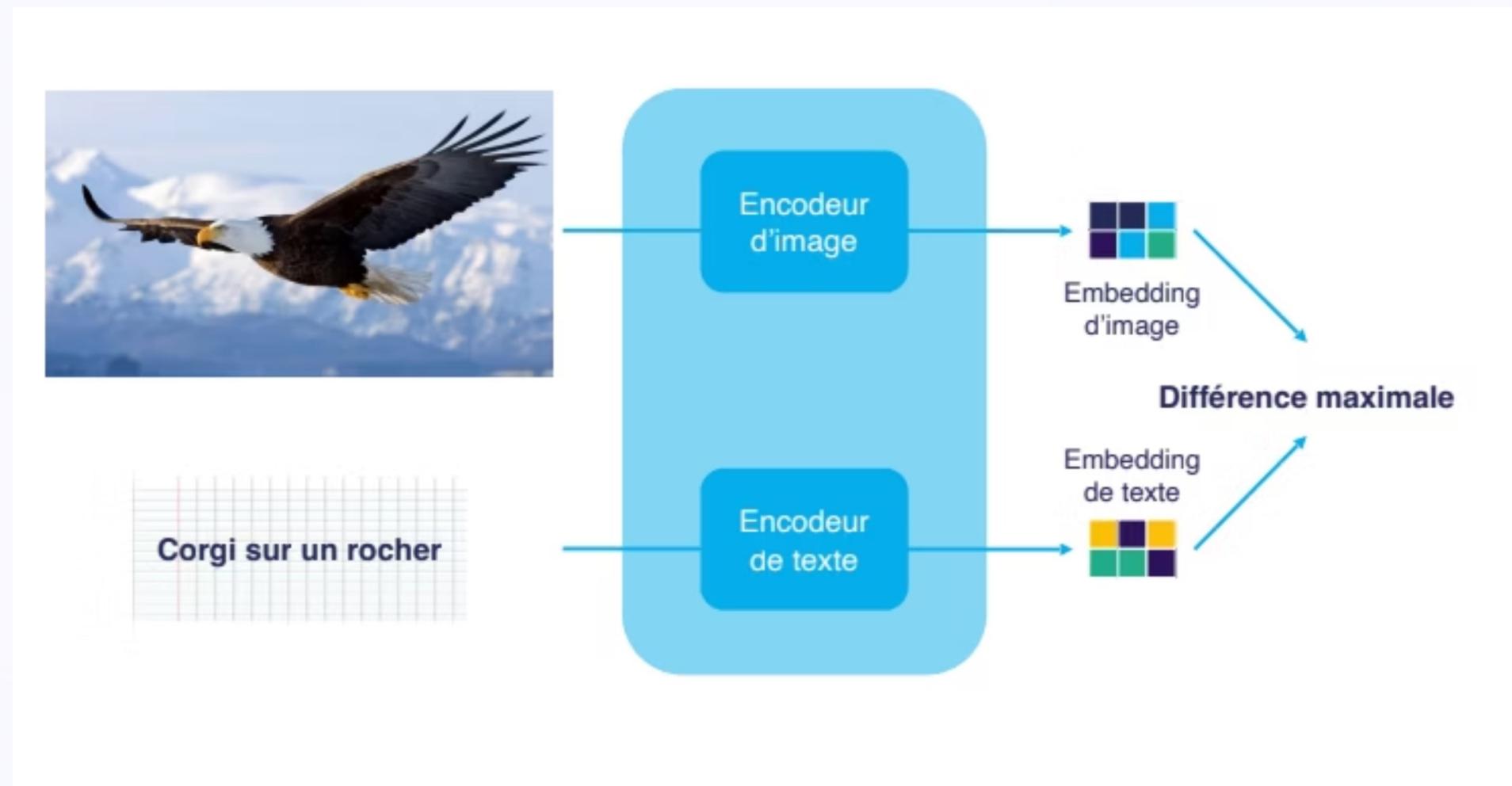
C'est grâce à cette mise en correspondance que le modèle pourra par la suite générer des images à partir de descriptions textuelles.

Processeur de texte (CLIP)



- Un encodeur pour le texte et un encodeur pour les images.
- L'objectif est d'obtenir une correspondance maximale entre les paires texte-image, afin que le modèle puisse apprendre à générer des images à partir de descriptions textuelles.
- En entraînant ces deux encodeurs de manière conjointe, on cherche à maximiser la similarité entre les représentations des paires texte-image authentiques.
- C'est cette correspondance étroite qui permettra au modèle de diffusion de générer des images cohérentes et fidèles aux descriptions fournies.

Processeur de texte (CLIP)

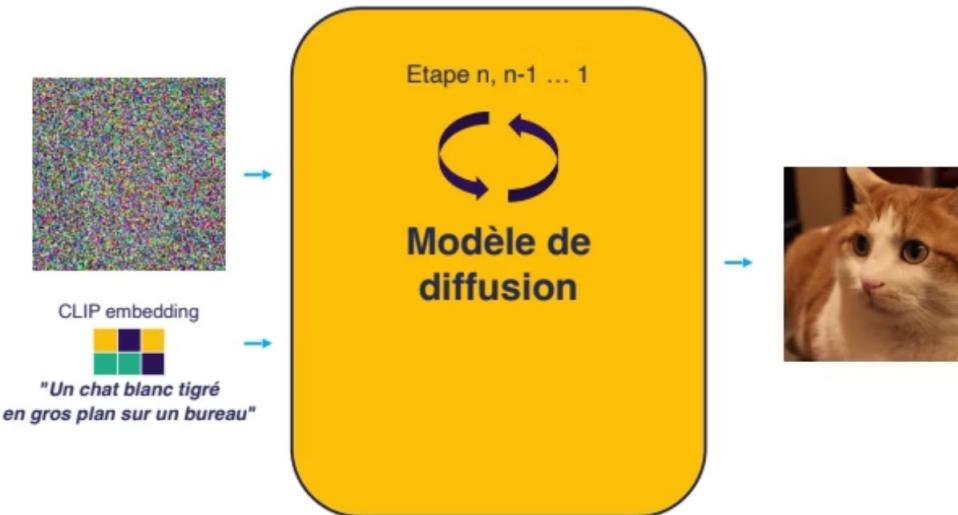


Mais on cherche aussi à maximiser la différence entre les embeddings sur les paires fictives.

Architecture Détailée du Modèle de Diffusion

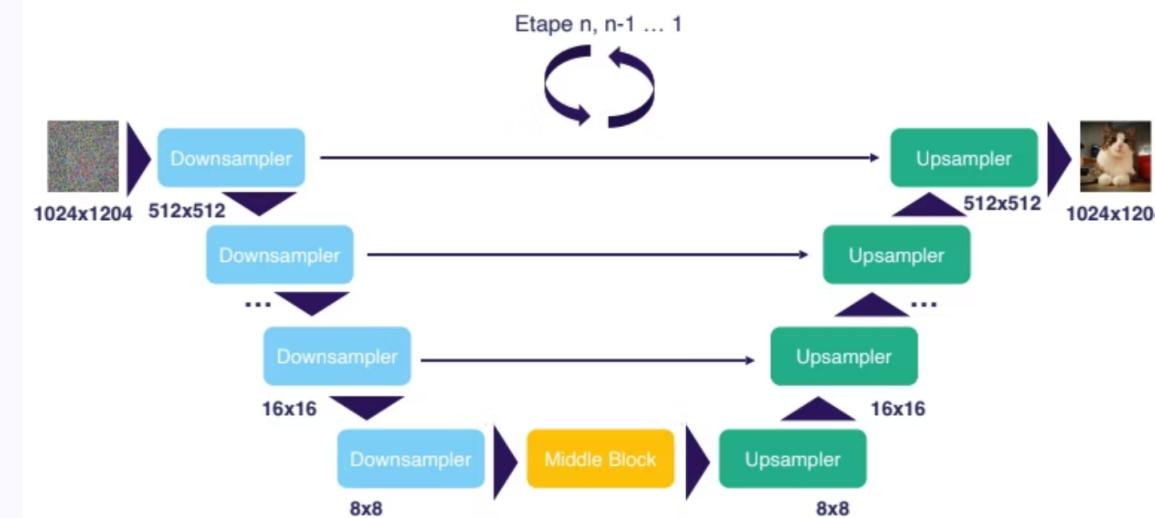
U-Net avec Attention

- Le cœur du modèle de diffusion est généralement un réseau U-Net modifié avec des mécanismes d'attention. Cette architecture permet de capturer des dépendances à longue distance dans l'image tout en préservant les détails locaux.
- Les blocs d'attention permettent d'intégrer le conditionnement textuel à différentes échelles de résolution, guidant ainsi le processus de débruitage.



Downsamplers

- Permettent de réduire progressivement la résolution des images tout au long du processus de génération.
- Cette réduction progressive de la résolution est cruciale pour permettre au modèle de capturer les structures à grande échelle de l'image, tout en générant graduellement les détails fins.
- Agissent comme une sorte d'entonnoir, forçant le modèle à se concentrer d'abord sur les éléments les plus importants avant d'ajouter les détails.

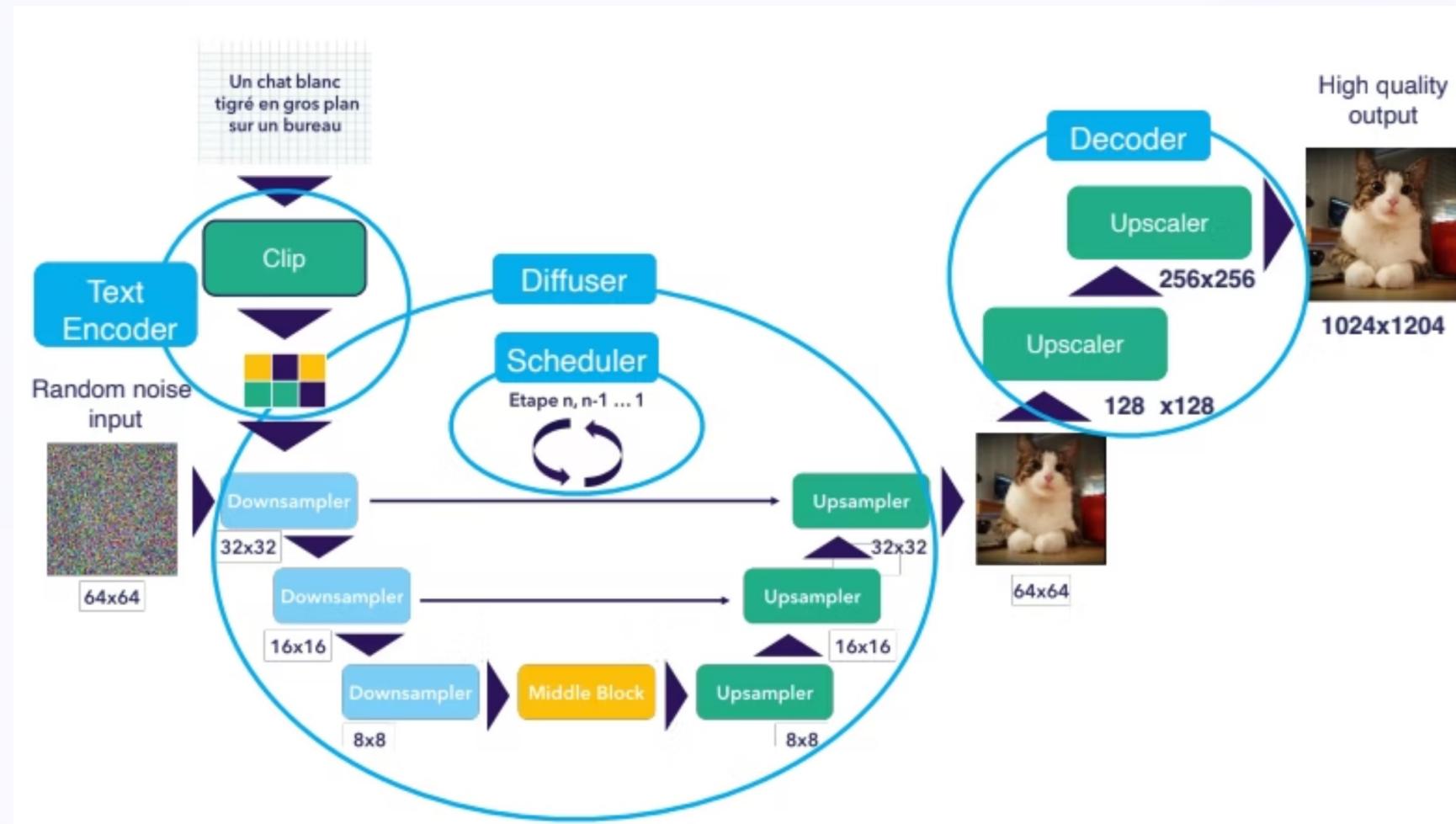




Étapes du Processus de Génération

-  **Initialisation avec du Bruit Aléatoire**
Le processus commence par générer une matrice de bruit gaussien de la taille de l'image souhaitée. Cette distribution aléatoire constitue le point de départ du processus de débruitage.
-  **Encodage du Prompt Textuel**
Le prompt textuel est traité par le modèle CLIP pour produire un embedding vectoriel qui capture la sémantique de la description. Cet embedding servira à guider le processus de génération.
-  **Débruitage Progressif**
À chaque étape t , le modèle prédit le bruit présent dans l'image bruitée x_t , conditionnellement à l'embedding textuel. Cette prédiction permet de calculer une estimation de x_{t-1} moins bruitée.
-  **Itération jusqu'à l'Image Finale**
Le processus se répète typiquement pour environ 50 étapes, chacune retirant progressivement du bruit jusqu'à obtenir une image claire correspondant à la description textuelle.

Architecture Détailée du Modèle de Diffusion



Ensemble, le diffuseur et le scheduler offrent une solution élégante pour générer des images de haute qualité, tout en réduisant considérablement les coûts de calcul par rapport à d'autres approches. C'est un élément clé de l'architecture des modèles de diffusion.

Diffuseur

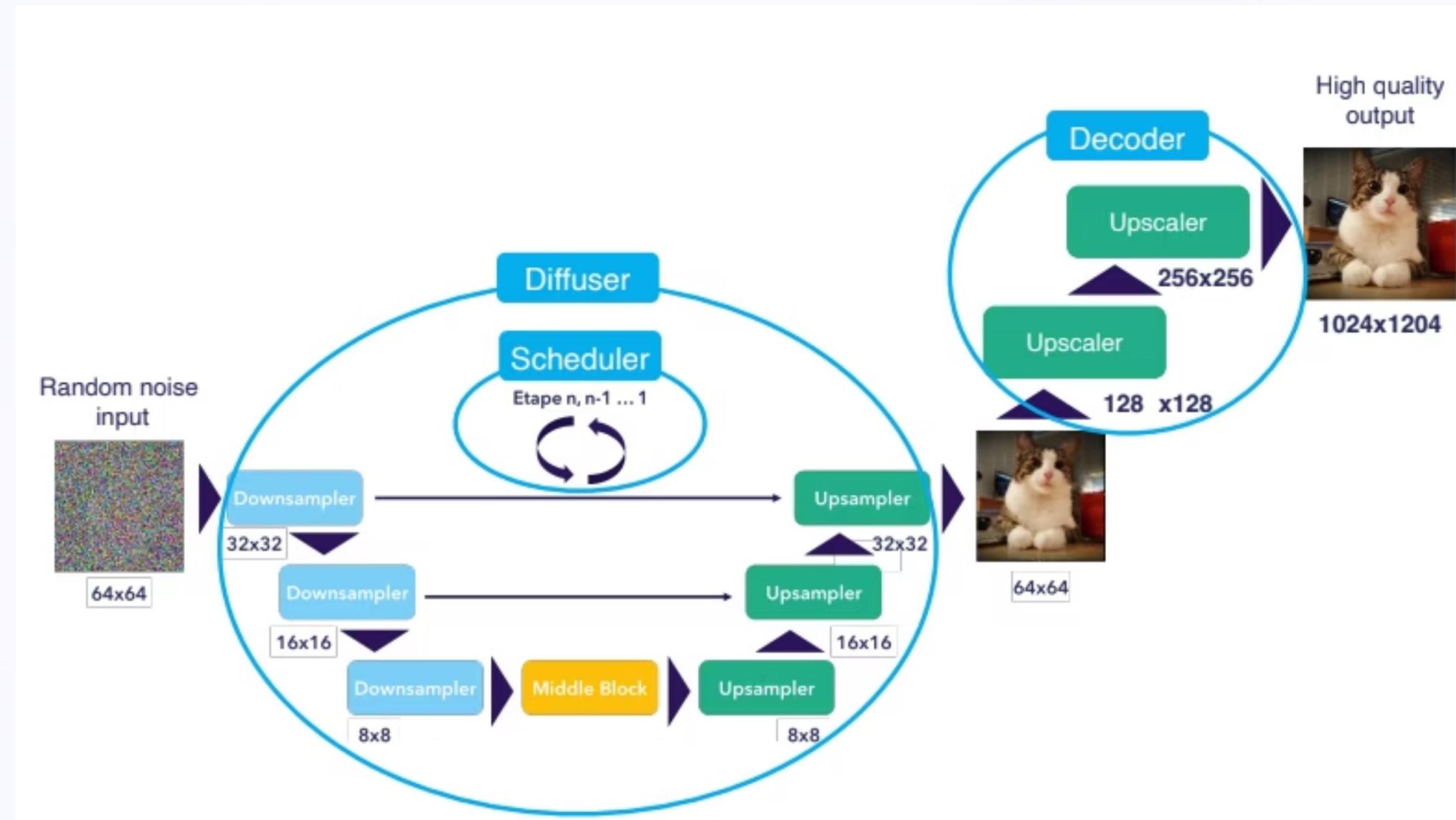
Chargé de la transformation progressive de l'espace latent, en réduisant petit à petit le niveau de bruit. Cette approche par étapes permet de capturer efficacement les structures complexes de l'image, tout en générant les détails fins de manière contrôlée.

Scheduler

Gère le rythme de cette transformation, en déterminant le nombre d'étapes et la vitesse de débruitage.

Un scheduler bien conçu permet de trouver un équilibre optimal entre la qualité de l'image générée et les ressources de calcul nécessaires.

Architecture Détailée du Modèle de Diffusion



Décodeur

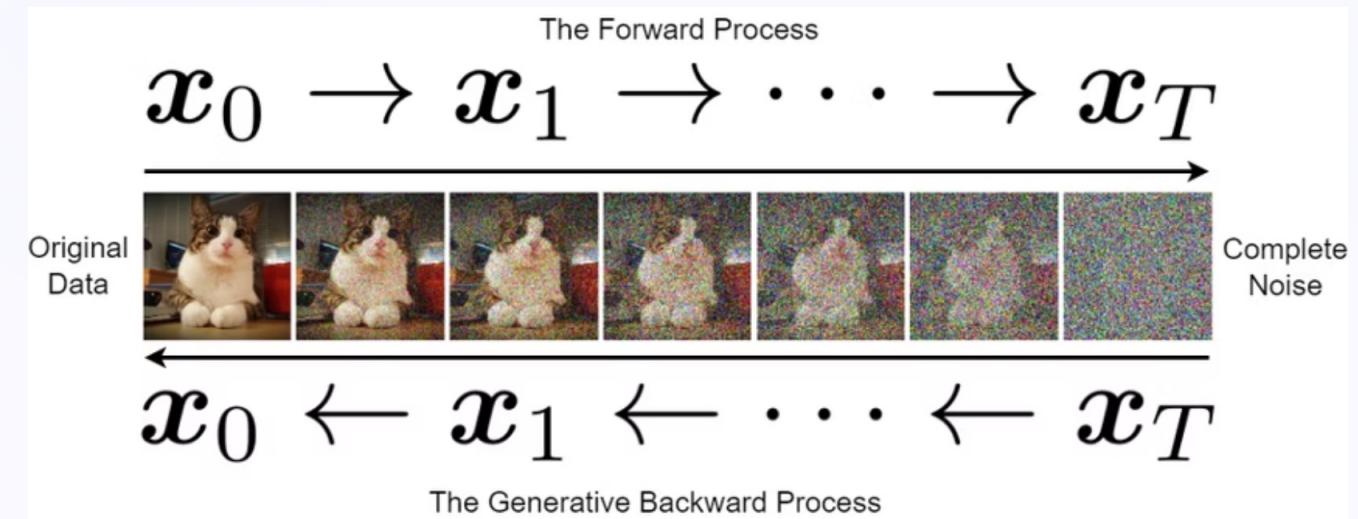
Responsable de la reconstruction progressive de l'image finale à partir des représentations intermédiaires générées tout au long du processus de débruitage.

Le choix de l'architecture du décodeur, de ses couches et de ses connexions, a un impact majeur sur les performances du modèle de diffusion.

Visualisation du Processus de Débruitage

Cette séquence illustre le processus de débruitage progressif, partant d'un bruit gaussien pur (à droite) et aboutissant à une image cohérente (à gauche). Observez comment les structures globales émergent d'abord, suivies par les détails fins.

Ce processus est guidé à chaque étape par l'embedding textuel qui oriente la génération vers une représentation visuelle correspondant à la description fournie.



Rôle du Scheduler dans la Diffusion

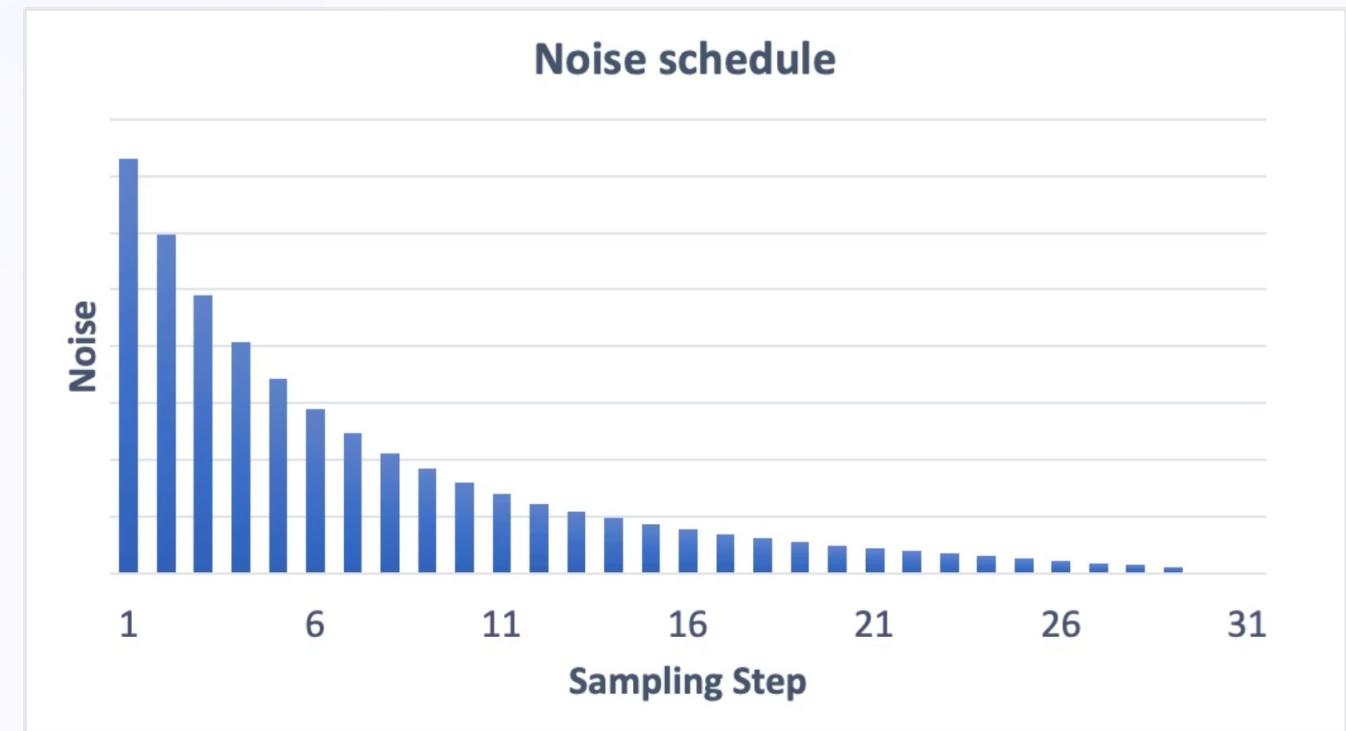
Fonction du Scheduler

Le scheduler détermine la séquence des coefficients de bruitage β_t et leurs dérivés, contrôlant ainsi la progression du processus de diffusion.

Différents schedulers peuvent produire des résultats visuellement distincts et affecter la stabilité du processus de génération.

Types de Schedulers

- **DDPM** (Diffusion probabilistic model): Le scheduler original, stable mais lent (nécessite ~1000 étapes)
- **DDIM** (Denoising Diffusion Implicit Model): Permet un échantillonnage déterministe et accéléré (~50 étapes)
- **PNDM**([Pseudo Numerical Methods for Diffusion Models on Manifolds](#)) : Utilise des méthodes numériques avancées pour améliorer la qualité
- **Euler a**: Offre un bon compromis entre vitesse et qualité



Evolution de la variance du bruit au cours des étapes de diffusion (30 étapes). Le choix du scheduler affecte significativement la qualité et l'efficacité du processus de génération.

Les schedulers modernes permettent de réduire considérablement le nombre d'étapes nécessaires (de ~1000 à ~50) tout en maintenant une qualité d'image élevée, rendant la génération d'images par diffusion beaucoup plus accessible en termes de ressources computationnelles.

Comparaison des Schedulers

Comparaison des Performances

Scheduler	Étapes typiques	Qualité	Déterminisme
DDPM	1000+	Excellente	Non
DDIM	50-100	Très bonne	Oui
Euler a	30-50	Bonne	Non
DPM-Solver	20-30	Bonne	Configurable

Le choix du scheduler représente un compromis entre vitesse, qualité et déterminisme. Les schedulers modernes comme DPM-Solver permettent une génération rapide tout en préservant une qualité d'image acceptable, rendant la diffusion stable plus accessible pour les applications en temps réel.

Input image



annotation
→
(canny edge
detector)



Stable
Diffusion



Conditionnement et Guidance

Conditionnement Spatial

Des extensions comme ControlNet permettent un conditionnement spatial précis, guidant la génération avec des cartes de profondeur, des poses, des croquis, etc.

Ces techniques offrent un contrôle beaucoup plus fin sur la composition de l'image générée, tout en préservant la cohérence avec le prompt textuel.

Classifier-Free Guidance

Une technique clé qui améliore l'adhérence au prompt textuel sans nécessiter un classificateur externe. Elle fonctionne en interpolant entre:

$$\tilde{\epsilon}_\theta(x_t, t, c) = w \epsilon_\theta(x_t, t, c) + (1 - w) \epsilon_\theta(x_t, t, \emptyset)$$

Où $w > 1$ est le poids de guidance et \emptyset représente un conditionnement vide.

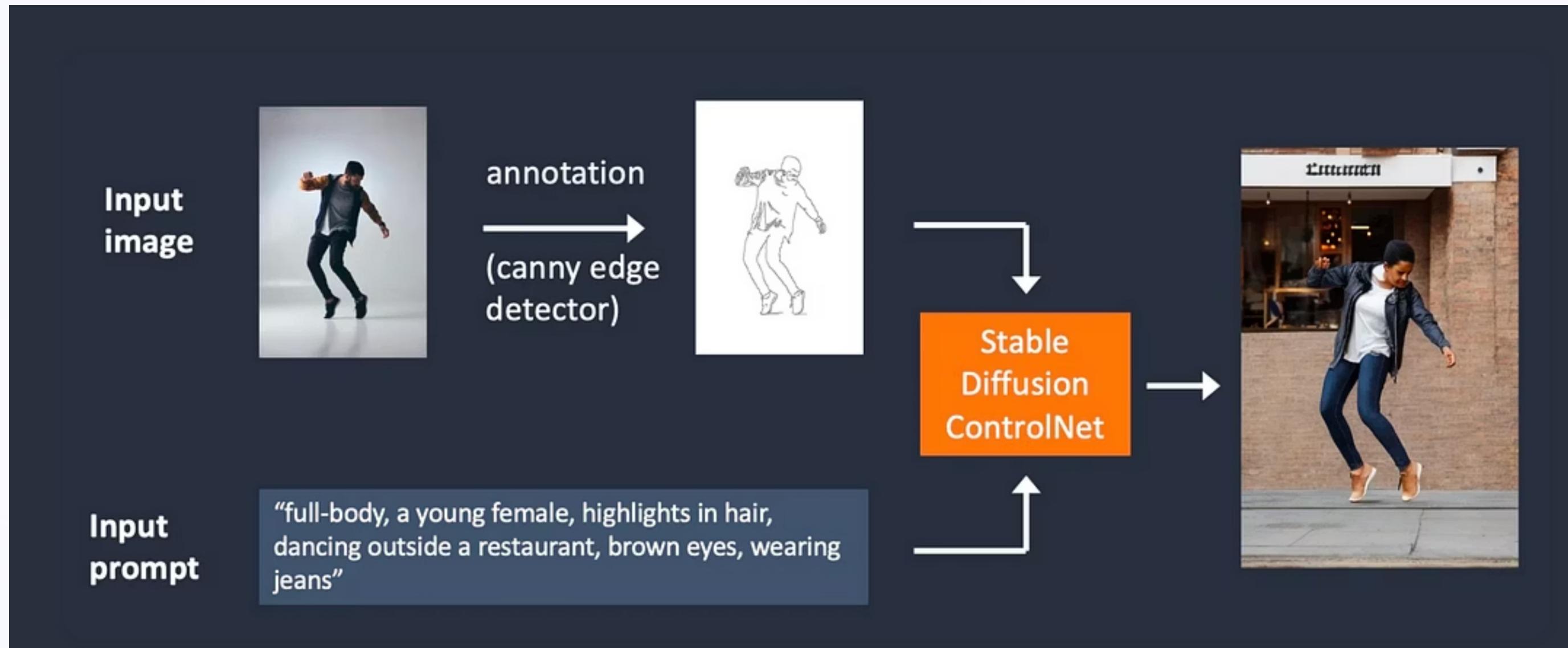
Des valeurs de w plus élevées produisent des images plus fidèles au prompt mais potentiellement moins diversifiées et naturelles.

Inversion et Édition

Des méthodes comme Textual Inversion permettent d'encoder des concepts visuels spécifiques dans l'espace d'embedding textuel, facilitant la génération d'images personnalisées.

Ces techniques ouvrent la voie à l'édition sémantique d'images existantes via le processus de diffusion.

Control Net





"Give him a cowboy hat"



"Give him a mustache"



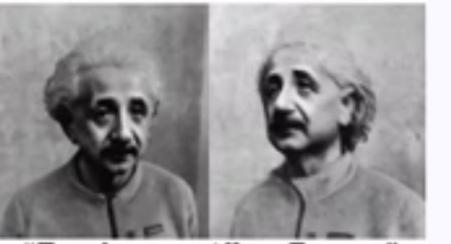
"Make him bald"



"Turn him into a clown"



"As a bronze bust"



"Turn him into Albert Einstein"



"Turn his face into a skull"

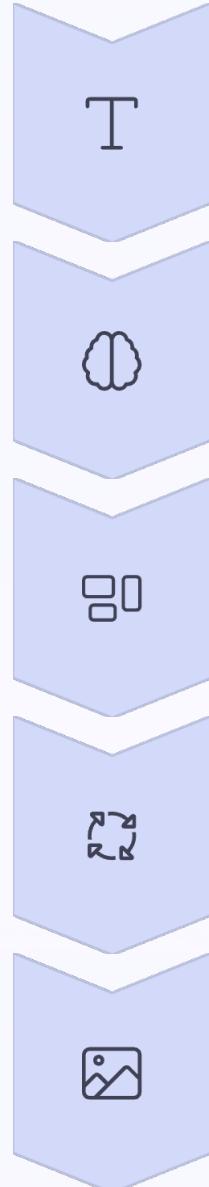


"Turn him into a Modigliani painting"



"Turn him into Batman"

Intégration de CLIP et du Modèle de Diffusion



Prompt Textuel

L'utilisateur fournit une description textuelle de l'image souhaitée, comme "un corgi sur un rocher".

Encodage CLIP

Le modèle CLIP transforme cette description en un vecteur d'embedding de haute dimension qui capture la sémantique du prompt.

Initialisation du Bruit

Une matrice de bruit gaussien est générée comme point de départ du processus de diffusion.

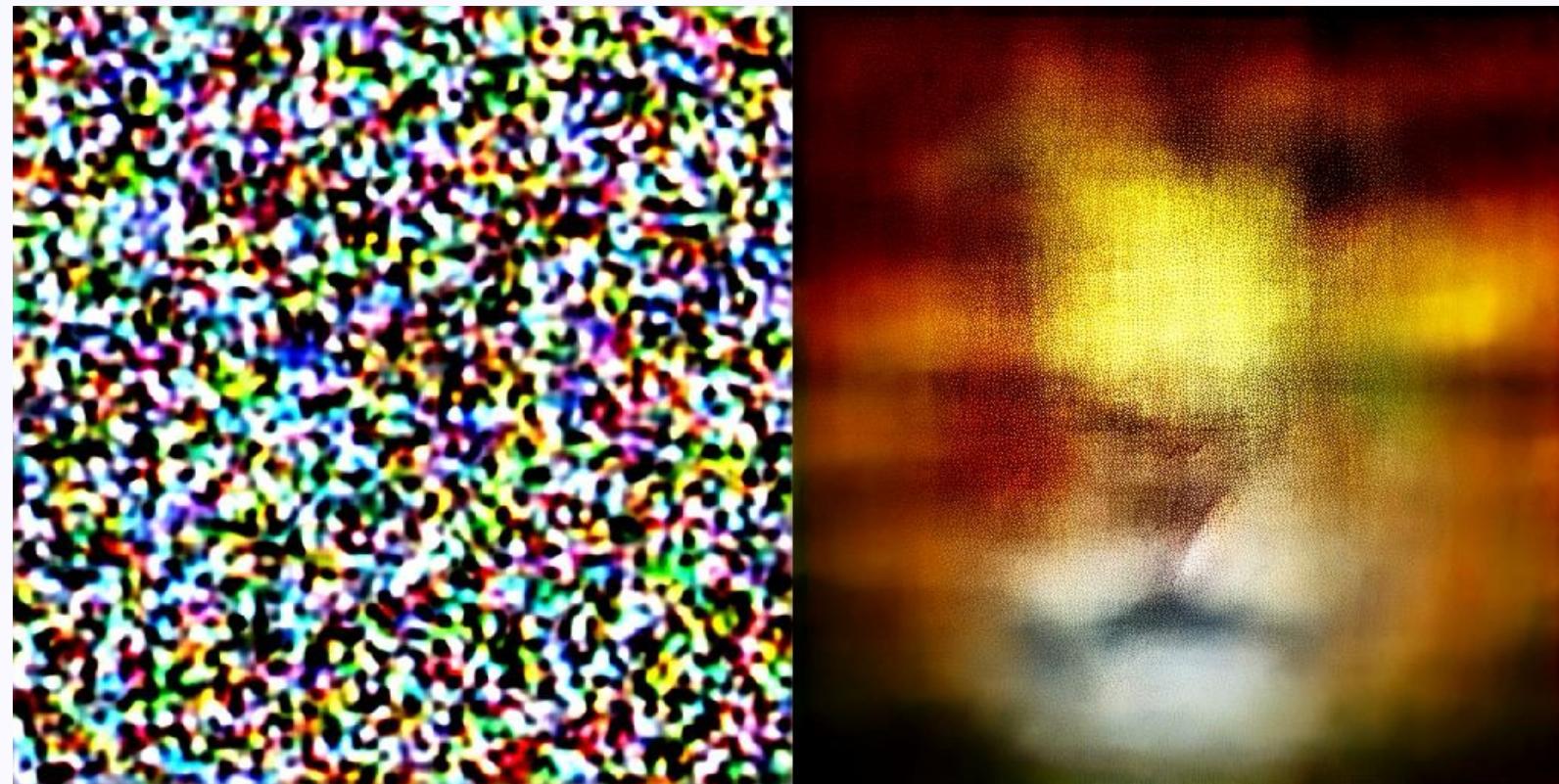
Débruitage Conditionné

Le modèle de diffusion prédit et retire progressivement le bruit, guidé par l'embedding CLIP à chaque étape.

Image Finale

Après ~50 étapes, une image claire émerge, correspondant sémantiquement au prompt textuel original.

Exemple de Génération: "The face of a lion, photography"



Analyse du Processus

Cette séquence illustre plusieurs aspects clés de la diffusion stable:

1. La composition globale émerge dans les dernières étapes ($t=50-40$)
2. Les formes distinctes lion se définissent ($t=40-20$)
3. Les détails comme la fourrure apparaissent ($t=20-10$)
4. Les nuances fines et les détails de surface sont affinés ($t=10-1$)

Visualisation des étapes intermédiaires du processus de diffusion pour le prompt "The face of a lion, photography". Observez comment les formes générales apparaissent d'abord, suivies par les détails spécifiques du chien et de son environnement.

Variations Stylistiques et Sémantiques



Variation générée en modifiant le prompt pour inclure différents styles artistiques ou contextes sémantiques, tout en conservant le sujet principal "corgi sur un rocher, digital art".



"Corgi sur un rocher, a la Van gogh"

Contrôle par le Prompt

Les modèles de diffusion comme DALL-E permettent un contrôle remarquable sur le style et le contenu via des modifications du prompt textuel:

- **Styles artistiques:** "dans le style de Van Gogh", "art digital", "photographie HDR"
- **Contextes:** "au coucher du soleil", "dans un paysage montagneux", "sous la pluie"
- **Aspects techniques:** "gros plan", "vue aérienne", "objectif grand angle"

Cette flexibilité découle de la richesse sémantique des embeddings CLIP et de la capacité du modèle de diffusion à interpréter ces nuances lors du processus de débruitage.

Défis Spécifiques: Génération de Mains

Problématique des Mains

La génération de mains humaines représente un défi particulier pour les modèles de diffusion, souvent caractérisé par:

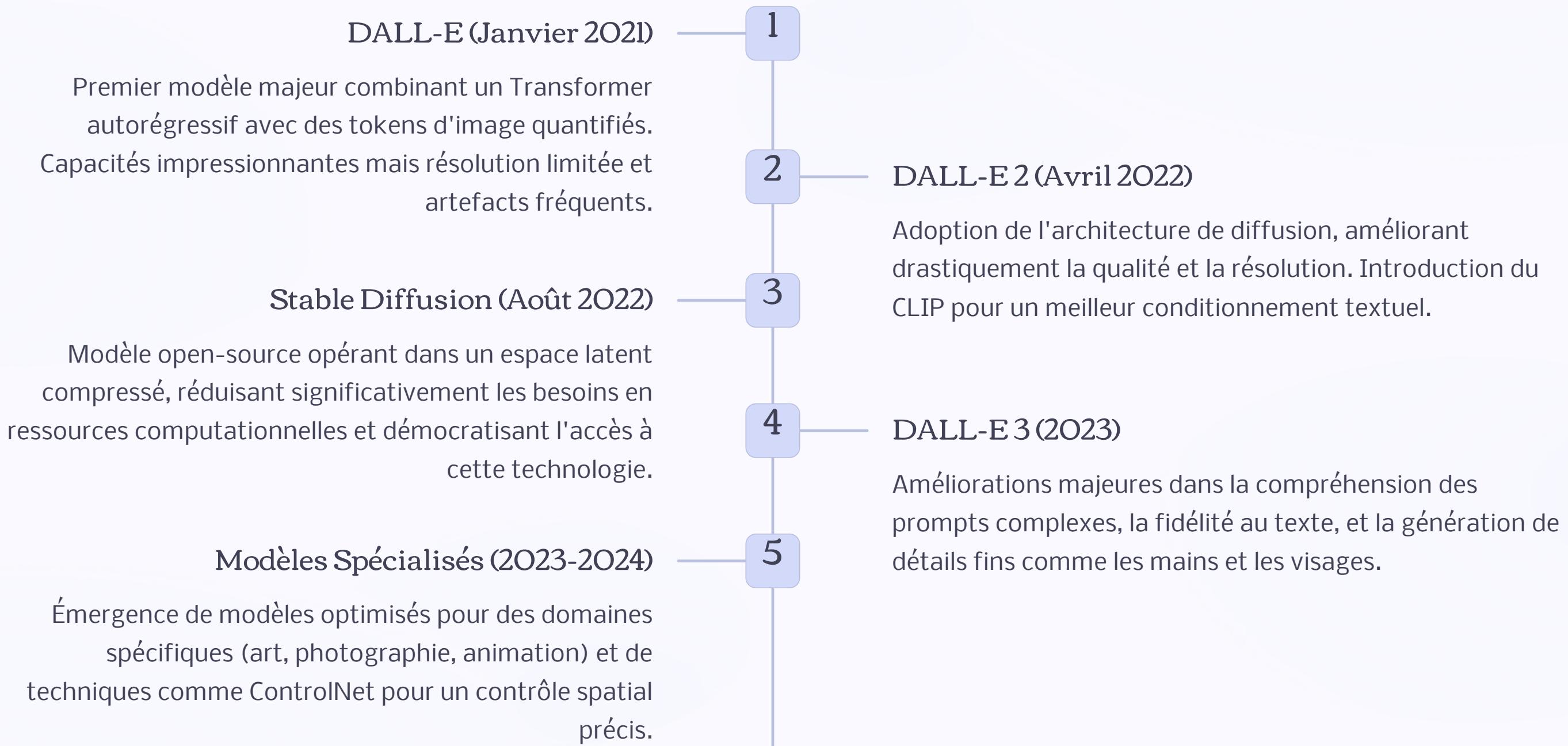
- Nombres incorrects de doigts
- Proportions anatomiques irréalistes
- Positions impossibles ou incohérentes

Ces difficultés proviennent de la complexité anatomique des mains, de leur grande variabilité de poses, et potentiellement d'une sous-représentation dans les données d'entraînement.



Exemples de générations de mains présentant diverses anomalies anatomiques, illustrant les limitations actuelles des modèles de diffusion pour certaines structures complexes.

Évolution et Améliorations des Modèles



Conclusion et Perspectives Futures

Synthèse

Les modèles de diffusion stable représentent une avancée fondamentale dans la génération d'images, offrant:

- Une qualité et une cohérence visuelle sans précédent
- Une flexibilité remarquable dans le conditionnement
- Un cadre mathématique élégant et bien fondé

Leur succès repose sur la combinaison de plusieurs innovations clés: le processus de diffusion progressive, l'encodage sémantique via CLIP, et les techniques avancées de conditionnement et d'échantillonnage.

Directions Futures

Plusieurs axes de développement prometteurs se dessinent:

- Génération vidéo par extension temporelle des modèles de diffusion
- Modèles 3D pour la création de scènes et objets tridimensionnels
- Contrôle plus fin et intuitif sur les générations
- Réduction des besoins computationnels pour des applications en temps réel
- Intégration avec d'autres modalités (audio, texte, interaction)

Ces avancées ouvrent des perspectives fascinantes pour la créativité assistée par IA et la démocratisation de la création visuelle.