Modèles de Diffusion

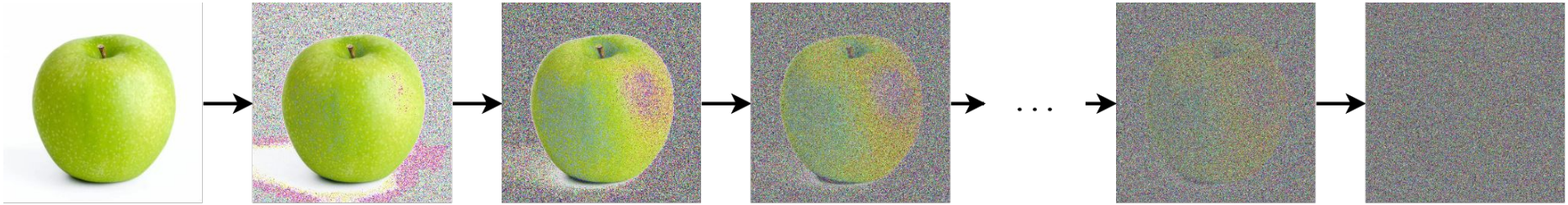
BONHOMME Romain
PIERRE Romain
EL MAZOUGUI Khawla
LAUGT Victor

Plan

- Modèle de Diffusion Classique
- Modèle de Diffusion Latent
- Conditionnement
- Optimisation de l’inférence

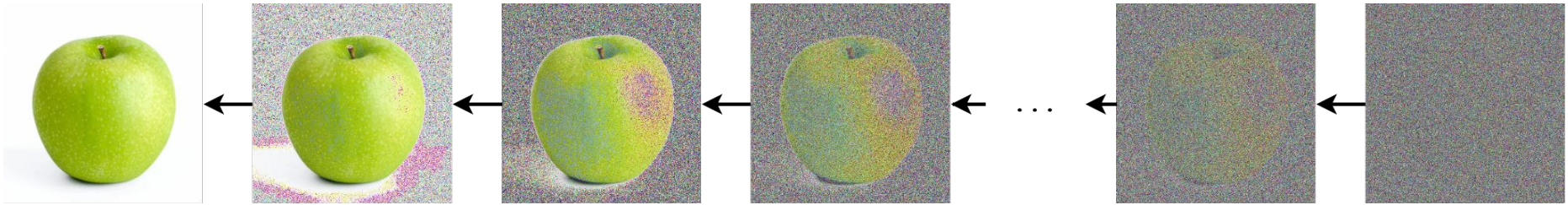
Modèle de Diffusion Classique

Diffusion



$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}z_t \qquad z_t \hookrightarrow \mathcal{N}(0, I)$$

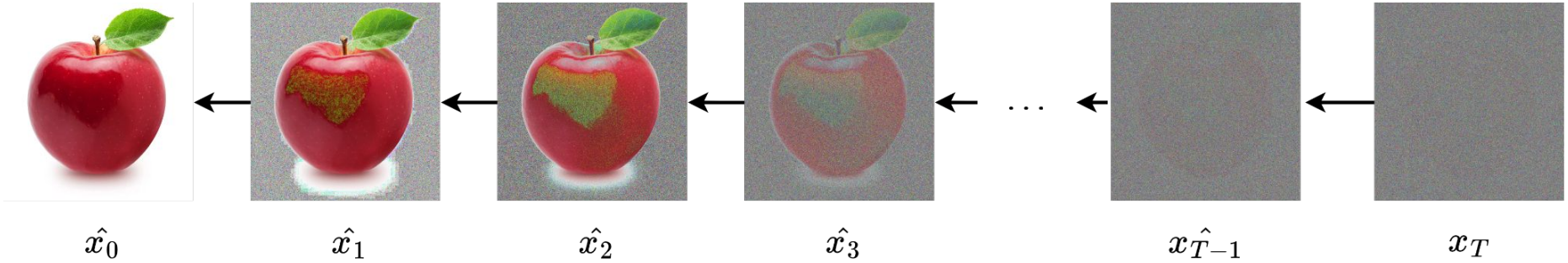
Denoising



$$\hat{x_{t-1}} = \text{UNet}(x_t, t) \quad \underset{\text{entraînement}}{\longrightarrow} \quad x_{t-1}$$

$$\text{minimiser} \quad \text{MSE}(\hat{x_{t-1}}, x_{t-1})$$

Génération d'une image



$x \leftarrow \text{tirage dans } \mathcal{N}(0, I)$

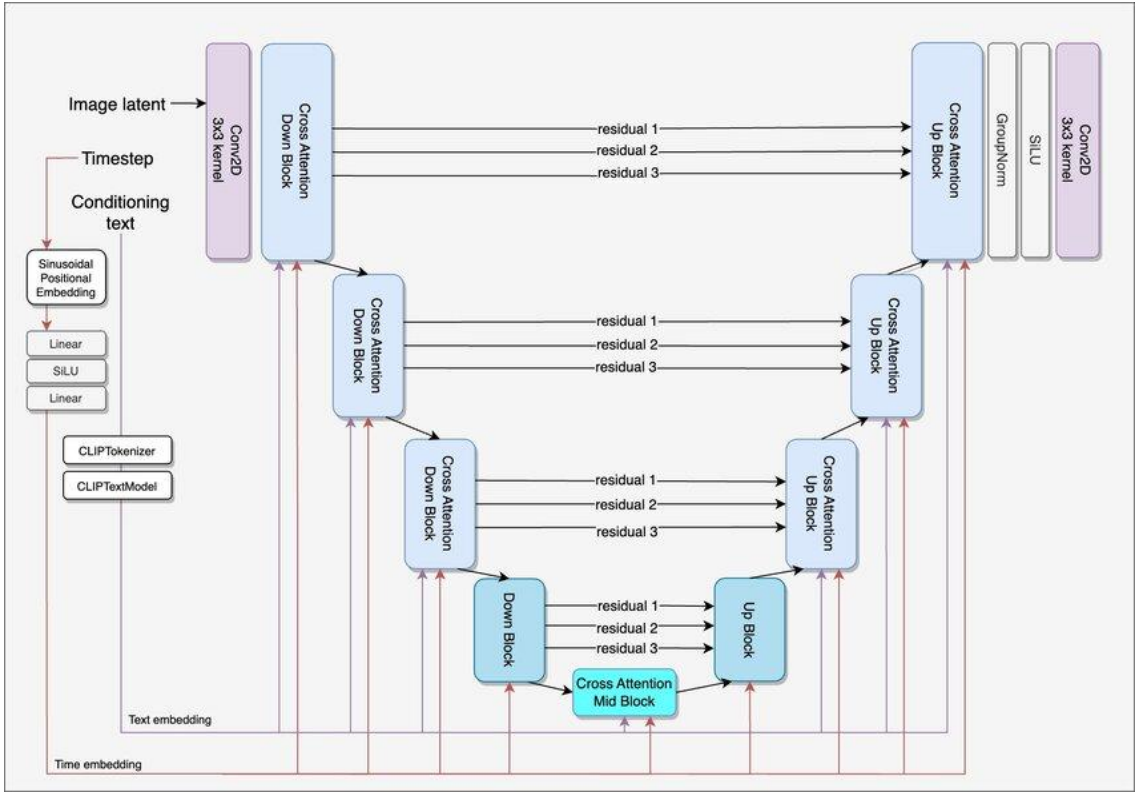
$\text{pour } t \text{ dans } T \dots 1, \quad x \leftarrow \text{UNet}(x, t)$

$\text{retourner } x$

Time conditional UNet

$$\mathrm{PE}_{(\mathrm{p},2i)} = \sin\left(\frac{p}{10000^{2i/d}}\right)$$

$$\mathrm{PE}_{(\mathrm{p},2i+1)} = \cos\left(\frac{p}{10000^{2i/d}}\right)$$



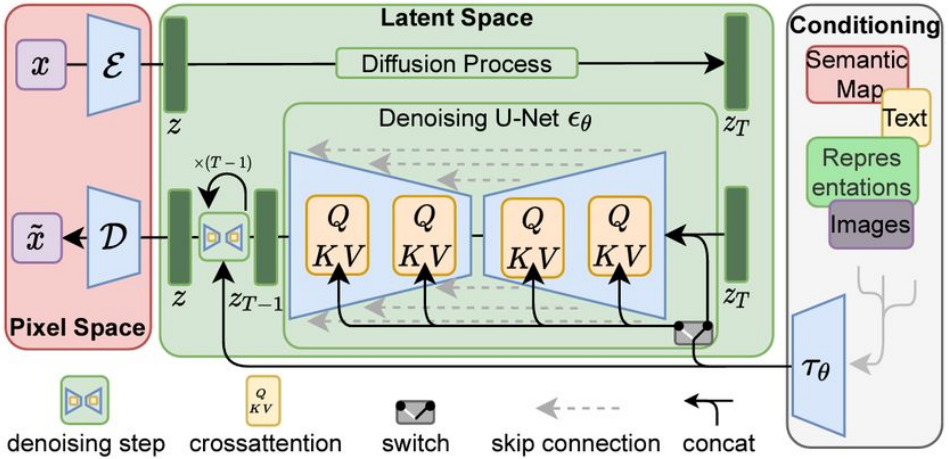
Modèle de Diffusion Latent (LDM)

Motivations

- Diffusion dans un espace latent et non dans celui des pixels → dimension réduite
- Chaque point = version condensée de l'information visuelle
- **Impact :** Réduction de la complexité de la tâche à apprendre

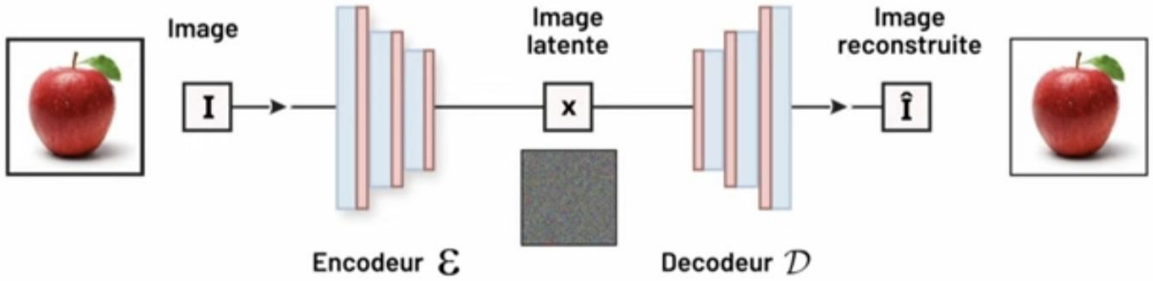
Architecture

- un Auto Encodeur Variationnel (VAE) (en entrée et sortie (Décodeur))
- un U-Net (au milieu)



High-Resolution Image Synthesis with Latent Diffusion Models, https://arxiv.org/pdf/2112.10752

Auto Encodeur Variationnel (VAE) et Décodeur



$$x \qquad z = f_\phi(x) \qquad \mathcal{L}_{AE} = \| x - \hat{x} \|^2$$

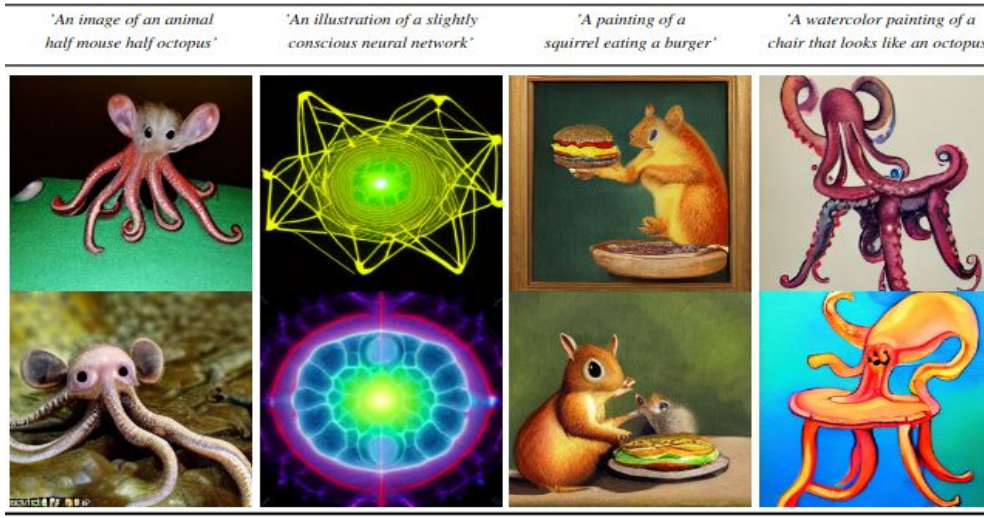
$$q_\phi(z \mid x) \sim \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x))$$

U-Net

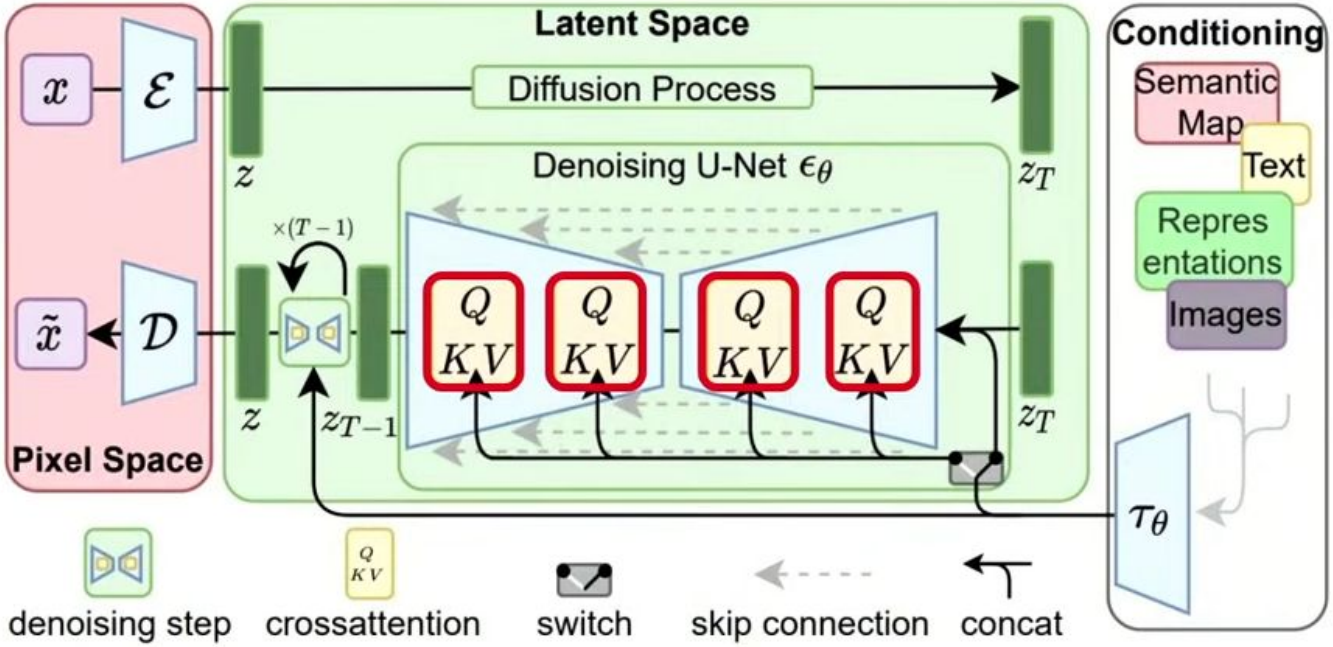
- **Entrées**
  - Image latente : tableau de dimensions (channels, largeur, hauteur)
  - Vecteur de bruit : représente le niveau de bruit (t)
  - Conditions : séquence d'embeddings supplémentaires (texte, style, etc.)
- **Processus**
  - Estimer la moyenne et la variance de la distribution Normale
  - Réduction progressive du bruit dans l’image latente
- **Sortie**
  - Image latente débruitée, décodée par le VAE en image finale

Conditionnement

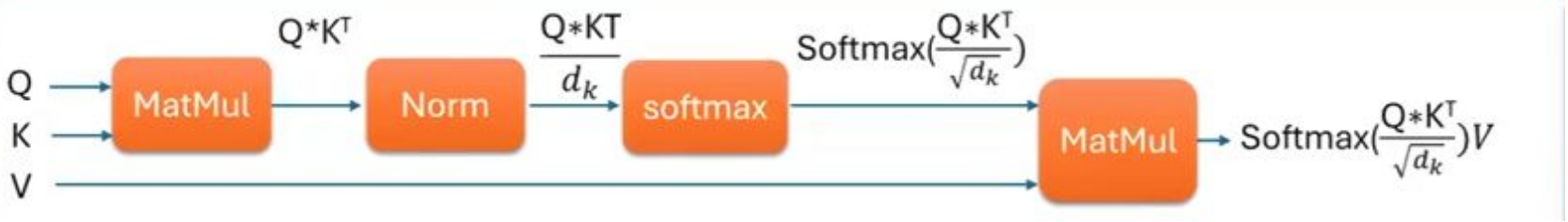
Introduction au Conditionnement



Principe du Conditionnement



Mécanisme d'Attention Croisée



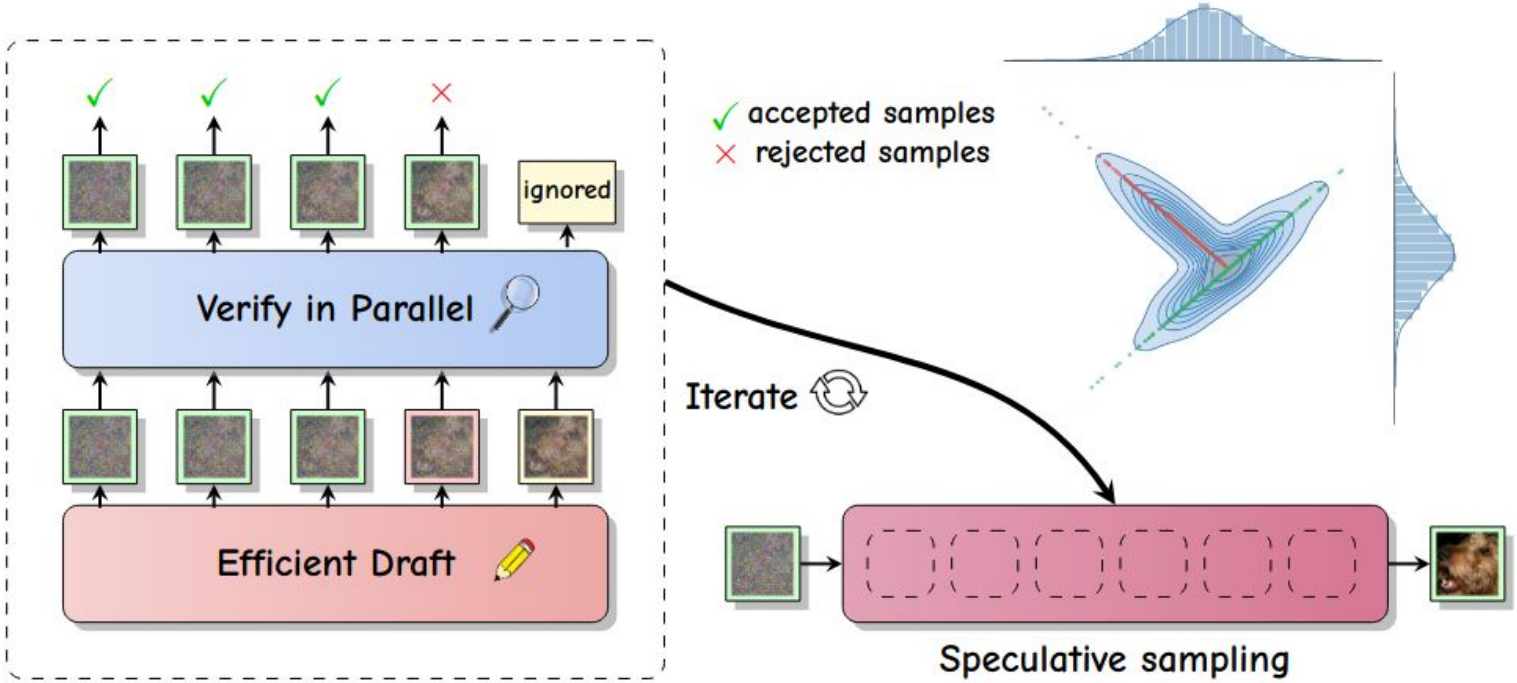
Optimisation de l’inférence

Différentes techniques existantes

- **Distillation**:
  - Entraînement d'un deuxième modèle simplifié qui approxime l'original
  - Perte de performance et entraînement supplémentaire
- **Amélioration des schémas d'échantillonnage**:
  - Réduire le nombre d'étapes nécessaires pour générer
  - Augmentation de la complexité computationnelle à chaque étape
- **Simulation parallèle**:
  - Diviser le processus de diffusion en sous-tâches exécutées en parallèles
  - Nécessite de meilleures ressources matérielles

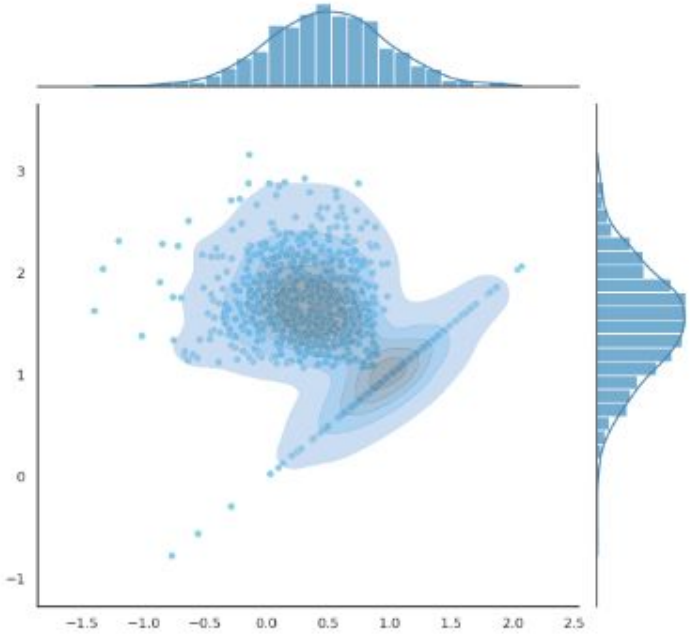
- **Échantillonnage spéculatif**: Technique d'optimisation issue des LLM, portée aux modèles de diffusion

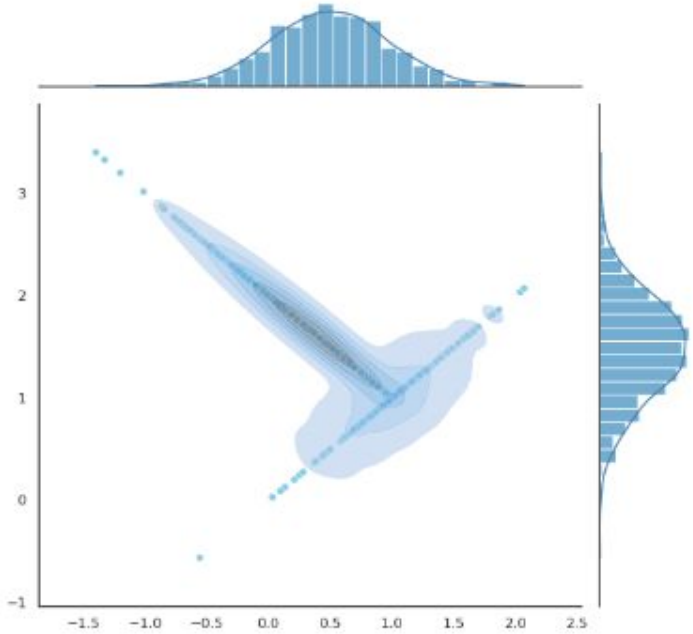
Échantillonnage spéculatif



Source: Accelerated Diffusion Models via Speculative Sampling, De Bartoli, Galashov, Gretton, Doucet (2025)

Couplage maximale





Source: Accelerated Diffusion Models via Speculative Sampling, De Bartoli, Galashov, Gretton, Doucet (2025)

Résultats

Dataset: CIFAR10 ((3, 32, 32) x 60 000)

Configuration	**Draft** (100 steps)		**Target** (100 steps)		**Target** (30 steps)		**Speculative**		
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	NFE ↓
$\varepsilon = 0.25, \tau = 2.0$	81.58	7.60	**2.45**	10.31	7.68	11.32	**2.34**	10.32	**35.40**

- **FID:** Similarité entre images réelles et générées
- **IS:** Qualité de l'image
- **NFE:** Nombre d'évaluation

Spéculatif: 35 appels → 2.34 FID
Classique: 100 appels → 2.45 FID
Soit **-65%** d'appels du modèle cible

L'échantillonnage spéculatif réduit généralement le nombre d'appel du modèle cible tout en préservant les performances

Source: Accelerated Diffusion Models via Speculative Sampling, De Bartoli, Galashov, Gretton, Doucet (2025)