

# Fusion Multimodale: Texte, Image et Audio

Master Data Science & Big Data / Module : Data Mining

Realisé Par:

**ABDELHAKIM ELGHAYOUBI  
OUSSAMA EL HAFIDI  
ILYASS BAZZI**

Encadré Par:

**Pr. NAWAL SAEL**





# Sommaire

**01**

Aperçu generale

**03**

Text to Audio

**02**

Text to Image



**04**

Demo & Conclusion



# 01

# Aperçu generale

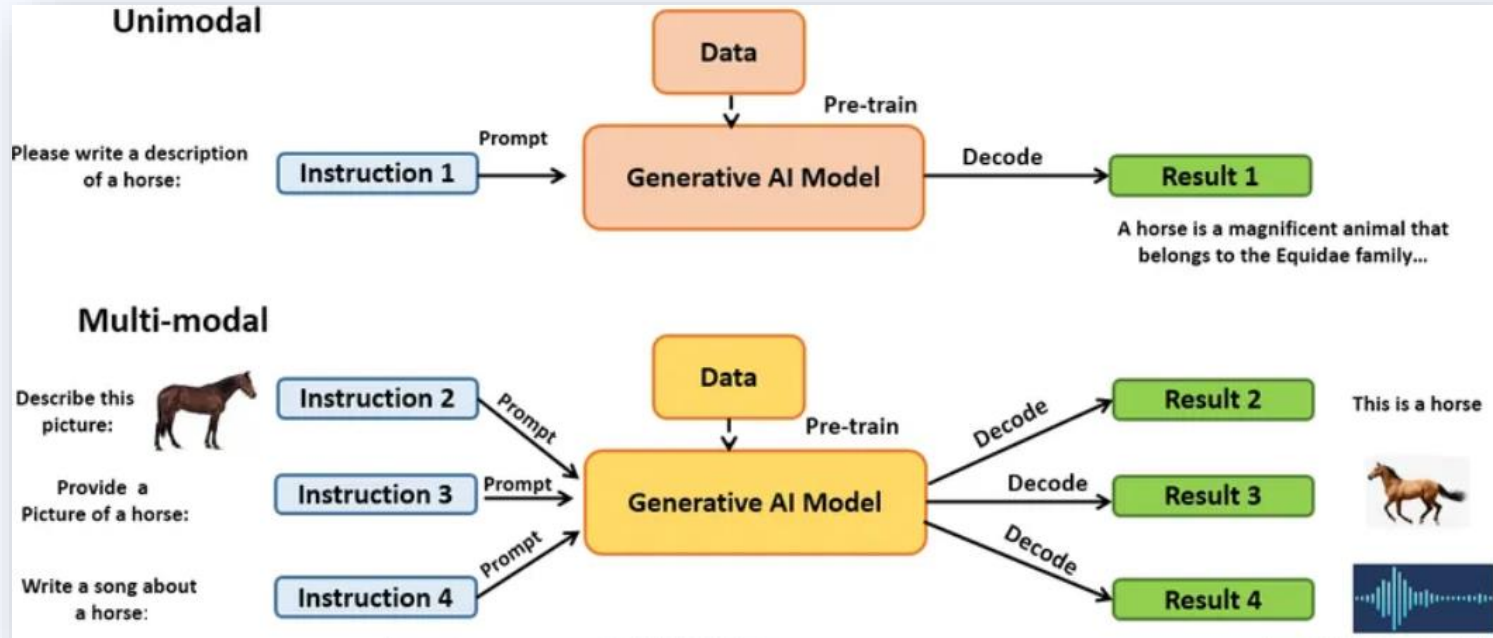
---

# Introduction à la Fusion Multimodale



- ❑ **Définition :** La fusion multimodale consiste à intégrer des données provenant de plusieurs modalités (texte, image, audio) dans un cadre unifié. Cela permet aux machines de comprendre et de relier des informations diversifiées, améliorant ainsi leur perception contextuelle.
- ❑ **Exemples :**
  - ❑ Générer des images ou de l'audio à partir de descriptions textuelles.
  - ❑ Créer des légendes textuelles pour des images.
- ❑ **Pourquoi est-ce important ?**
  - ❑ Cela imite la perception humaine qui combine la vision, le langage et les sons.
  - ❑ Utilisations dans les industries créatives, les outils d'accessibilité et les systèmes interactifs.

# Fonctionnement Technique



# Comment Fonctionnent les Modèles Multimodaux ?

- ❑ **Espaces d'Embedding** : Le texte, l'image et l'audio sont convertis en représentations numériques (embeddings). CLIP aligne ces embeddings pour des tâches croisées.
- ❑ **Pipelines de Génération** : Stable Diffusion et AudioLDM utilisent ces embeddings pour décoder des sorties (images/audio) à partir d'un prompt.

## Exemples de Workflow :

- ❑ **Texte → Image** :
  - ❑ Entrée : "Une cabane chaleureuse dans les bois sous la neige."
  - ❑ Processus : Génération d'embeddings → Décodage pour créer une image.
- ❑ **Texte → Audio** :
  - ❑ Entrée : "Un marché animé avec des conversations."
  - ❑ Processus : Génération d'embeddings → Synthèse audio des sons correspondants.

# Applications de l'IA Multimodale

## Domaines Créatifs :

- ❑ Génération artistique et musicale à partir de prompts textuels.
- ❑ Création d'histoires avec des éléments visuels et sonores synchronisés.

## Accessibilité :

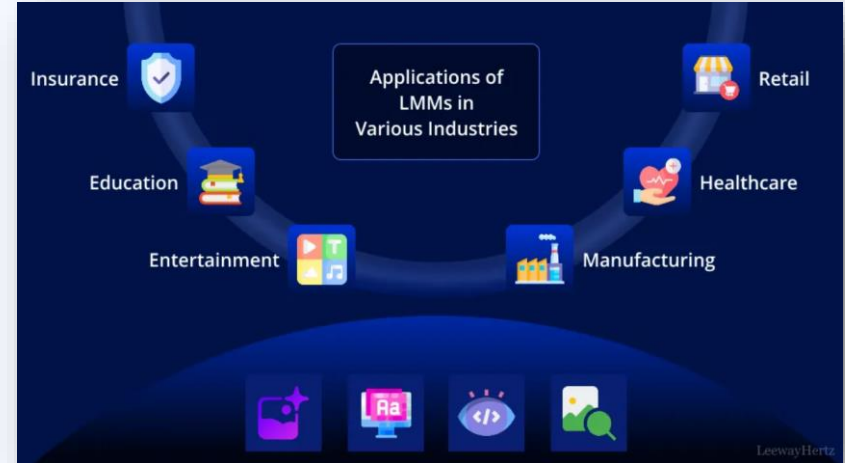
- ❑ Aides pour les personnes ayant un handicap visuel ou auditif, comme la génération de paysages sonores ou de descriptions visuelles.

## Éducation :

- ❑ Matériel d'apprentissage immersif combinant visuels, narration et effets sonores.

## Interaction Homme-Machine :

- ❑ Assistants virtuels enrichis capables de reconnaître des objets visuels et de répondre de manière contextuelle.





# Points forts et limites

## Points forts :

- ❑ **Compréhension enrichie** : Combine plusieurs modalités pour une analyse plus complète.
- ❑ **Interactions naturelles** : Approche plus intuitive, imitant la perception humaine.
- ❑ **Applications variées** : Utilisations dans la création, l'éducation et l'accessibilité.
- ❑ **Polyvalence** : Modèles adaptés à une large gamme de tâches sans apprentissage supplémentaire.
- ❑ **Personnalisation** : Contenus adaptés aux préférences et contextes des utilisateurs.

## limites :

- ❑ **Alignement des Données** : S'assurer de la cohérence entre les modalités (par exemple, une image générée correspond au contexte sonore).
- ❑ **Ressources** : Les modèles sont gourmands en calcul et en mémoire.
- ❑ **Subjectivité des Résultats** : Les modèles peuvent interpréter différemment des prompts similaires, par exemple, des termes comme "joyeux" ou "sombre".





# 02

# Text to Image

---

# Aperçu

Un modèle de génération de texte à image utilise l'apprentissage automatique pour créer des images à partir de descriptions textuelles. Il repose sur des réseaux de neurones profonds pour comprendre les liens entre texte et visuels, produisant des images alignées avec les descriptions.

## Applications

- ❑ **Industries Créatives** : Génération rapide d'idées visuelles (ex : designs de mode).
- ❑ **Marketing** : Création de visuels pour campagnes à partir de descriptions de produits.
- ❑ **E-commerce** : Amélioration des listes de produits avec des images générées.
- ❑ **Éducation** : Visualisation de concepts complexes pour faciliter l'apprentissage.

# Modèles populaires

## ❑ DALL-E (OpenAI)

- ❑ **2021** : Pionnier de la génération de texte à partir d'images avec une compréhension du texte basée sur CLIP.
- ❑ **DALL-E 2** : Amélioration de la résolution, du réalisme et de l'alignement du texte.
- ❑ **DALL-E 3** : Amélioration de la compréhension du contexte et de la fidélité des messages.

## ❑ Diffusion stable (Stability AI)

- ❑ **2022** : Modèle open-source basé sur la diffusion, connu pour ses résultats de haute qualité et son accessibilité.
- ❑ Excelle dans la génération d'images artistiques et photoréalistes.















## ❑ Imagen (Google)

- ❑ Combine de puissants encodeurs de texte avec un accent sur le photoréalisme et les détails fins.
- ❑ Remarquable pour sa grande fidélité et son alignement sur les invites.

## ❑ MidJourney

- ❑ Populaire pour la génération d'images stylisées et artistiques avec une gestion intuitive des invites.
- ❑ Largement utilisé dans les industries créatives et les médias sociaux.

# Benchmark

Leaderboard Image Arena Personal Arena Leaderboard				
TEXT TO IMAGE MODEL LEADERBOARD				
CREATOR	NAME	ARENA ELO	ARENA WIN RATE	# SELECTIONS
 Midjourney	Midjourney v6	1171	71%	13129
 Stability.ai	Stable Diffusion 3	1154	68%	10153
 Playground AI	Playground v2.5	1099	61%	11475
 OpenAI	DALLE 3 HD	1094	61%	11349
 OpenAI	DALLE 3	1073	58%	10771
 Stability.ai	Stable Diffusion 3 Turbo	1038	52%	7621
 Stability.ai	Stable Diffusion 3 Medium	1033	53%	2997
 Stability.ai	Stable Diffusion 1.6	1028	51%	7475
 Stability.ai	SDXL Lightning	1009	49%	9033
 Amazon	Amazon Titan G1	1008	49%	4244
 Stability.ai	Stable Diffusion XL 1.0	980	45%	8357
 OpenAI	DALLE 2	819	24%	4440
 Stability.ai	Stable Diffusion 2.1	819	24%	4395
 Stability.ai	Stable Diffusion 1.5	674	16%	829

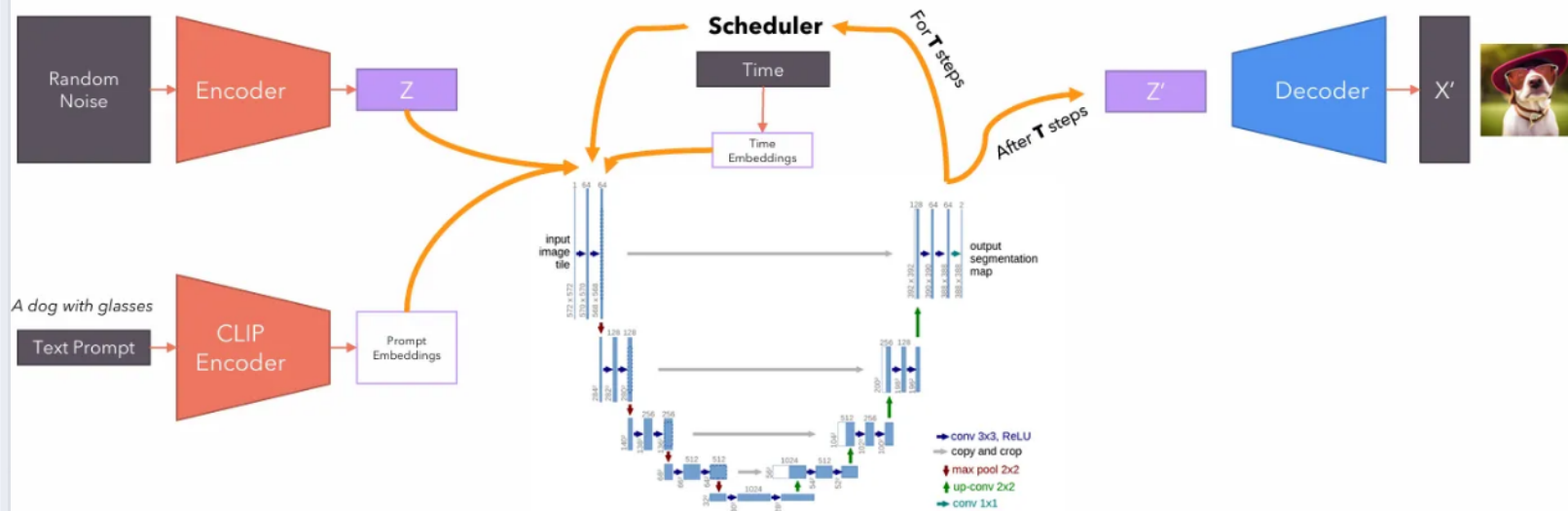
# Grandes questions à reponde

- ❑ Méthode d'apprentissage permettant de générer de nouveaux éléments à partir de nombreux exemples → **Forward/Reverse Diffusion**
- ❑ Méthode de compression des images (pour accélérer la formation et la génération) → **Latent Space/VAE**
- ❑ Moyen de lier le texte et les images → **CLIP**
- ❑ Moyen a guide la generation d'image→ **U-Net & Cross attention**

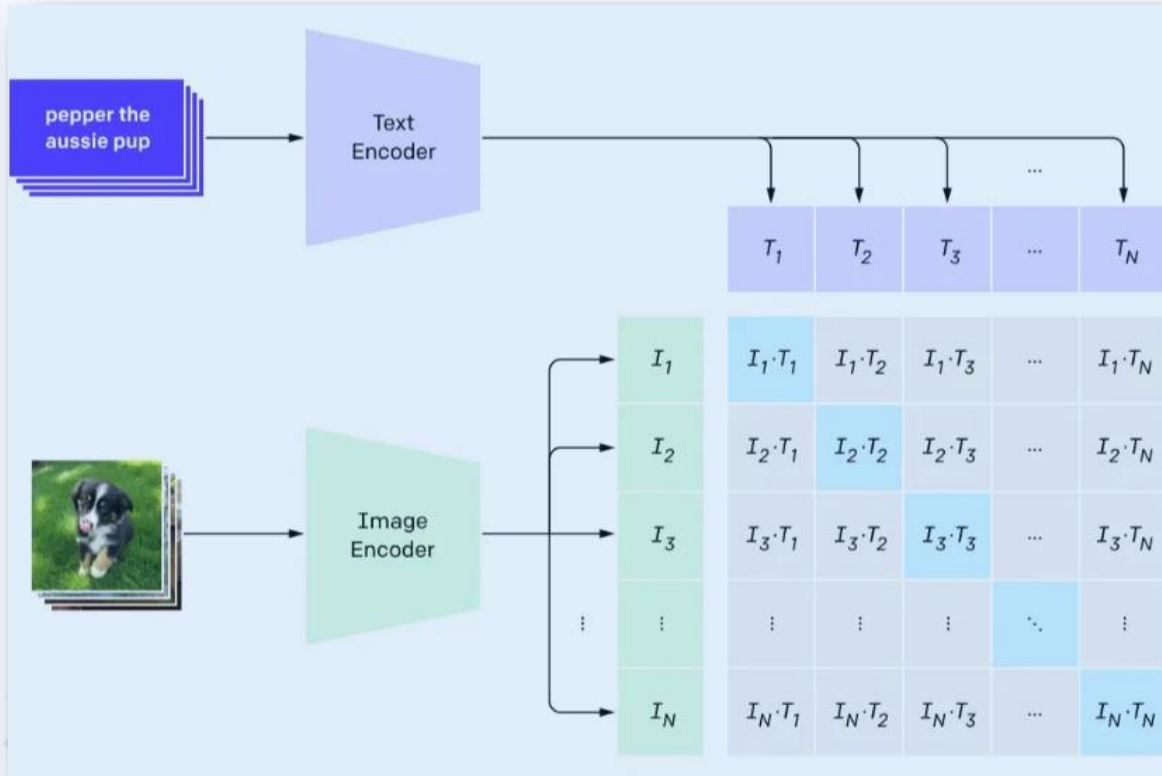


# Architecture Générale

## Architecture (Text-To-Image)

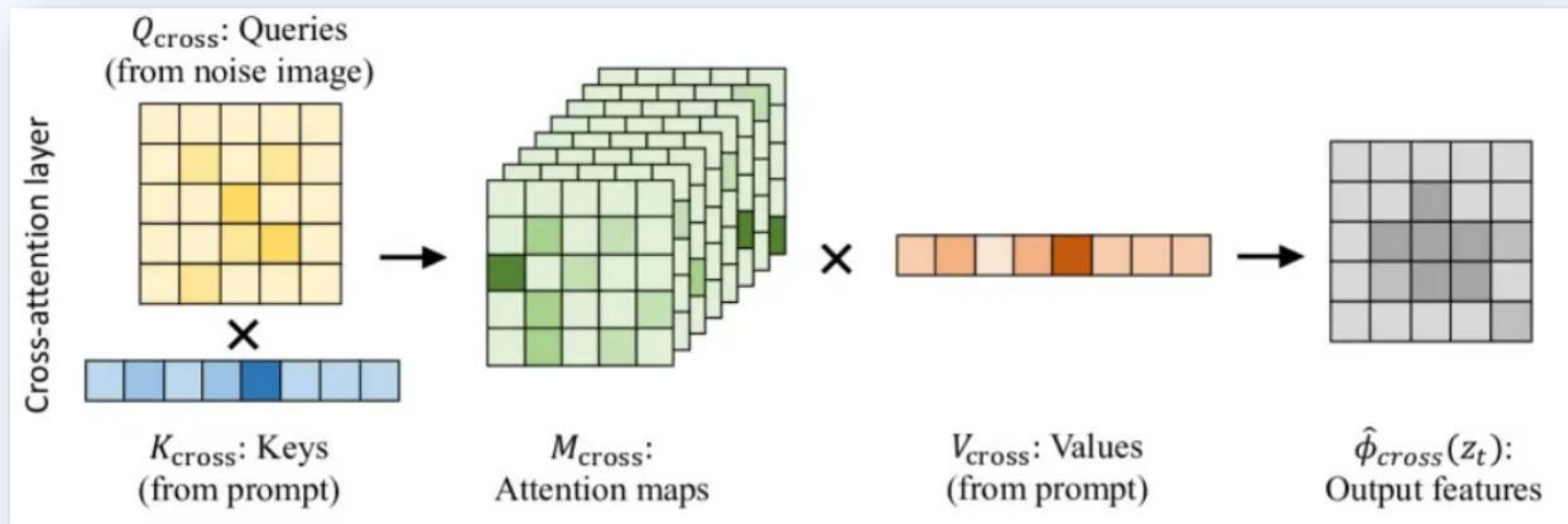


# Étape 1 : Encodage du Texte avec CLIP



# Étape 1 : Encodage du Texte avec CLIP

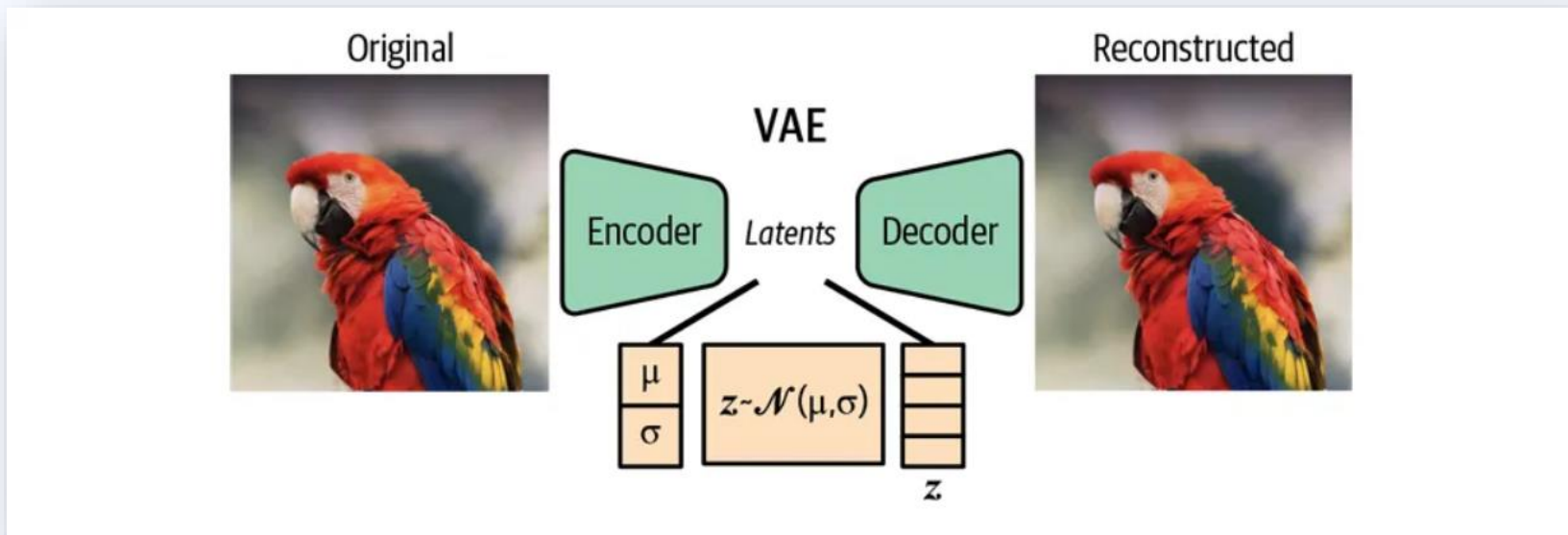
## Mécanisme de Cross-Attention





## Étape 2 : Initialisation de l'Espace Latent (Bruit Aléatoire)

Espace Latent



# Étape 2 : Initialisation de l'Espace Latent (Bruit Aléatoire)

## Initialisation par Bruit Aléatoire

$$g(x, y) = f(x, y) + n(x, y)$$

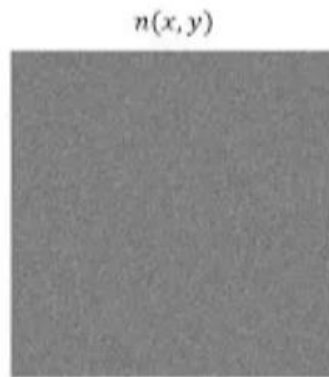


=



Pixel value: [0~1] or [0~255]

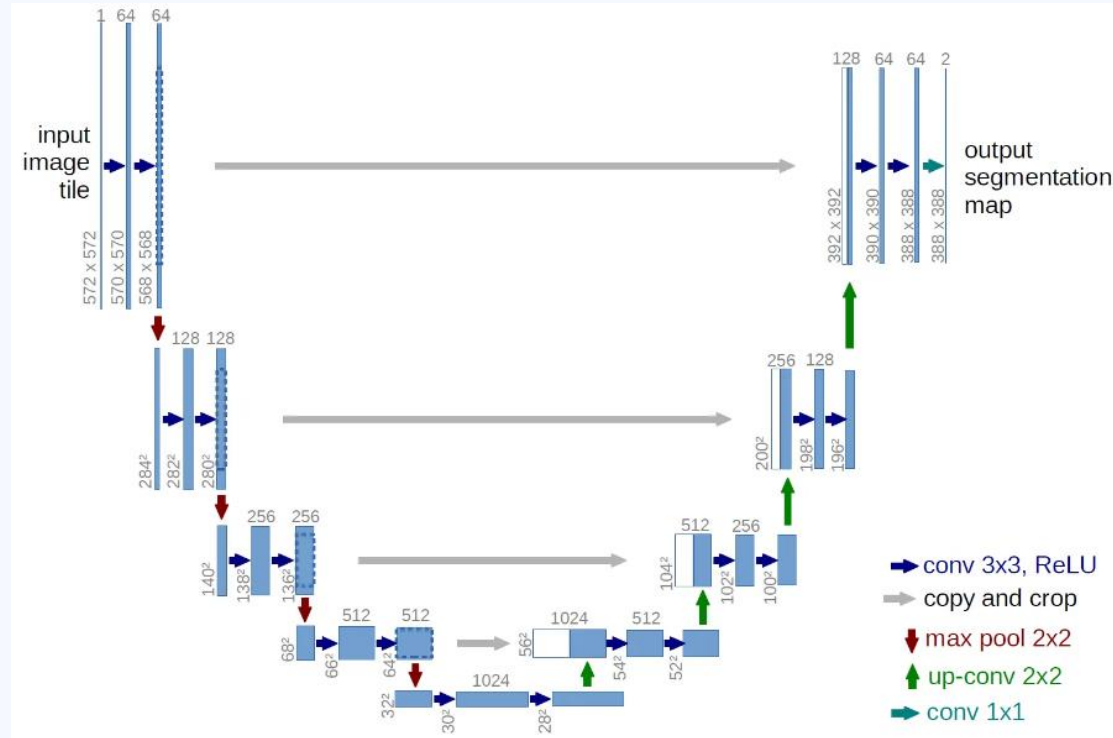
+



Pixel value: random number

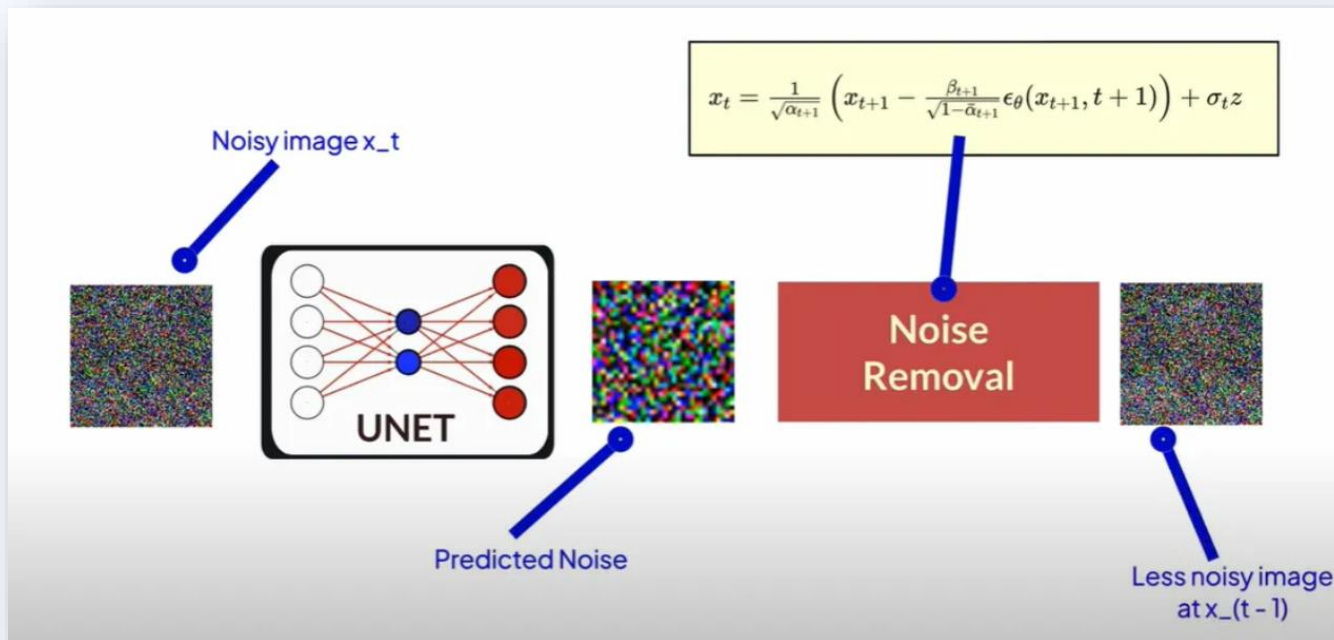
# Étape 3 : Dé-bruitage avec U-Net et Cross-Attention (Generation d'image)

## U-Net



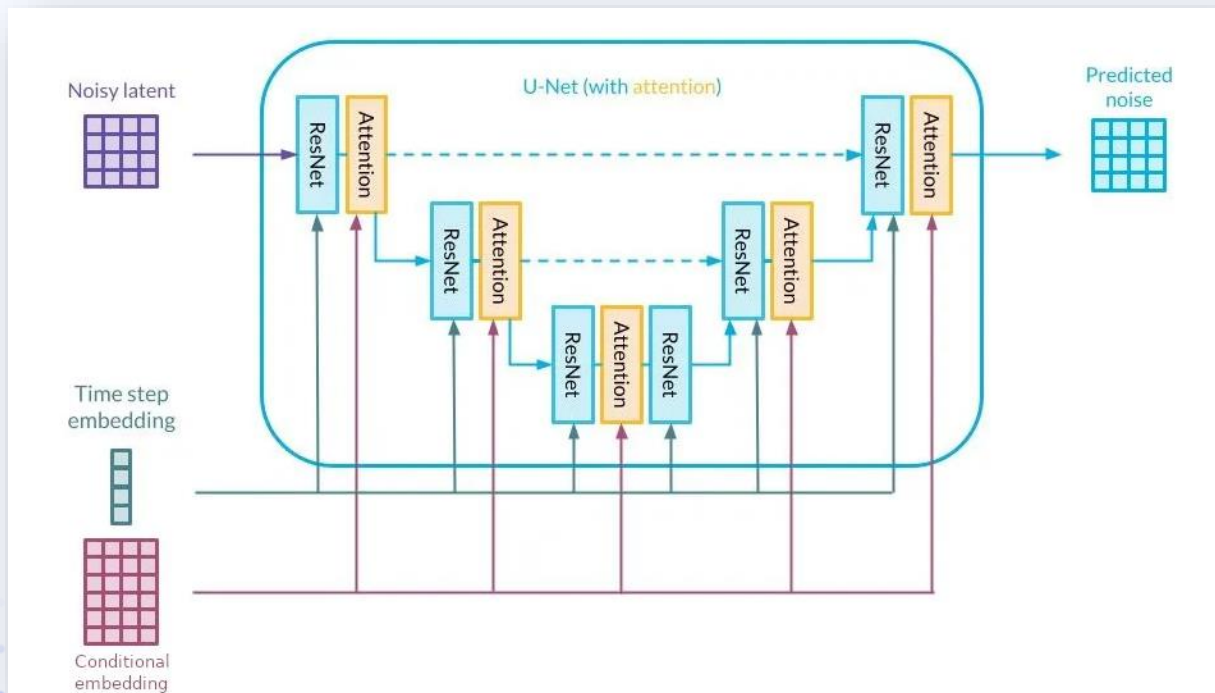
# Étape 3 : Dé-bruitage avec U-Net et Cross-Attention (Generation d'image)

## Processus de Diffusion Inverse (Débruitage)

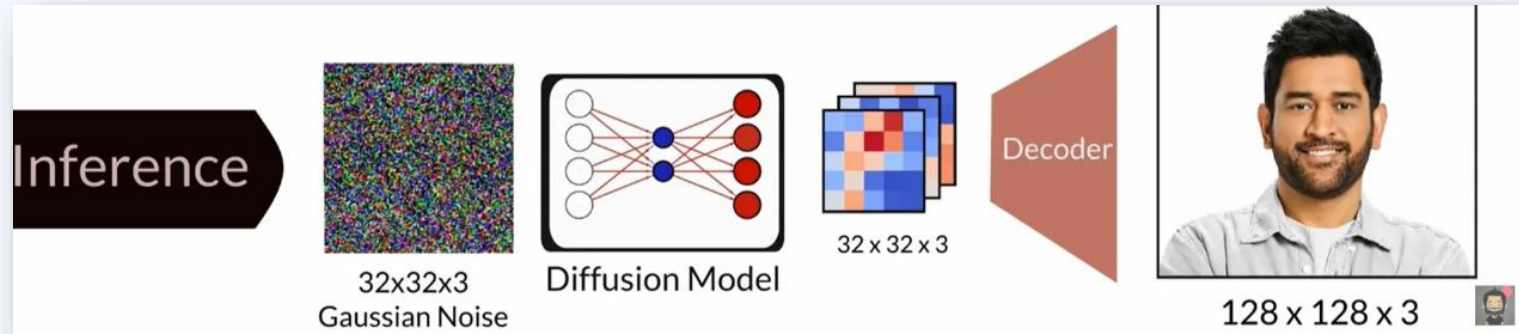


# Étape 3 : Dé-bruitage avec U-Net et Cross-Attention (Generation d'image)

## Processus Itératif de Diffusion Inverse



## Étape 4 : Décodage de l'Image avec le Décodeur VAE (Étape Finale)



- ❑ Le **décodeur VAE** est responsable du **décodage de l'espace latent** en une image dans l'**espace pixel**.
- ❑ Il prend la représentation latente débruitée et la reconstruit en une image haute résolution, transformant les informations abstraites et compressées en une **image visuellement significative**.

# Résultat Final

- ❑ Après l'étape finale de débruitage et de décodage, le résultat est une **image haute résolution** qui correspond étroitement au prompt texte.
- ❑ Par exemple, le prompt "***A serene lake surrounded by snow-capped mountains at sunrise***" générerait une **image réaliste** d'un lac entouré de montagnes, illuminé par un lever de soleil.





# 03

# Text to Audio

---



# Aperçu

## ❑ AudioLDM :

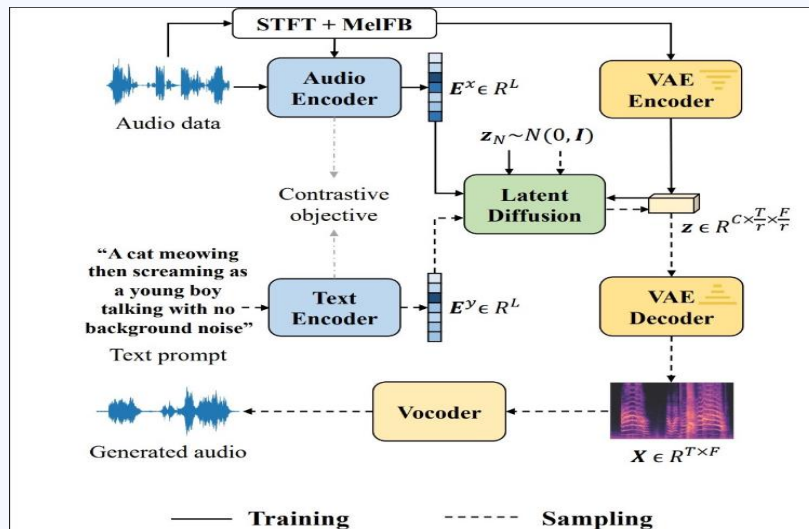
- Proposé dans l'article "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models".
- C'est un modèle de diffusion latent : technique utilisant un processus progressif de bruitage et de débruitage pour apprendre une représentation compacte des données.
- Il permet de générer des audios à partir de descriptions textuelles.
- Applications : Production d'effets sonores, voix humaines ou musique à partir de prompts textuels.





# Architecture AudioLDM

L'équipe de développement a exploré diverses combinaisons et a déterminé que l'architecture ci-dessous était la plus performante. Les critères d'évaluation étaient guidés par une philosophie de conception visant non seulement à améliorer les performances du modèle, mais également à garantir que sa conception reste évolutive, efficace en termes de coûts et optimisée pour la mémoire.



# Architecture AudioLDM

## 1. Prétraitement des données audio

- **STFT (Short-Time Fourier Transform) :**

- ✓ La transformation STFT décompose le signal audio brut en un spectrogramme qui capture les informations de fréquence et de temps.
- ✓ Elle permet de convertir un signal audio temporel en une représentation dans le domaine fréquentiel, plus adaptée pour l'analyse et la génération.

- **MelFB (Mél-filterbank) :**

- ✓ Applique une banque de filtres pour générer un spectrogramme Mel, qui est une représentation compacte et perceptuellement pertinente (inspirée de la perception humaine des sons).

## 2. Audio Encode

L'encodeur audio transforme les spectrogrammes Mel en une représentation compacte dans un espace latent.

- **Entrée : Spectrogramme Mel**

Obtenu à partir du signal audio brut via la STFT et une banque de filtres Mel, le spectrogramme contient des informations temporelles et fréquentielles pertinentes.

- **Modèle d'encodage :**

Utilise des architectures comme :

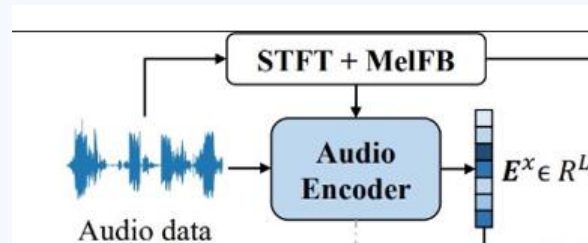
**CNNs** : Pour capturer les motifs locaux du spectrogramme.

**Transformers avec attention** : Pour analyser les relations à long terme entre les fréquences et les temps.

**RNNs** : Pour modéliser les dépendances séquentielles.

- **Sortie : Vecteur latent**

Ce vecteur encode les caractéristiques essentielles (tonalité, timbre, variations temporelles) dans un format compact aligné avec les représentations textuelles pour un objectif contrastif.

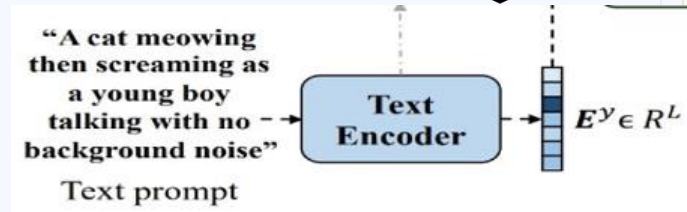


# Architecture AudioLDM

## 3. Text Encoder

Le rôle de l'encodeur de texte est de convertir une description textuelle en un vecteur latent  $E^y \in R^L$

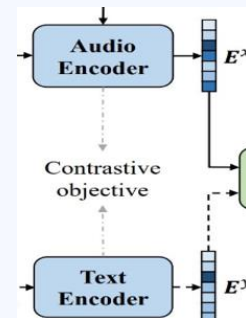
Pour ce faire, l'encodeur utilise un modèle transformer qui repose sur le mécanisme d'attention.



**Résultat :** Une sortie sous forme de vecteur latent  $E^y$  alignable avec les représentations audios via un objectif contrastif.

## 4. Contrastive Objective

- Cette étape vise à aligner les représentations des données audio  $E^x$  et texte  $E^y$  dans le même espace latent.
- En alignant ces deux espaces, le modèle apprend à relier les descriptions textuelles aux caractéristiques des signaux audio correspondants.



# Architecture AudioLDM



## 5. Latent Diffusion Model (LDM)

Le **LDM** repose sur un processus de diffusion progressif qui consiste en deux étapes clés : **le bruitage** et le **débruitage**.

- **Bruitage (Diffusion forward) :**

Le processus de diffusion commence par l'ajout progressif de bruit à la représentation latente initiale  $z_0$  au fil des étapes  $t$ . À chaque étape  $t$ , un bruit gaussien  $\epsilon$  est ajouté à la représentation latente, perturbant ainsi les données de manière contrôlée.

**Formule** mathématique de la diffusion (**bruitage**) :

$$z_t = \sqrt{\alpha t} z_0 + \sqrt{1 - \alpha t} \epsilon, \epsilon \sim N(0, I)$$

$z_t$  est la représentation bruitée à l'étape  $t$ ,

$\alpha t$  est un hyperparamètre qui contrôle l'ampleur du bruit ajouté à chaque étape,

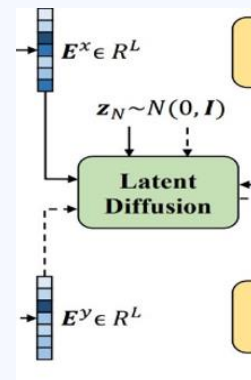
$z_0$  est l'information initiale ou la représentation latente sans bruit,

- **Débruitage (Diffusion backward) :**

Le modèle apprend à inverser ce processus de diffusion, c'est-à-dire à **débruiter progressivement  $z_t$**  pour récupérer **la représentation latente originale  $z_0$** . Le processus inverse (ou débruitage) est optimisé de manière à permettre au modèle de reconstruire  $z_0$  à partir de  $z_t$  même lorsqu'il est fortement bruité.

- L'objectif ici est de restaurer les informations de manière précise, tout en apprenant à supprimer le bruit de façon contrôlée, étape par étape, pour obtenir une représentation latente propre et significative.

**Sortie du modèle:** Le modèle, après débruitage, **génère un vecteur latent** conservant les caractéristiques essentielles de l'audio d'origine.

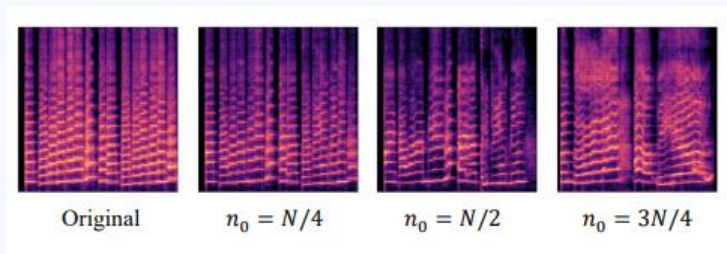
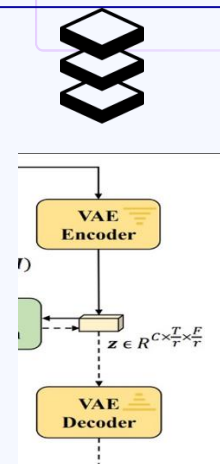


# Architecture AudioLDM

## 6. VAE (Variational Autoencoder)

Le VAE (Autoencodeur Variationnel) est un modèle génératif probabiliste utilisé ici pour encoder et décoder les spectrogrammes audios.

- **Encodeur VAE** : Comprime les données en un espace latent en apprenant une distribution gaussienne  $z \sim N(\mu, \sigma)$  plutôt qu'une représentation déterministe.
- **Décodeur VAE** : Reconstitue les données originales (spectrogramme).



## 7. Vocoder

- Transforme le spectrogramme généré par le décodeur VAE en signal audio brut (onde sonore).

# Architecture AudioLDM



## Flux global du processus

- **Entraînement:**
  - Le modèle est entraîné à relier les représentations latentes audio et texte, tout en apprenant à reconstruire les données audio à partir du bruit ajouté dans l'espace latent.
- **Génération/Sampling :**
  - Donne un texte en entrée au Text Encoder, puis utilise le LDM et le VAE pour produire un spectrogramme, qui est finalement converti en audio avec le vocoder.

## Fonction de perte :

- Contraste pour aligner audio et texte :

$$L_{contrastive} = -\log\left(\frac{\exp(s(a, t))}{\sum t' \exp(s(a, t'))}\right)$$

où  $s(a, t)$  est la similarité entre audio et texte.

- Diffusion : Maximiser la vraisemblance des données.

# Tableau des Configurations des Checkpoints



Checkpoint	Étapes d'entraînement	Conditionnement audio	Dimension audio CLAP	Dimension UNet	Paramètres
<a href="#">audioldm-s-full</a>	1,5M	Non	768	128	421M
<a href="#">audioldm-s-full-v2</a>	> 1,5M	Non	768	128	421M
<a href="#">audioldm-m-full</a>	1,5M	Oui	1024	192	652M
<a href="#">audioldm-l-full</a>	1,5M	Non	768	256	975M

- **Conditionnement audio** : Indique si le modèle prend en compte des données audio comme condition d'entrée.
- **Dimension audio CLAP** : Taille de la représentation latente audio/textuelle utilisée par le modèle.
- **Dimension UNet** : Taille de l'architecture UNet utilisée pour la diffusion latente.



# 04

## Demo & Conclusion

---

# Conclusion

La fusion multimodale, combinant texte, image et audio, transforme les interactions avec l'IA en les rendant plus riches et immersives. Des modèles comme **CLIP**, **Stable Diffusion** et **AudioLDM** repoussent les frontières de l'intelligence artificielle en permettant la génération et l'analyse de contenus variés. Les applications couvrent des domaines comme les industries créatives, l'éducation et l'accessibilité, tout en ouvrant des perspectives pour des systèmes plus synchronisés et personnalisés. Bien que des défis technologiques subsistent, cette approche promet un avenir où l'IA interagit de manière intuitive et proche des capacités humaines.



**Thanks for your  
attention 😊!**