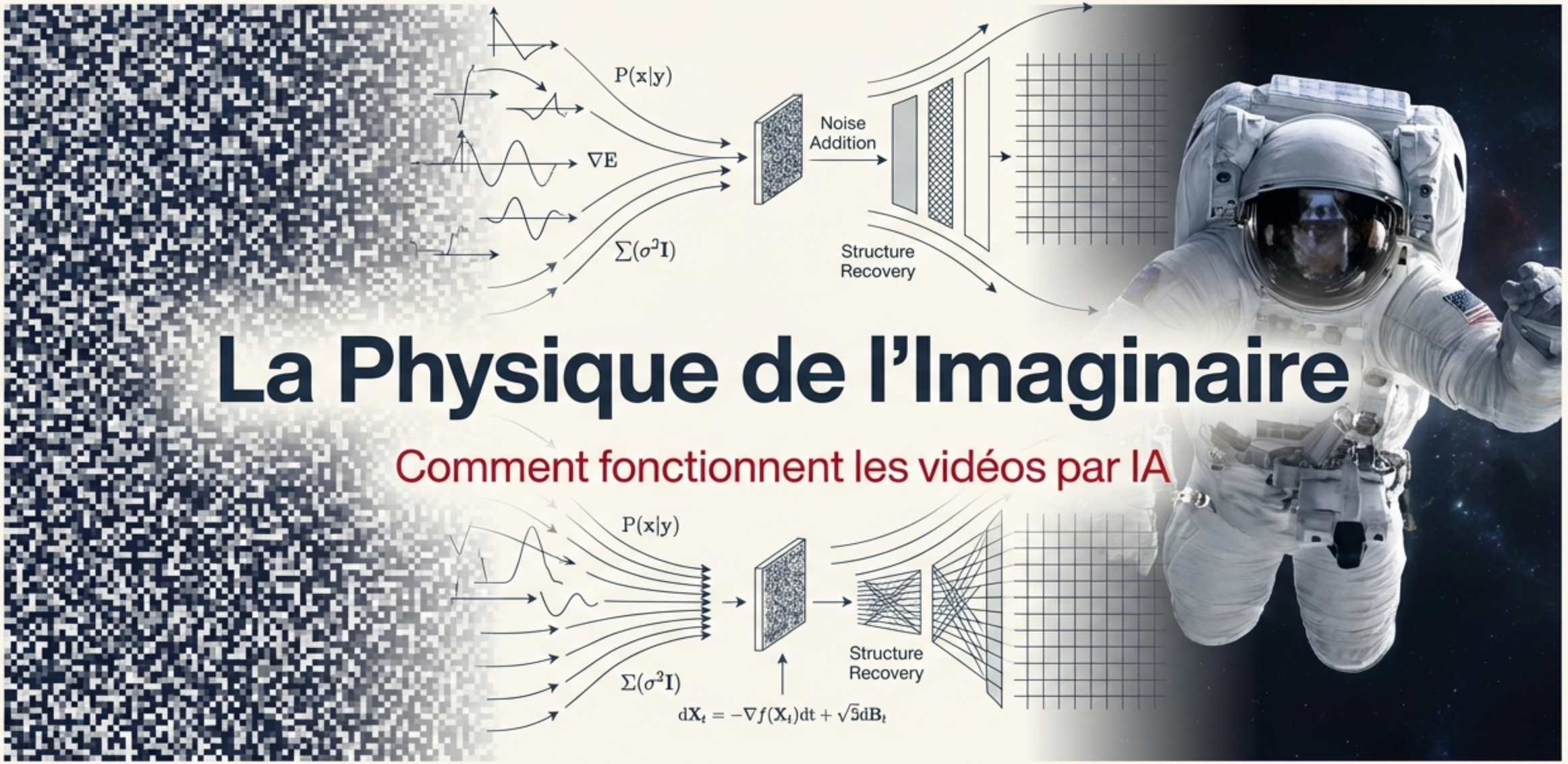


La Physique de l'Imaginaire

Comment fonctionnent les vidéos par IA

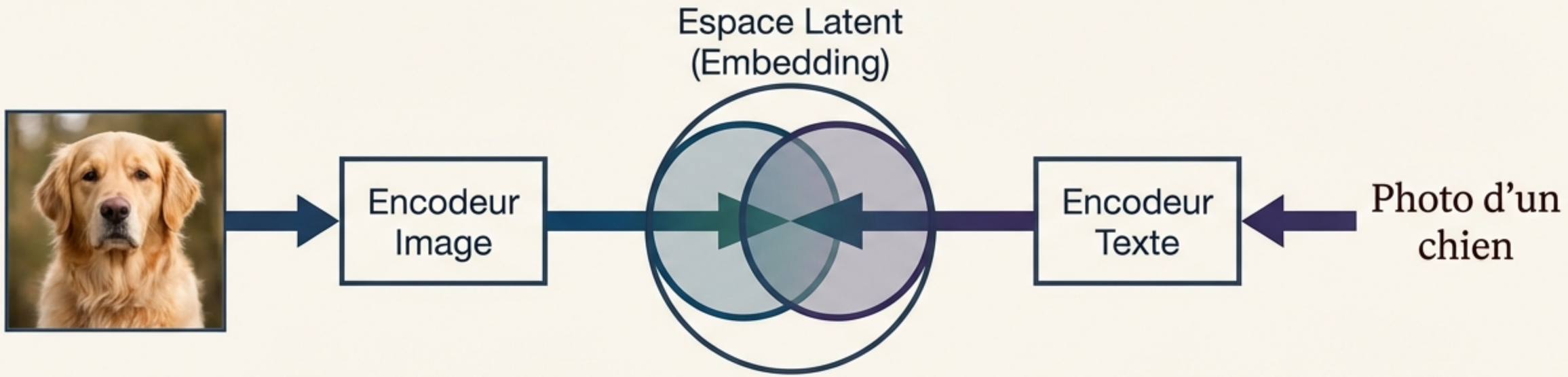
Une exploration visuelle des modèles de diffusion

Sculpter le bruit

Bruit Pur (t=T) t=40 t=25 t=10 Vidéo Finale (t=0)

Le modèle ne crée pas l'image d'un seul coup. Il prédit et retire le bruit, itération après itération, pour révéler une structure cachée.

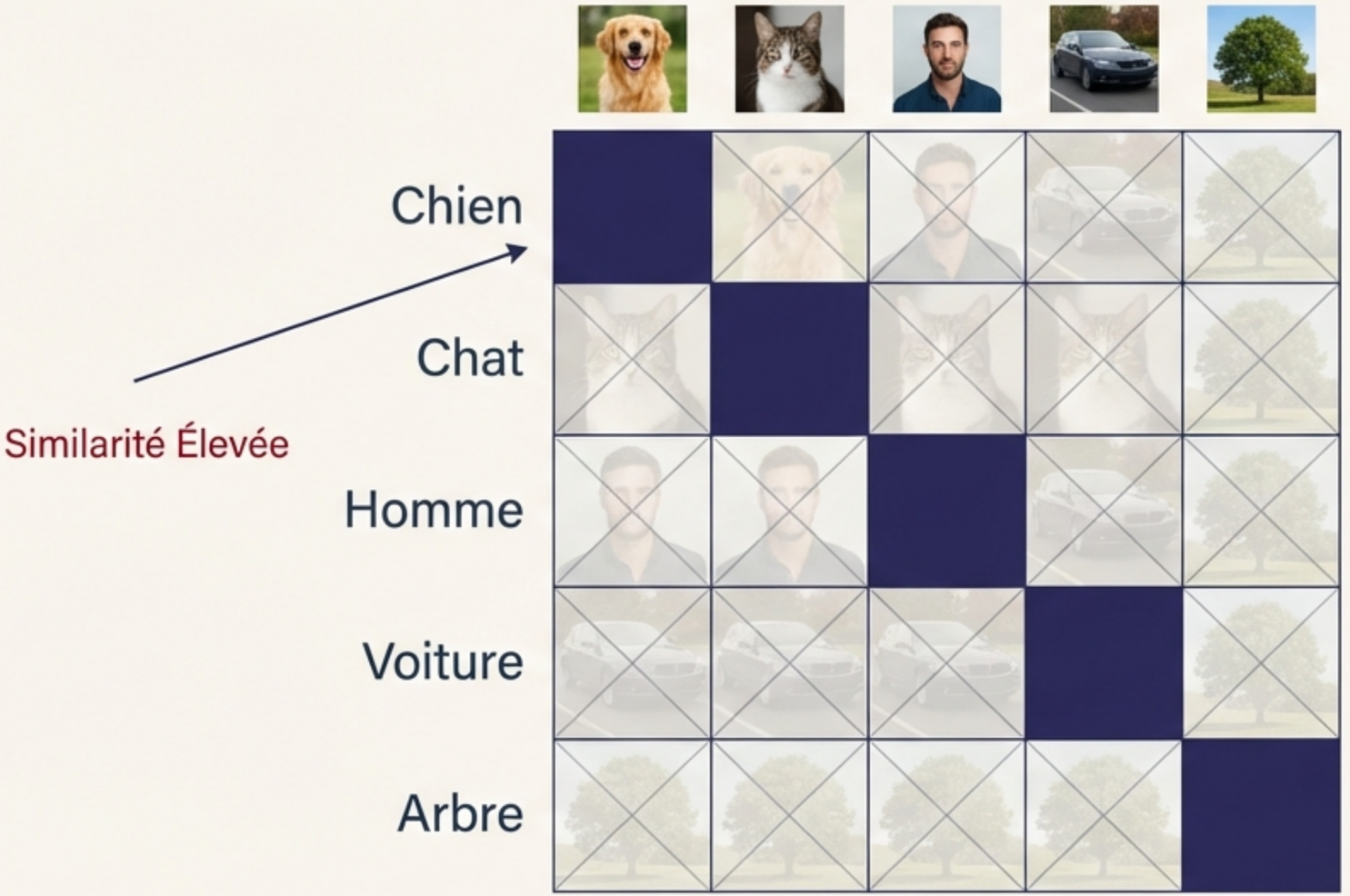
Le langage des images : CLIP



Pour générer, il faut d’abord comprendre. CLIP projette les images et les textes dans un même espace mathématique, créant un pont entre le visuel et le sémantique.

L’apprentissage contrastif

L'objectif : maximiser la similarité sur la diagonale (paires correctes) et la minimiser partout ailleurs.



L’algèbre des concepts

Vecteur
(Homme + Chapeau)

—

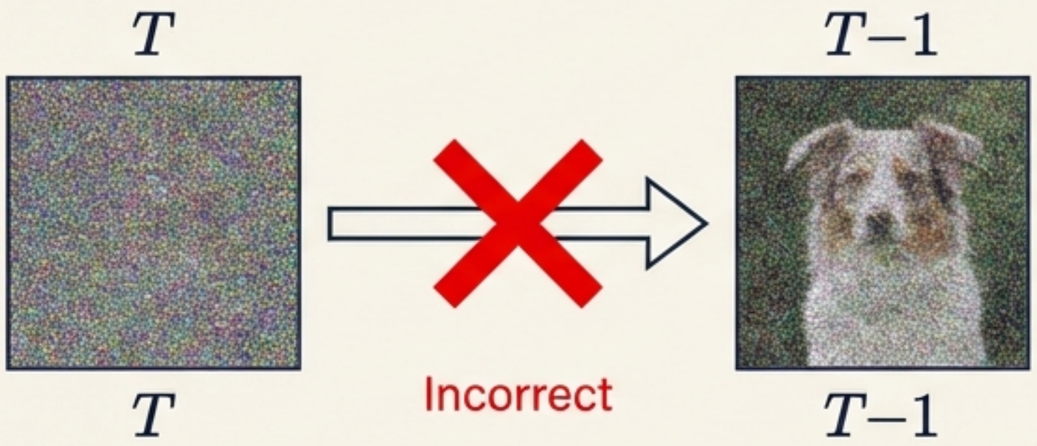
Vecteur
(Homme)

≈ ⟶ CHAPEAU

L'espace latent capture le sens. En soustrayant le vecteur 'Moi' du vecteur 'Moi avec chapeau', on isole le concept mathématique pur du chapeau.

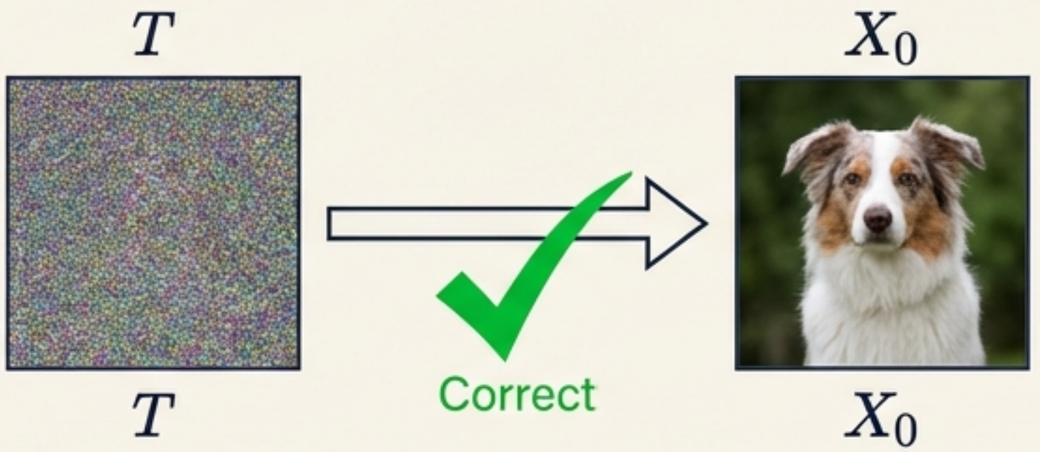
L’intuition vs La Réalité

Approche Naïve



Nettoyage étape par étape

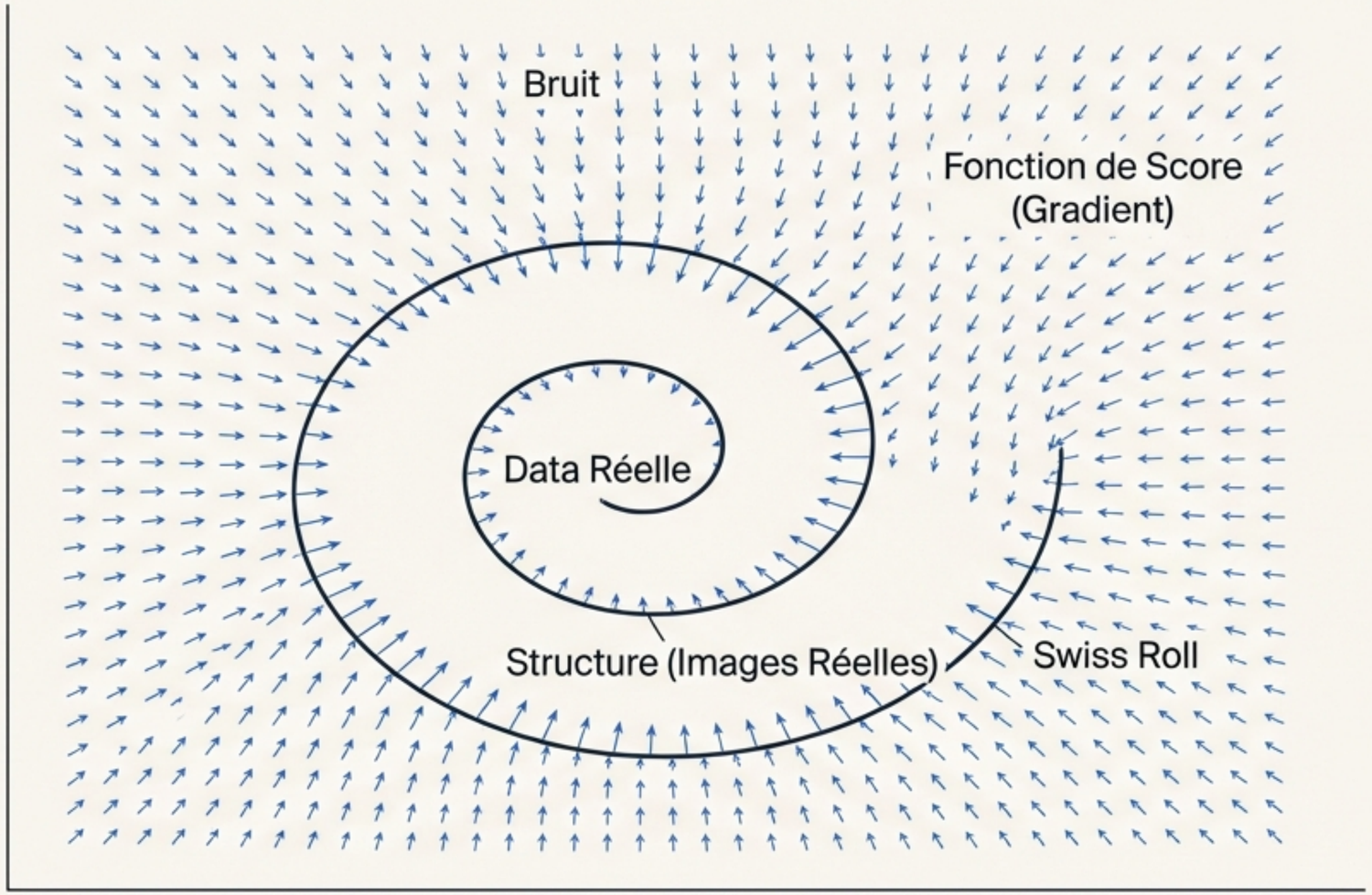
Réalité Mathématique



Prédire le bruit total (X0)

Les modèles modernes n’apprennent pas à nettoyer l’image petit à petit. Ils apprennent à prédire l’image finale ($X0$) à partir du bruit actuel.

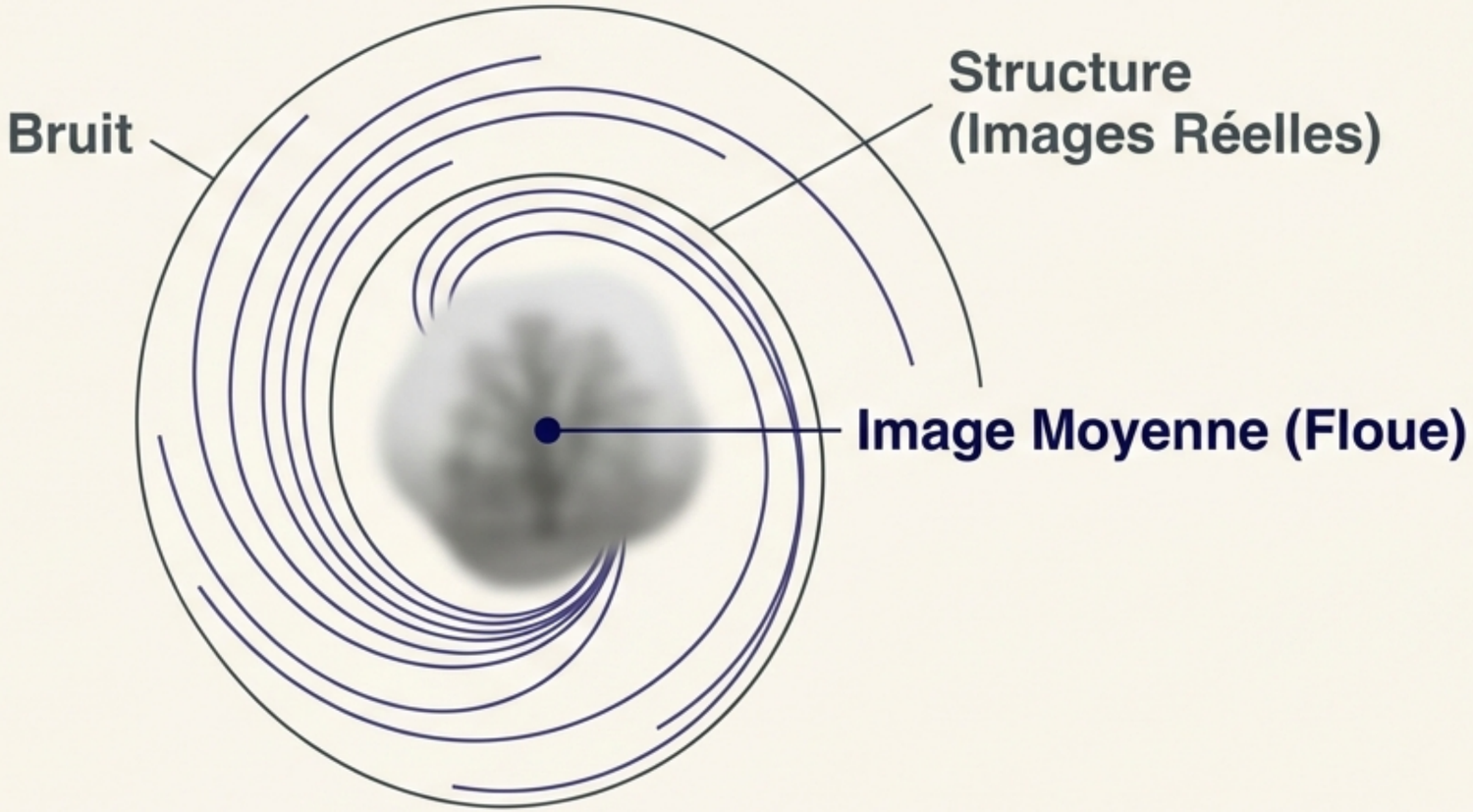
L’analogie de la spirale



Imaginons toutes les images possibles sur une spirale 2D.

Le modèle apprend un “champ de vecteurs” : où que l’on soit dans le bruit, il indique la direction pour revenir vers la structure.

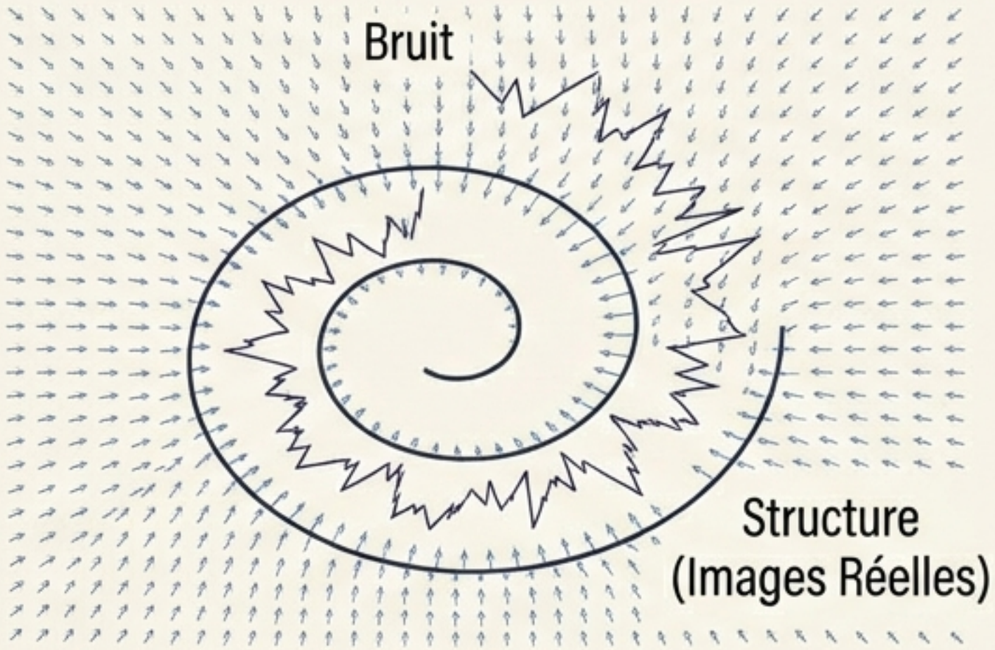
Le piège de la moyenne



Sans bruit, le modèle prédit la moyenne de toutes les possibilités. Cela mène au centre vide de la spirale, créant une image floue.

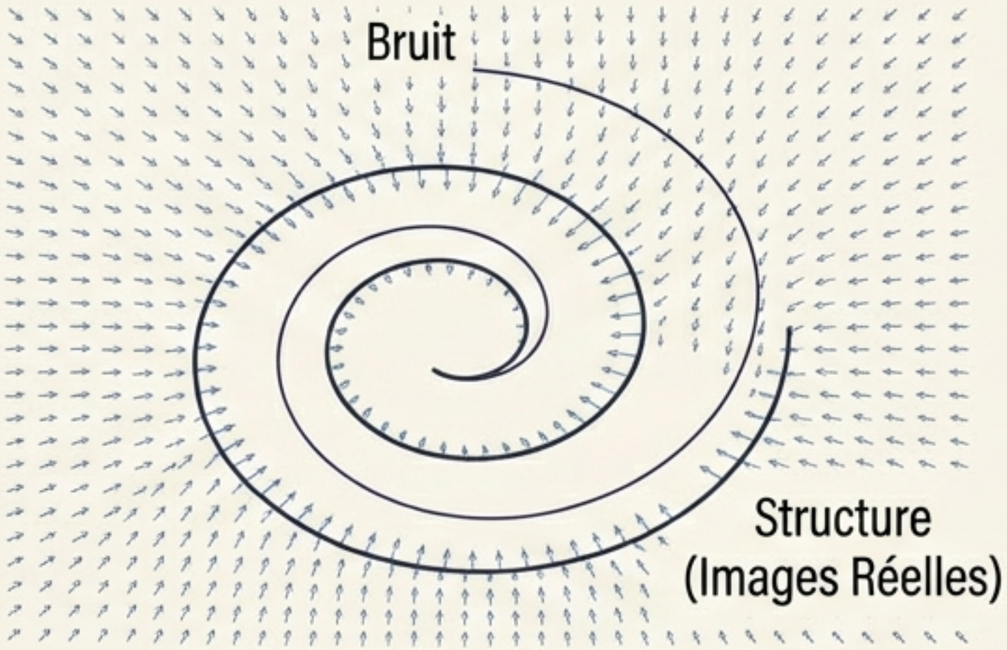
Deux chemins vers la netteté

DDPM - Stochastique



Marche Aléatoire (Lent)

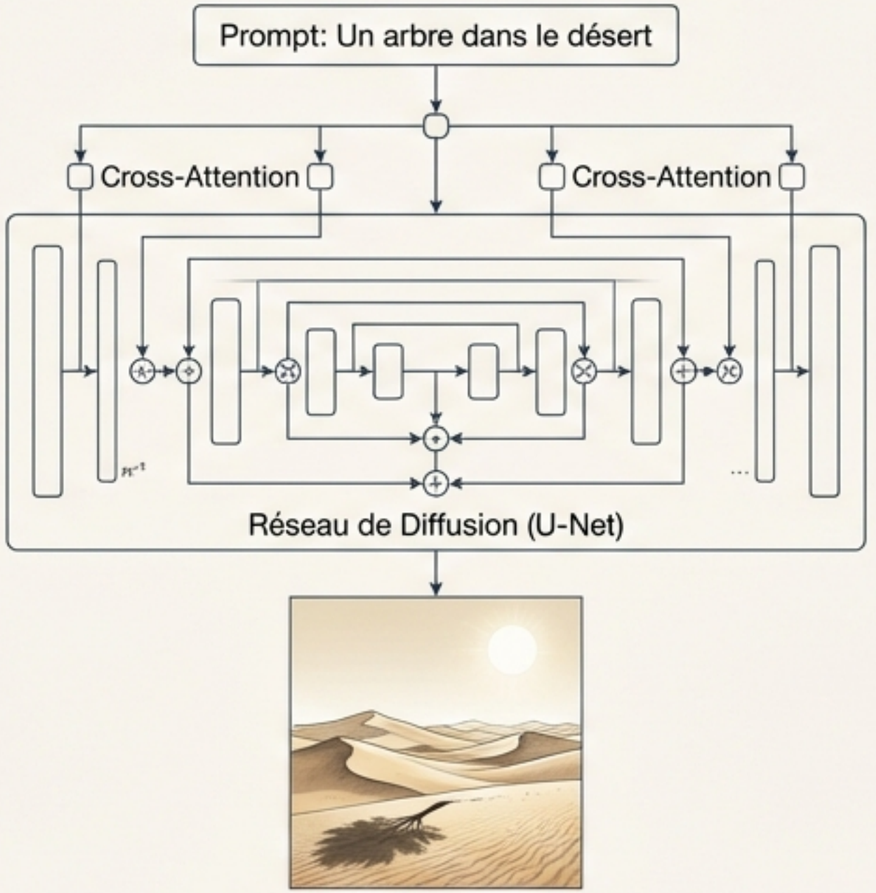
DDIM / Flow Matching



Équation Différentielle (Rapide)

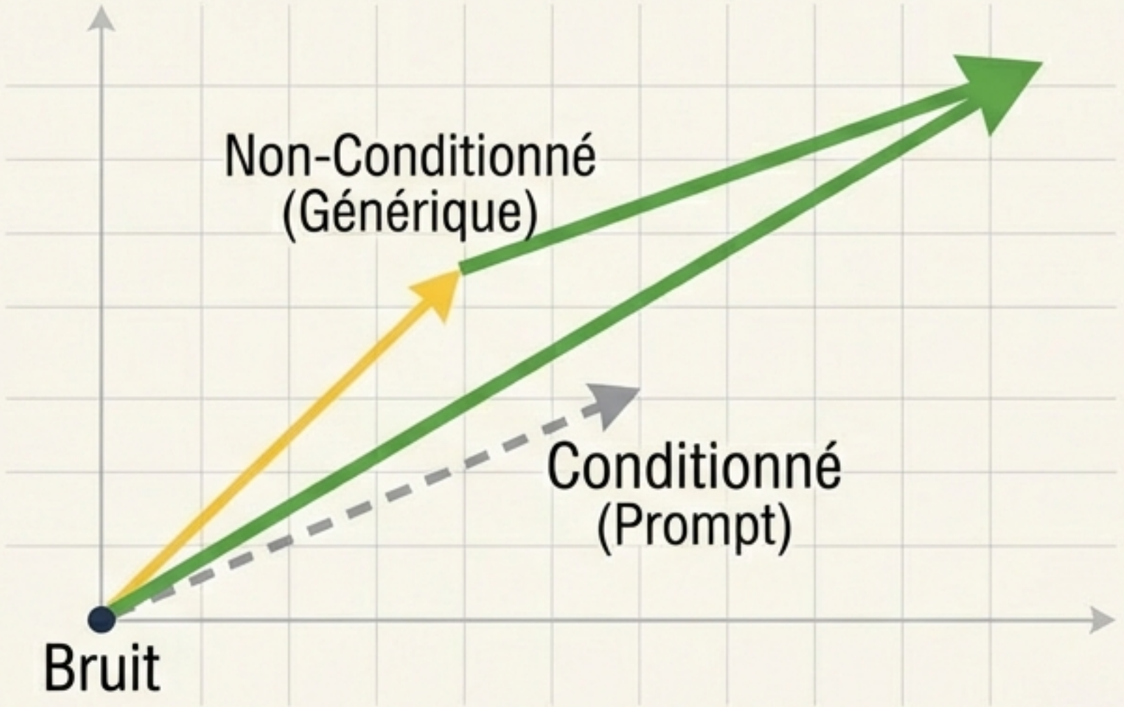
DDPM explore via le hasard. DDIM suit le flux direct.
C'est la base des modèles vidéo modernes comme WAN.

Le conditionnement



Donner une carte au modèle ne suffit pas toujours. Ici, le modèle a généré un désert générique, ignorant l'arbre demandé.

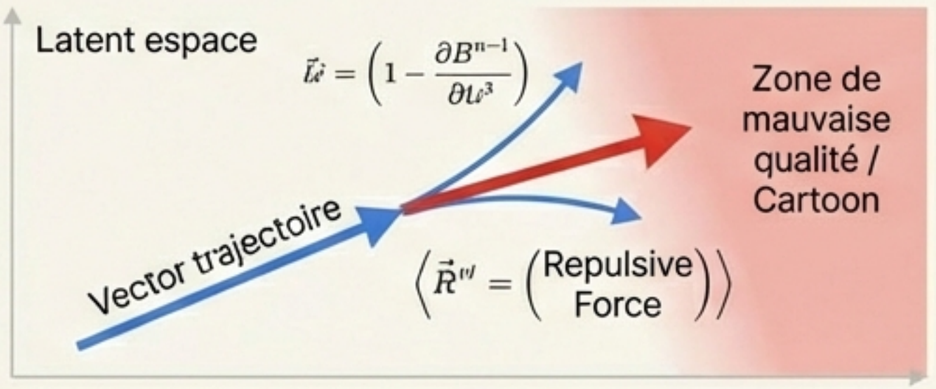
Classifier-Free Guidance



Résultat = Non_Cond + échelle * (Cond - Non_Cond)

On amplifie la différence entre 'ce que veut le prompt' et
'ce que ferait le modèle sans instruction'.

Les prompts négatifs

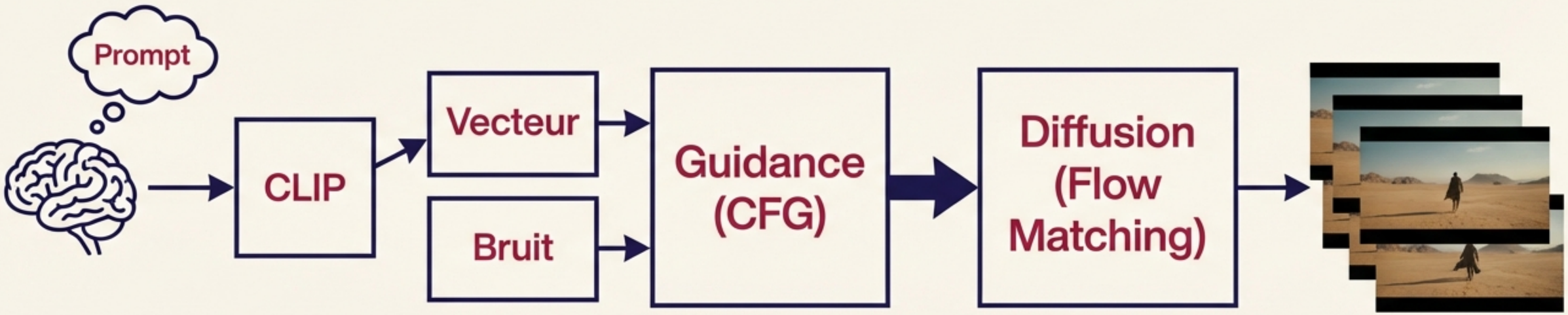


Sans prompt négatif

Avec prompt négatif

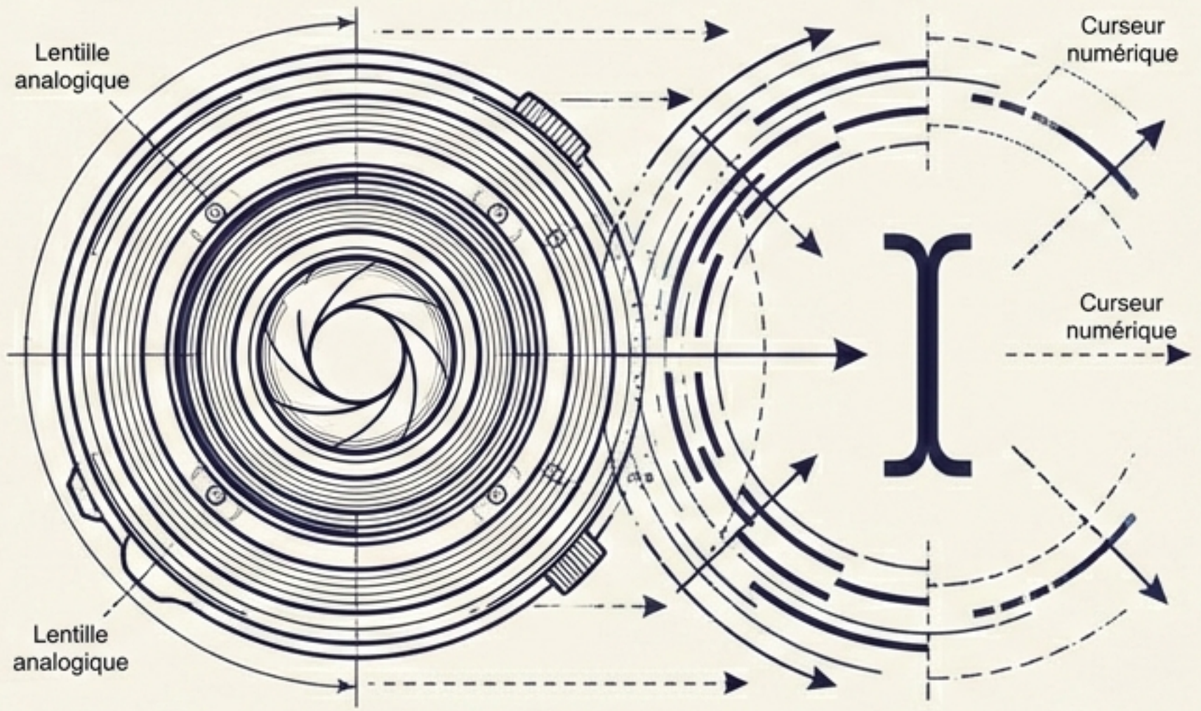
Mathématiquement, on soustrait le vecteur des concepts indésirables (ex: ‘doigts en trop’) pour éloigner la génération de ces zones.

La synthèse : De la pensée aux pixels



Le texte guide, la guidance amplifie, et la diffusion sculpte. C'est l'union de la compréhension sémantique et de la physique statistique.

Une nouvelle classe de machine



Nous ne capturons plus des photons. Nous naviguons dans un espace mathématique. Pour créer, nous n’avons plus besoin de caméras. Il suffit de mots.

Sources & Remerciements

Basé sur l’analyse technique de Stephen Welsh pour la chaîne 3Blue1Brown.

Vidéo : ‘But how do AI images and videos actually work?’

Présentation générée par IA.