Comment une IA crée des images : le cours complet, zone 3 Expert

Tout ce que le jeu explique, station par station, à lire d'une traite. Zone 3, niveau expert.

Cours complet de La ligne IA · laligne-ia.fr · version expert (Zone 3) · mis à jour le 20 novembre 2026. Le jeu gratuit pour comprendre comment marche une IA.

A RER A : Les bases, sous le capot

Pixels : Une image, c'est des nombres

Pour une IA, une photo n'est qu'un grand tableau de **pixels**. Chaque pixel est fait de trois nombres, entre 0 et 255 : la quantité de rouge, de vert et de bleu. Une IA ne voit jamais de « tarte » : elle voit des millions de nombres.

Une photo de 12 mégapixels contient 36 millions de nombres, trois par pixel. C'est beaucoup trop pour un modèle : la plupart des IA d'image commencent donc par compresser l'image dans une version plus petite et plus abstraite, qu'on appelle l'espace latent. Elles travaillent dans cet espace, puis reconvertissent le résultat en pixels à la fin.

> **Sous le capot :** Une image de 1024 × 1024 pixels, c'est plus de 3 millions de nombres. Les générateurs travaillent souvent dans un espace latent compressé, environ 8 fois plus petit de chaque côté, pour aller beaucoup plus vite.

> **À retenir : Pour une IA, une image est un tableau de nombres : trois par pixel, pour le rouge, le vert et le bleu.**

Morceaux : Découper l'image en carrés

Comme pour le texte, l'IA découpe l'image en petits morceaux : des carrés de quelques pixels, appelés **patchs**. Chaque carré devient l'équivalent d'un mot. Puis, grâce à l'attention, l'IA regarde quels carrés vont ensemble.

C'est le principe des « Vision Transformers » : la même architecture que pour le texte, appliquée à une suite de carrés au lieu d'une suite de mots. C'est ce qui permet à un même modèle de lire une photo et une phrase, et de répondre à une question sur une image. Les vraies IA ne comparent pas que la couleur : elles repèrent des formes, des textures et des objets.

> **Sous le capot :** Les modèles de vision de type ViT découpent l'image en patchs, par exemple de 16 × 16 pixels, transformés en vecteurs puis traités par un Transformer, exactement comme des tokens de texte.

> **À retenir : L'IA traite une image comme une phrase : une suite de carrés qu'elle relie entre eux.**

Associer : Relier les images et les mots

Pour qu'une IA crée « une tarte aux fraises », elle doit d'abord savoir à quoi ça ressemble. On l'entraîne avec des centaines de millions d'images accompagnées de leur **légende**. Elle apprend à placer l'image et sa description au même endroit sur sa carte du sens.

Une méthode célèbre, CLIP, a été entraînée sur des centaines de millions de paires image-légende récupérées sur internet. Le modèle apprend à rapprocher une image de sa vraie légende et à l'éloigner des autres. Résultat : les mots et les images partagent un même espace. C'est ce pont qui permet ensuite de guider la création d'une image avec du texte.

> **Sous le capot :** Des modèles comme CLIP apprennent sur des centaines de millions de paires image-légende à placer une image et sa description au même endroit d'un espace commun. On parle d'apprentissage contrastif.

> **À retenir : L'IA apprend le lien entre les mots et les images en voyant des millions d'images légendées.**

B RER B : Fabriquer une image

Le bruit : Abîmer des images exprès

Pour apprendre à créer des images, on fait d'abord l'inverse : on prend de vraies photos et on les **noie dans le bruit**, petit à petit, jusqu'à ce qu'il ne reste qu'une neige de télé. À chaque étape, on demande au modèle de deviner le bruit qu'on vient d'ajouter.

On répète l'exercice sur des centaines de millions d'images, avec toutes les doses de bruit possibles. Le modèle devient un expert pour repérer, dans une image brouillée, ce qui est du bruit et ce qui est de l'image. Une fois entraîné, il peut partir d'un bruit pur et faire le chemin inverse. C'est la station suivante.

> **Sous le capot :** On ajoute un bruit gaussien selon un calendrier précis, sur des centaines d'étapes. Le réseau, souvent un U-Net ou un Transformer de diffusion, apprend à prédire le bruit ajouté à chaque étape.

> **À retenir : Pour apprendre à créer, l'IA s'entraîne à reconnaître le bruit ajouté à de vraies images.**

Débruiter : L'image sort du brouillard

Pour créer une image, l'IA part d'une neige de bruit tirée au hasard. À chaque étape, elle devine quelle part est du bruit, et en retire un peu. Après quelques dizaines d'étapes, une image nette apparaît. On appelle ça la **diffusion**.

Il faut en général de 20 à 50 étapes. Les grandes formes et les couleurs apparaissent d'abord, les détails à la fin. Beaucoup de générateurs d'images, comme Stable Diffusion, reposent sur ce principe. Certains générateurs récents combinent d'autres méthodes, mais l'idée du brouillard qui se dissipe reste la plus répandue.

> **Sous le capot :** Pour générer, on part d'un bruit pur et on applique le réseau de 20 à 50 fois, en retirant à chaque passe une partie du bruit prédit. Un décodeur repasse ensuite de l'espace latent aux pixels.

> **À retenir : Une IA d'image part d'un bruit au hasard et le retire étape par étape, du flou vers le net.**

Guider : Le prompt tient le volant

Si on part toujours du bruit, comment obtenir une tarte plutôt qu'une pizza ? Le texte de la demande, le **prompt**, guide chaque étape du débruitage. Et le bruit de départ, tiré au hasard, s'appelle la **graine**.

La demande est transformée en vecteurs, comme pour les IA de texte, grâce au pont entre mots et images vu à la station Associer. À chaque étape, le modèle compare ce qu'il ferait avec et sans la demande, et exagère la différence pour coller au texte. Ce réglage, souvent appelé « guidance », peut être poussé : trop fort, les images deviennent saturées et caricaturales.

> **Sous le capot :** Le texte encodé guide chaque étape. Le guidage sans classifieur compare la prédiction avec et sans le texte pour pousser l'image vers la demande. La graine fixe le bruit de départ : même graine et même prompt donnent la même image.

> **À retenir : Le prompt guide chaque étape ; la graine décide du point de départ. Même graine et même demande donnent la même image.**

Décrire : Écrire un bon prompt

Plus la demande est précise, plus l'image ressemble à ce qu'on imagine. Un bon prompt donne le **sujet**, mais aussi la **lumière**, le **cadrage** et le **style**. Exactement les mots d'un photographe.

Les générateurs ont appris sur des images légendées par des humains : ils réagissent donc au vocabulaire de la photo et du dessin (contre-jour, vue de dessus, faible

profondeur de champ, aquarelle…). Une demande vague laisse l'IA choisir à ta place, en général le cliché le plus courant. Une demande précise, c'est toi qui décides.

> **Sous le capot :** Les mots du prompt deviennent des vecteurs qui pèsent sur l'image. Les termes précis (« éclairage rasant », « vue en plongée ») déplacent bien plus le résultat que les adjectifs vagues comme « beau ».

> **À retenir : Un bon prompt décrit le sujet, la lumière, le cadrage et le style. Le vague donne du banal.**

C RER C : Limites, sécurité, société

Animer : Une vidéo, c'est des images qui se tiennent

Une vidéo, c'est une suite d'images, 24 ou 30 par seconde. Pour en générer une, l'IA doit débruiter toutes ces images en même temps, et surtout les garder **cohérentes** : même assiette, même nombre de fraises, même lumière d'une image à l'autre.

Les IA vidéo traitent le temps comme une dimension de plus : elles découpent la vidéo en petits blocs d'espace et de temps, et l'attention relie ces blocs entre eux. C'est beaucoup plus coûteux qu'une image, d'où des vidéos souvent courtes. Les défauts typiques : objets qui apparaissent ou disparaissent, mains qui changent de forme, physique approximative.

> **Sous le capot :** Les modèles vidéo débruitent un bloc d'images en même temps, avec une attention qui relie les images entre elles dans le temps : c'est elle qui assure la cohérence d'une image à la suivante.

> **À retenir : Générer une vidéo, c'est générer des dizaines d'images qui doivent rester cohérentes entre elles.**

L'empreinte : Le coût caché des images et des vidéos

Générer une image consomme bien plus d'énergie qu'une réponse en texte. Et une vidéo, c'est des dizaines d'images à débruiter ensemble : l'addition explose. Électricité, **eau** pour refroidir les machines, fabrication des puces : chaque création a une **empreinte**.

Selon une enquête du MIT Technology Review, une image en haute définition consomme à peu près l'énergie d'un micro-ondes allumé quelques secondes, et une vidéo de 5 secondes jusqu'à l'équivalent d'une heure de micro-ondes avec certains modèles. Ces chiffres varient beaucoup selon les modèles et sont discutés, mais l'ordre de grandeur est clair : la vidéo coûte des centaines de fois plus qu'une image. Une partie de cette énergie est perdue en essais ratés : un prompt bien écrit dès le départ, c'est aussi moins de régénérations.

> **Sous le capot :** Une image demande des dizaines de passes du réseau ; une vidéo, des centaines d'images. La consommation grimpe avec la résolution, le nombre d'étapes et la durée.

> **À retenir : Image et surtout vidéo coûtent cher en énergie et en eau. Un bon prompt évite de régénérer dix fois.**

Les ratés : Les erreurs typiques des IA d'image

Les IA d'image font des erreurs typiques : des **mains** avec trop ou pas assez de doigts, du **texte** sans queue ni tête, des **ombres** ou des reflets impossibles. Elles reproduisent l'apparence des choses, sans toujours en comprendre la logique.

Ces erreurs viennent de ce que l'IA a appris : elle connaît l'allure générale d'une main ou d'un panneau, mais pas les règles strictes (cinq doigts, une orthographe, une seule source de lumière). Les modèles récents se sont nettement améliorés sur les mains et le texte. Ces indices sont donc de moins en moins fiables pour repérer une image générée.

> **Sous le capot :** Le modèle apprend des corrélations visuelles, pas des règles : nulle part il n'est écrit qu'une main a cinq doigts. Plus de données et de plus gros modèles réduisent ces erreurs sans les supprimer.

> **À retenir : L'IA imite l'apparence des choses sans toujours en respecter la logique. Mais elle progresse vite.**

Le faux : Vraie photo ou image générée ?

Les images et vidéos générées sont devenues très réalistes : on parle de **deepfakes** quand elles imitent de vraies personnes ou de vrais événements. À l'œil, c'est de plus en plus difficile. Il faut donc s'appuyer sur des indices plus solides que l'impression.

Des systèmes de marquage se développent : certaines IA ajoutent un filigrane invisible dans leurs images, et le standard C2PA (« Content Credentials ») permet d'attacher à un fichier l'historique de sa création. Mais ces marques peuvent être absentes ou retirées, et les détecteurs automatiques se trompent souvent. Le meilleur réflexe reste de remonter à la source.

> **Sous le capot :** La norme C2PA ajoute au fichier un historique signé, les Content Credentials, et certains générateurs insèrent un filigrane invisible. Mais une simple capture d'écran efface souvent ces traces.

> **À retenir : À l'œil, on ne peut plus trancher. Les vrais indices : la source, l'historique de l'image, et les marquages de provenance.**

Les sources : D'où viennent les images d'entraînement ?

Pour apprendre, les IA d'image ont vu des centaines de millions d'images, souvent collectées sur internet : photos, illustrations, œuvres d'art. Beaucoup ont des **auteurs** qui n'ont jamais été consultés. C'est l'un des grands débats autour de ces IA.

En Europe, la loi autorise la fouille de textes et de données sur les contenus accessibles en ligne, sauf si l'auteur s'y est opposé de façon lisible par les machines (dans les conditions du site, les métadonnées ou les fichiers techniques). Le règlement européen sur l'IA impose en plus aux fournisseurs de grands modèles de respecter ces refus et de publier un résumé de leurs données d'entraînement. Plusieurs procès opposent des auteurs ou des banques d'images aux éditeurs d'IA. Au Royaume-Uni, Getty Images a en grande partie perdu en 2025 face à Stability AI : la justice a estimé qu'un modèle qui ne stocke pas les images n'en est pas une copie. Aux États-Unis, des auteurs ont obtenu un accord de 1,5 milliard de dollars avec Anthropic. Le sujet est loin d'être tranché, et il se joue pays par pays.

> **Sous le capot :** Le débat juridique porte sur deux moments : la copie des images pour l'entraînement, encadrée en Europe par l'exception de fouille de données et le droit d'opposition, et les images produites, quand elles ressemblent trop à une œuvre existante.

> **À retenir : Les IA d'image ont appris sur des œuvres de vrais auteurs. En Europe, un auteur peut s'opposer à cette utilisation, et le débat juridique continue.**

Le débat : Créer des images : progrès ou menace ?

Les IA d'image ouvrent des possibilités incroyables : visualiser une idée en quelques secondes, restaurer de vieilles photos, rendre la création accessible à tous. Mais elles posent aussi des **questions** : le travail des illustrateurs et des photographes, la confiance dans les images, la mémoire visuelle collective, l'énergie consommée.

Les images ont toujours servi de preuve : « je l'ai vu en photo ». Avec la génération d'images, cette évidence disparaît, et les archives de demain risquent de mélanger vrai et faux. Pour les métiers de l'image, les commandes simples (illustrations d'articles, visuels de catalogue) sont les premières concernées, tandis que le travail d'auteur, le regard et la direction artistique restent difficiles à remplacer. Les règles se construisent : obligation d'indiquer les contenus générés dans certains cas en Europe, marquages de provenance, droit d'opposition des auteurs.

> **Sous le capot :** Les enjeux : la valeur du travail des créateurs, la confiance dans les images comme preuve, et l'énergie. Le règlement européen sur l'IA prévoit de signaler clairement les contenus générés de type deepfake.

> À retenir : Les IA d'image sont de formidables outils et de vrais défis. Ce qu'on en fera dépend des règles, et de l'usage de chacun.

Les autres cours :