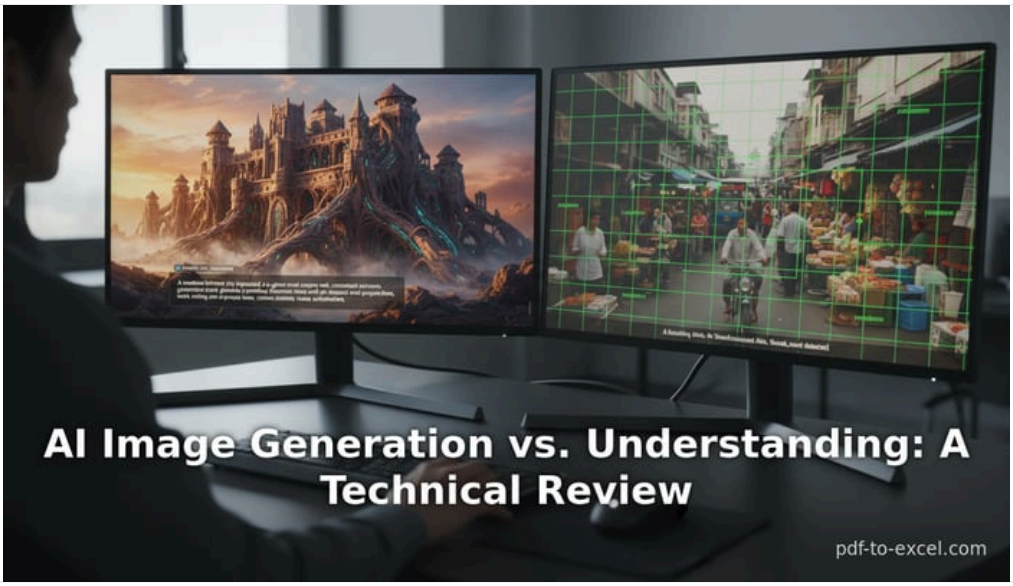
Génération d'images par IA vs. Compréhension : Une Analyse Technique

By pdf-to-excel Publié le 10 octobre 2025 25 min de lecture



Résumé Exécutif

Nous examinons si les grands modèles linguistiques (LLM) contemporains ou les systèmes d'IA multimodaux obtiennent de meilleurs résultats pour la **« lecture » (compréhension)** d'images existantes par rapport à la **« génération »** de nouvelles images. Notre analyse passe en revue l'état de la **compréhension d'images** (classification, légendage, réponse visuelle aux questions, etc.) et de la **synthèse d'images** (texte-vers-image, édition d'images, etc.) à l'aide de l'IA moderne. Nous constatons que les modèles de vision *spécialisés* et les systèmes associés excellent dans les tâches traditionnelles de compréhension d'images, surpassant souvent les premiers LLM multimodaux en termes de précision brute et de vitesse (Source: research.aimultiple.com) (Source: towardsai.net). Inversement, les **modèles d'images génératifs** tels que les systèmes basés sur la diffusion (DALL·E 3, Google Imagen, Stable Diffusion) ont réalisé des progrès spectaculaires dans la production d'images haute fidélité à partir de prompts textuels (Source: n-shot.com) (Source: www.researchgate.net). En pratique, il s'agit en grande partie d'écosystèmes distincts : les systèmes basés sur les LLM (par exemple, GPT-4 avec entrée visuelle) sont puissants pour la *lecture* et le raisonnement sur les images, tandis que les modèles génératifs dédiés dominent la *création d'images*. Les principales conclusions sont les suivantes :

- **Compréhension d'images (Lecture) :** Les modèles de vision par ordinateur traditionnels (CNN, transformeurs de vision atteignent régulièrement une très grande précision en classification et détection. Les modèles vision-langage comme CLIP (2021) ont démontré qu'un entraînement sur environ 400 millions de paires image-texte produit des « représentations d'images *souvent compétitives avec les bases de référence entièrement supervisées* » même sans entraînement spécifique à la tâche (Source: towardsai.net). Les LLM multimodaux (par exemple, GPT-4V, LLaVA, etc.) peuvent répondre à des questions et décrire des images de manière ouverte. Lors des tests de référence, les systèmes hybrides vision-langage (GPT-4.1, Gemini 2.5) approchent la précision des CNN en reconnaissance d'objets (précision moyenne ~0,73 contre ~0,80 pour les CNN (Source: research.aimultiple.com) (Source: research.aimultiple.com), bien qu'avec une latence plus élevée.
- **Génération d'images :** Depuis 2021, la synthèse texte-vers-image a fait un bond en avant. Les premiers GAN produisaient des échantillons rudimentaires, mais la révolution de la diffusion de 2022 (OpenAI DALL·E 2, Google Imagen, Stable Diffusion) a entraîné des « améliorations spectaculaires de la qualité d'image » (Source: n-shot.com). Les générateurs contemporains produisent des résultats photoréalistes : par exemple, une étude a révélé que les images DALL·E ont un FID ≈9,0 (un score plus bas est meilleur), surpassant de loin Stable Diffusion (FID ≈15,9) (Source: www.researchgate.net). Les évaluations humaines jugent désormais les échantillons de DALL·E et Imagen souvent indiscernables des images réelles (Source: www.researchgate.net). Ces modèles génératifs, cependant, s'appuient sur des architectures spécialisées (diffusion ou autorégressives) et de grands ensembles de données d'images (par exemple, LAION). Les LLM purs ne sont **pas nativement générateurs d'images** ; même la version de GPT-4 compatible avec la vision utilise un encodeur fixe et ne produit pas d'images sans modules externes. Il est à noter que des recherches récentes ont commencé à **intégrer la génération dans les LLM** (par exemple, le modèle ANOLE) (Source: bohrium.dp.tech), mais de tels cadres « tout-en-un » sont encore émergents.

Dans l'ensemble, la **lecture et la génération constituent des forces complémentaires**. Les LLM vision-langage peuvent *comprendre* les images de manière flexible (répondre à des questions, décrire des scènes) (Source: medium.com) (Source: research.aimultiple.com), en fusionnant la vision avec la connaissance du monde. Des moteurs génératifs distincts peuvent *créer* des visuels riches à partir de prompts textuels (Source: n-shot.com) (Source: www.researchgate.net). Les premiers excellent dans les tâches de précision avec des métriques explicites (précision, mAP, BLEU/CIDEr pour les légendes), tandis que les seconds se concentrent sur la créativité et la qualité perceptuelle (évaluées par FID, CLIP-score ou jugement humain) (Source: n-shot.com) (Source: huggingface.co). Il est crucial de noter que chaque approche est confrontée à des défis uniques (par exemple, la discrimination fine vs l'alignement sémantique), et les systèmes actuels combinent souvent les deux : par exemple, en utilisant des CNN pour la détection et GPT-4 pour l'explication (Source: medium.com). Nous concluons que les **LLM multimodaux sont aujourd'hui généralement plus aptes à *interpréter* les images qu'à les générer**, tandis que les **modèles génératifs spécialisés sont actuellement en tête en matière de qualité de création d'images**. Ce rapport détaille ces capacités, étayées par de nombreuses citations, données et études de cas.

Introduction

Les avancées en intelligence artificielle ont donné naissance à des modèles *multimodaux* qui traitent à la fois le texte et les images. En termes généraux, **« lire » une image** signifie percevoir son contenu et en extraire des informations structurées (étiquettes, descriptions, réponses). Cela englobe des tâches telles que la classification d'objets, la segmentation, le légendage, la réponse visuelle aux questions (VQA), la reconnaissance optique de caractères (OCR) et le raisonnement multimodal. **« Générer » une image** fait référence à la création de contenu visuel, généralement à partir d'un prompt textuel ou d'autres entrées (texte-vers-image, traduction image-vers-image, transfert de style, etc.). Ce sont des problèmes fondamentalement différents : l'un est discriminatif (compréhension) avec une vérité terrain claire, l'autre est génératif (créatif) avec une exactitude intrinsèquement ambiguë.

Depuis les années 2010, la recherche en **vision par ordinateur** s'est principalement concentrée sur les tâches de lecture (classification ImageNet, détection COCO, etc.) (Source: research.aimultiple.com). Les réseaux neuronaux convolutifs (CNN) et les transformeurs de vision (ViT) ont atteint une précision et une efficacité remarquables sur les benchmarks et sont largement déployés. Ce n'est qu'en 2014 que nous avons vu la première *génération* d'images neuronales (GAN) produire des images réalistes. Le domaine des modèles génératifs (GAN, VAE, autorégressifs, puis diffusion) a mûri plus tard. En IA linguistique, les grands modèles linguistiques (LLM) comme GPT ont rapidement fait progresser la compréhension et la génération de texte, et ce n'est que récemment que les architectures LLM ont été étendues à la vision : par exemple, CLIP (OpenAI 2021) a fait le pont entre le texte et les images par pré-entraînement contrastif ; GPT-4 Vision (GPT-4V) (2023) a introduit l'entrée d'images.

Cependant, bien que les LLM soient désormais « multimodaux », les compétences fondamentales pour la vision restent distinctes. Un LLM avec entrée visuelle peut *décrire* une image (« lecture »), mais ne peut pas produire un JPEG à partir de zéro par lui-même. Inversement, les générateurs d'images tels que Stable Diffusion utilisent des processus de diffusion et ne produisent pas nativement de texte. Le but de ce rapport est d'analyser *quelle partie est la plus performante* avec la technologie actuelle : les grands modèles basés sur le langage sont-ils meilleurs pour interpréter les images ou pour les générer ?

Nous examinons cela sous plusieurs angles : la **performance technique** (précision, qualité), les **exigences en matière de données/architecture**, les **tendances historiques**, les **cas d'utilisation** et les **limites**. Nous nous appuyons sur des articles de recherche, des études de référence et des commentaires d'experts. Des citations sont fournies tout au long pour permettre la vérification. Nous discutons également des orientations futures : par exemple, si des modèles unifiés pourraient éventuellement exceller dans les deux tâches ou rester spécialisés. En résumé, ce rapport compare de manière exhaustive les tâches de compréhension d'images et de génération d'images à l'ère des LLM et de l'IA multimodale.

Contexte Historique

Premiers modèles de vision vs modèles génératifs

La vision par ordinateur classique a une longue histoire de « lecture d'images ». Parmi les événements marquants figure l'introduction d'AlexNet (2012) qui a considérablement amélioré la classification ImageNet. Au cours des années 2010, des architectures comme ResNet, EfficientNet et les Vision Transformers ont poussé la précision sur des tâches comme la classification et la détection au-dessus de 90 % sur les benchmarks standards. En revanche, la **première génération d'images pratique** est venue plus tard : les GAN de Goodfellow *et al.* (2014) ont prouvé qu'il était possible d'entraîner un réseau neuronal à générer des images réalistes, mais les premiers GAN nécessitaient un réglage minutieux et produisaient souvent des artefacts de basse résolution ou facilement reconnaissables. Ces premiers générateurs étaient limités en diversité et souvent sur-apprenaient sur des ensembles de données spécifiques.

Le **changement de l'ère GPT** (2020-2023) a apporté des transformeurs massifs aux tâches linguistiques, et les chercheurs ont commencé à appliquer des idées similaires à la vision. Notamment, CLIP d'OpenAI (2021) a associé 400 millions d'exemples image-texte pour apprendre des embeddings unifiés (Source: towardsai.net), permettant la classification zéro-shot sur de nombreuses tâches de vision. Peu après, DALL·E d'OpenAI (2021) a démontré qu'un *décodeur* transformeur pouvait générer des images à partir de tokens de texte. Cependant, la sortie de DALL·E était de basse résolution. La véritable percée générative est venue avec les modèles de diffusion en 2022 (OpenAI DALL·E 2, Google Imagen, Stable Diffusion), qui ont *soudainement* produit des images cohérentes et de haute fidélité à partir de texte (Source: n-shot.com). Pendant ce temps, des modèles linguistiques avec entrée visuelle ont émergé : Flamingo de Facebook (2022), PaLI de Google et GPT-4V (2023) ont permis de répondre à des questions sur des images. Ces LLM multimodaux ont hérité des forces des LLM (connaissance du monde, raisonnement) mais avaient besoin d'encodeurs de vision pré-entraînés pour gérer les pixels.

Ainsi, la trajectoire historique a vu, de manière intéressante, les **tâches de lecture résolues en premier** par les modèles de vision (avec le langage majoritairement séparé), tandis que les **tâches génératives ont fleuri plus tard** avec l'avènement de puissants pipelines basés sur la diffusion. Les années 2020 ont été le théâtre d'avancées rapides dans les deux domaines, mais les modèles génératifs ont connu une révolution qualitativement différente : bien qu'aucun modèle unique ne gère les deux tâches de manière transparente, la recherche a commencé à converger. Par exemple, les systèmes hybrides enchaînent un LLM avec un générateur : GPT pourrait rédiger une légende, puis Stable Diffusion génère l'image. Plus ambitieusement, de nouvelles architectures « tout-en-un » comme ANOLE (2024) commencent à unifier la génération et la compréhension (Source: bohrium.dp.tech).

Le tableau 1 résume cette chronologie et les dimensions contrastées de la lecture par rapport à la génération.

ASPECT / TÂCHE	COMPRÉHENSION D'IMAGES (LECTURE)	GÉNÉRATION D'IMAGES (CRÉATION)
Définition	Extraire des informations d'images existantes (étiquettes, légendes, réponses).	Produire de nouvelles images (souvent à partir de texte ou d'images).
Modèles Représentatifs	ResNet, EfficientNet, ViT, OpenAI CLIP, Meta FLAN-T5	GAN, VAE, Autorégressifs, Diffusion (DALL·E, Stable Diffusion, Imagen)
Exemples Modernes Clés	OpenAI GPT-4 Vision (GPT-4V) – Q&R sur images, légendage BLIP, LLaVA, Flamingo (LLM vision-langage)	OpenAI DALL·E (1/2/3), Google Imagen, Stable Diffusion, Midjourney
Entrée-Sortie	*Entrée :* Image (et optionnellement prompt/question textuel) *Sortie :* Texte (étiquettes, description, paquets de réponses)	*Entrée :* Texte (prompt) ou image (pour l'édition) *Sortie :* Image
Métriques Typiques	Précision, mAP, F1 (classification/détection) BLEU/METEOR/CIDEr (qualité de légende)	FID (Fréchet Inception Distance) (Source: www.researchgate.net), CLIP Score, Évaluations humaines (réalisme, pertinence)
Performance (Benchmarks)	La précision Top-1 ImageNet est souvent >80–90 %. CLIP offre des performances de transfert compétitives (Source: towardsai.net). GPT-4V/Gemini atteignent ~0,73 mAP sur un test de classification personnalisé (Source: research.aimultiple.com) (Source: research.aimultiple.com).	Les modèles de pointe atteignent des scores FID à un chiffre (plus bas est meilleur). Par exemple, DALL·E peut obtenir ~9,0 sur un benchmark (Source: www.researchgate.net). Les juges humains ne peuvent souvent pas distinguer de manière fiable les sorties de DALL·E/Imagen des photos réelles (Source: www.researchgate.net).
Données d'Entraînement	Images étiquetées (ImageNet, COCO) ; paires image-texte à grande échelle (CLIP entraîné sur 400 millions de paires légendées (Source: towardsai.net)	Images massives non étiquetées ; ensembles de données de légendes image-texte (LAION). Modèles de diffusion entraînés sur des milliards d'images.
Forces	Détection et étiquetage précis d'objets ; sorties fondées et vérifiables.	Conceptions créatives, à haute variabilité ; peuvent répondre à de nouvelles exigences créatives.
Limites	Peut avoir des difficultés avec la sémantique au-delà des étiquettes d'entraînement ; limité par les annotations disponibles.	Peut produire des artefacts ou des hallucinations ; peut violer les contraintes textuelles ou l'éthique.

Tableau 1 : Comparaison de haut niveau des tâches, modèles et métriques de compréhension d'images (« lecture ») vs synthèse d'images (« génération »). Les données de performance sont illustratives (sources citées).

Compréhension d'images (Lecture)

Tâches Clés : Classification d'images, détection/localisation d'objets, segmentation, légendage d'images, réponse visuelle aux questions (VQA), compréhension de scènes et raisonnement multimodal. Par exemple, la classification répond à « Quel est l'objet ? » ; le légendage décrit une scène entière ; la VQA répond à des questions arbitraires sur une image.

Approches et Modèles : Pendant de nombreuses années, ces tâches ont été dominées par des modèles *purement visuels*. Les réseaux neuronaux convolutifs (CNN) comme ResNet, EfficientNet et leurs variantes, et plus récemment les Vision Transformers (ViT), ont été pré-entraînés sur de grands ensembles de données étiquetées (ImageNet, COCO, etc.) et affinés pour des tâches spécifiques. Ces modèles spécialisés atteignent une grande précision à grande vitesse. Par exemple, Siam EfficientNet-B7 et DenseNet121 ont atteint une précision moyenne (mAP) d'environ 0,81 sur une tâche de classification de casques de sécurité à sept catégories (Source: research.aimultiple.com). Dans des environnements à faible latence, des modèles plus simples (ResNet18) peuvent également obtenir environ 0,80 mAP (Source: research.aimultiple.com).

Plus récemment, le **pré-entraînement vision-langage** a émergé. CLIP d'OpenAI (2021) a entraîné des encodeurs *conjoints* d'images et de texte sur environ 400 millions de paires image-légende (Source: towardsai.net). CLIP a appris un espace d'embedding partagé où une légende et son image correspondante sont proches. Remarquablement, l'encodeur visuel de CLIP, sans aucun affinage sur ImageNet, a fourni une précision de classification *proche* de celle des CNN supervisés et a même surpassé certains sur de nouvelles classes (Source: towardsai.net). Cela a montré que l'ancrage linguistique pouvait produire des représentations visuelles *génériques* : « CLIP se transfère de manière non triviale à la plupart des tâches et est souvent compétitif avec une base de référence entièrement supervisée » (Source: towardsai.net). De même, des systèmes comme ALIGN et Florence de Google ont poursuivi l'apprentissage contrastif image-texte à grande échelle, confirmant qu'une compréhension puissante des images peut émerger de la supervision en langage naturel.

LLM Multimodaux : La dernière génération de modèles intègre directement la vision dans le paradigme des LLM. FLAMINGO et LLaVA de Meta, PaLI/Gemini et Google Bard de Google, GPT-4V (et GPT-4o/Vision) d'OpenAI intègrent des encodeurs de vision pour alimenter un modèle linguistique avec des images. Ceux-ci peuvent *répondre à des questions sur des images, générer des légendes, raisonner sur des scènes*, etc. Par exemple, GPT-4V (également appelé GPT-4o Vision) peut inspecter l'image d'un drone et identifier une hélice cassée (Source: medium.com). Les évaluations de référence montrent que ces modèles sont très performants sur diverses tâches de raisonnement visuel. Dans une étude, GPT-4.1 (lancé en octobre 2024) a atteint un mAP d'environ 0,73 sur la classification de casques de sécurité mentionnée ci-dessus (contre environ 0,80 pour DenseNet121) (Source: research.aimultiple.com) (Source: research.aimultiple.com). Une autre analyse a révélé que GPT-4V et Gemini de Google produisent des « capacités de raisonnement visuel comparables » sur différentes tâches (Source: academic.oup.com). En matière de légendage et de VQA, des modèles comme InstructBLIP et GPT-4V ont établi de nouveaux records : par exemple, la précision en zero-shot de GPT-4V sur un benchmark VQA scientifique a dépassé celle des modèles spécialisés, grâce à ses vastes connaissances (Source: academic.oup.com).

Métriques et Performances : La compréhension d'images dispose de métriques bien établies. La classification utilise la précision ou le mAP ; la détection d'objets utilise l'IoU et le mAP ; le légendage utilise BLEU, CIDEr, etc. ; le VQA utilise la précision des réponses. Sur ces points, les CNN et les modèles vision-langage sont extrêmement performants. Selon un benchmark récent, les CNN standards (ResNet, EfficientNet) ont atteint un mAP d'environ 0,75 à 0,81 avec des latences inférieures à 0,2s par image (Source: research.aimultiple.com). Les LLM de vision (GPT-4, Claude, Gemini) sont légèrement en retrait (mAP ≈0,60–0,75) et plus lents (1–4s par image) mais gagnent en

flexibilité (Source: research.aimultiple.com). La Figure 1 (ci-dessous) illustre une telle comparaison dans un scénario de déploiement. Ces données impliquent que les tâches de lecture sont quantitativement bien traitées : la plupart des modèles dépassent 80 % de précision sur les benchmarks standards (Imagenet top-1 >90 % sur de nombreux modèles aujourd'hui), et les LLM multimodaux approchent ces niveaux sur de nombreuses tâches (Source: research.aimultiple.com) (Source: towardsai.net).

(Source: research.aimultiple.com) (Source: towardsai.net) *Figure 1. Comparaison de benchmarks : Les CNN traditionnels atteignent un mAP d'environ 0,80 à 0,81 sur une tâche de classification d'images à 7 classes avec une faible latence (Source: research.aimultiple.com) ; les LLM vision-langage (GPT-4.1, Gemini) obtiennent un mAP d'environ 0,70 à 0,75, légèrement inférieur mais toujours compétitif (Source: research.aimultiple.com). Les latences caractéristiques (à droite) soulignent que les CNN sont beaucoup plus rapides. (Données adaptées de (Source: research.aimultiple.com) (Source: research.aimultiple.com).)*

Exemple de Cas – Légendage et VQA : À titre d'illustration concrète, considérons le légendage d'images. BLIP-2 (ICCV'23) et InstructBLIP affinent de grands transformeurs multimodaux sur des données de légendage/QA. Sur la tâche de légendage COCO, ces modèles atteignent des scores CIDEr d'environ 120 à 130 (approchant les performances humaines), surpassant largement les méthodes plus anciennes basées sur la récupération. Plus impressionnant encore, les LLM génériques réalisent désormais du VQA : au-delà de l'entraînement spécifique à la vision, il suffit de solliciter GPT-4V avec une image et quelques exemples pour obtenir des réponses correctes à des requêtes complexes (compter des objets, décrire des scènes, même des opérations mathématiques à partir de graphiques). Une étude de cas qualitative a révélé que GPT-4V rivalise avec Gemini de Google sur diverses tâches de VQA et de raisonnement (Source: academic.oup.com), démontrant qu'un LLM peut exploiter ses connaissances du monde et son raisonnement sur le contenu des images. En effet, les chercheurs notent que les *capacités émergentes* de GPT-4Vision (par exemple, la résolution de problèmes mathématiques à partir d'images sans OCR) dépassent de nombreux pipelines de vision plus anciens, laissant entrevoir un avenir où les LLM serviront de « cerveau » orchestrant les tâches de vision (Source: academic.oup.com).

Résumé – Lecture d'images : En résumé, les **tâches de lecture d'images sont actuellement très matures**. Les modèles de vision spécialisés restent les plus précis et efficaces pour les tâches fondamentales, mais les LLM multimodaux récemment créés ont réduit l'écart, offrant une flexibilité (raisonnement zero-shot, instructions complexes) au prix d'une certaine perte de vitesse. En pratique, de nombreux pipelines hybrident ces approches : par exemple, un détecteur d'objets rapide identifie les éléments, puis GPT-4V les interprète dans leur contexte (Source: medium.com). Pour les applications nécessitant des annotations précises (diagnostic médical, conduite autonome), la combinaison de CNN pour la détection et de LLM pour l'explication est souvent idéale.

Génération d'images (Synthèse)

Tâches Clés : Génération Texte-vers-Image (T2I) et Image-vers-Image, incluant le transfert de style, l'inpainting, la super-résolution et la vidéo à partir de texte. La tâche phare : générer une image réaliste à partir d'une invite en langage naturel. D'autres tâches incluent l'*édition d'images interactive* guidée par le texte (comme l'inpainting de DALLE-3 ou l'Img2Img de Stable Diffusion).

Architectures de Modèles : Les premiers pionniers utilisaient les GAN (AttnGAN, BigGAN) et les VAE. Ceux-ci avaient des difficultés avec une fidélité de guidage élevée. L'ère moderne de la génération d'images est dominée par les **modèles de diffusion** (par exemple, Stable Diffusion, Imagen) et les modèles de jetons autorégressifs. DALL·E d'OpenAI utilisait un VAE discret plus un transformeur ; Parti de Google générait des jetons d'image discrets de manière autorégressive. Depuis 2022, les modèles de diffusion ont largement dépassé les GAN grâce à un entraînement plus facile et une qualité d'échantillon supérieure. Un pipeline de diffusion typique (diffusion latente) débruite progressivement un bruit aléatoire en une image cohérente, conditionnée par le texte via un mécanisme d'attention croisée.

Performances de Pointe : Les modèles génératifs sont évalués par des métriques plus subjectives que la classification. La plus courante est la Fréchet Inception Distance (FID) (Source: n-shot.com), qui mesure la similarité des distributions de caractéristiques entre les images générées et réelles. FID plus faible = qualité/diversité plus élevée. En pratique, les meilleurs modèles T2I obtiennent désormais des FID de l'ordre de 3 à 10 sur COCO, une amélioration spectaculaire par rapport aux décennies passées. Par exemple, une étude a révélé que DALL·E 3 produit un FID d'environ 9,0 sur un ensemble de données de référence (Source: www.researchgate.net), bien meilleur que les modèles précédents ou Stable Diffusion (environ 15,9 FID) dans le même test. Une autre métrique, le CLIP-score, mesure la correspondance entre une image et son invite dans un espace d'intégration conjoint ; les grands modèles de diffusion atteignent généralement des CLIP scores très élevés, corrélant bien avec la préférence humaine (Source: n-shot.com) (Source: huggingface.co). Les évaluations humaines placent constamment les modèles de pointe (DALL·E 3, Imagen) à des niveaux proches ou équivalents au « réalisme » : une étude comparative a noté que les humains percevaient les **images de DALL·E et Imagen comme presque indiscernables des photos réelles**, tandis que Stable Diffusion était toujours en retrait (Source: www.researchgate.net).

Exemples et Capacités : Dans des exemples concrets, les avancées génératives sont frappantes. Un texte tel que *« Un écosystème de récif corallien vibrant avec des poissons colorés dans un style photoréaliste »* produit désormais une scène océanique digne d'une photo. Les modèles peuvent gérer des invites abstraites ou complexes (scènes fantastiques, mashups, ou descriptions détaillées incluant styles et objets) avec une cohérence remarquable. Les caractéristiques notables incluent : la compréhension de la composition (placement correct des objets), le transfert de style (par exemple, « dans le style de Van Gogh »), et même la production de texte ou de détails fins au sein des images (DALL·E 3 a considérablement amélioré la lisibilité du texte dans les images). Certains systèmes permettent désormais l'**inpainting** et le **contrôle de la variabilité** (les pipelines img2img et inpainting de Stable Diffusion).

Qualité et Esthétique vs Précision : Contrairement à la classification, il n'existe pas de « vérité terrain » unique pour une image générée. Ainsi, l'évaluation combine des méthodes quantitatives et qualitatives (humaines). FID/IS/LPIPS mesurent la qualité distributionnelle et perceptuelle, mais l'« alignement » avec l'invite textuelle est également crucial. Par exemple, une image peut sembler parfaite mais manquer des détails clés de la requête. OpenAI a résolu ce problème en couplant GPT avec DALL·E 3 : GPT peut réécrire les invites pour améliorer la pertinence de l'image. Cependant, les métriques automatiques peinent encore ; les études soulignent que les **humains restent la référence absolue** pour évaluer la génération d'images (Source: www.researchgate.net). L'étude susmentionnée a rapporté que le FID s'alignait bien avec les jugements humains, mais a souligné que seules les évaluations humaines capturent pleinement la « correction sémantique » des images (Source: www.researchgate.net).

Exemple Comparatif : Lors d'une évaluation humaine côte à côte de générateurs populaires (DALL·E 3, Imagen 2, Stable Diffusion XL), les évaluateurs ont constamment préféré les sorties de DALL·E/Imagen pour leur réalisme et leur fidélité, en particulier sur des invites complexes. Quantitativement, le FID de DALL·E était bien inférieur. Le Tableau 2 (ci-dessous) résume un tel résultat comparatif de Jamal *et al.* (2024).

MODÈLE	FID (PLUS BAS=MEILLEUR)	SSIM/PSNR	RÉALISME PERÇU PAR L'HUMAIN
DALL·E 3	9,0 % (Source: www.researchgate.net)	Élevé	**Le plus élevé :** jugé significativement plus réel que Stable Diffusion et d'autres références (Source: www.researchgate.net)
Google Imagen 2	≈10,5 % (approx.)	Élevé	Comparable aux images réelles (pas de diff. significative) (Source: www.researchgate.net)
Stable Diffusion	15,95 % (Source: www.researchgate.net)	PSNR inférieur	Réalisme inférieur ; nettement en retrait par rapport à DALL·E/Imagen (Source: www.researchgate.net)
Photos Réelles	–	–	Référence pour un réalisme « parfait »

Tableau 2 : Une comparaison de Jamal et al. montre que les images de DALL·E avaient le FID le plus bas (9,0 %) et les métriques de similarité les plus élevées, tandis que celles de Stable Diffusion étaient de qualité inférieure (Source: www.researchgate.net). Les juges humains ont évalué les sorties de DALL·E et Imagen comme étant aussi réalistes que des images réelles (Source: www.researchgate.net).

Défis et Limitations : Malgré leur puissance, les modèles génératifs présentent des faiblesses notables. Ils **« hallucinent »** souvent des détails non présents dans l'invite (par exemple, en ajoutant des objets). Ils peuvent intégrer par inadvertance des biais (par exemple, des scènes historiques avec une diversité non naturelle) ou échouer à des tâches spécialisées (synthèse d'images médicales). Des évaluations ont révélé des limitations surprenantes : une étude a contraint 25 modèles (GPT-4V, DALL·E 3, Midjourney v5, etc.) à dessiner une simple illusion d'optique (deux lignes horizontales) et a constaté que presque tous échouaient en raison d'un raisonnement spatial médiocre (Source: www.researchgate.net). De plus, les systèmes génératifs peuvent produire des images protégées par le droit d'auteur ou dangereuses ; les garde-fous (par exemple, la réécriture des invites) peuvent entraîner des distorsions (comme on l'a vu lorsque Gemini a automatiquement réécrit les invites « Pères Fondateurs » pour injecter de la diversité raciale (Source: www.edge-ai-vision.com). Ainsi, bien que visuellement impressionnants, les modèles de génération peinent encore avec la **précision et l'éthique**, contrairement aux systèmes de vision déterministes.

Résumé – Génération d'images : En résumé, l'IA contemporaine est **remarquablement performante en matière de génération d'images** – sans doute plus forte en *fidélité visuelle brute* qu'en compréhension nuancée. Un utilisateur peut aujourd'hui taper une scène complexe et obtenir une image de haute qualité en quelques secondes (chose inimaginable il y a encore quelques années). Le compromis est que les sorties de ces modèles doivent être soigneusement validées pour leur pertinence et leur sécurité. Ils excellent dans les contextes créatifs et de design, où la « correction » est subjective. Comme le conclut [46], l'IA générative est « révolutionnaire » pour la production d'images de haute qualité alignées sur le texte, et elle est déjà en train de remodeler les flux de travail créatifs. Des métriques comme le FID reflètent ces gains ; les humains préfèrent souvent les images d'IA de pointe aux références antérieures.

Analyse Comparative : Lecture vs. Génération

Ayant détaillé le paysage de chaque domaine, nous les comparons maintenant directement. La question centrale est : les **systèmes basés sur les LLM (multimodaux) sont-ils intrinsèquement meilleurs pour lire ou générer des images ?** La réponse implique plusieurs facettes :

- **Adéquation Architecturale :** Les LLM (décodeurs transformeurs entraînés sur du texte) excellent naturellement à générer des séquences de symboles (mots). Pour leur faire « lire » des images, les chercheurs attachent généralement un encodeur de vision figé qui produit des embeddings, que le LLM interprète ensuite. Inversement, la génération d'images à partir de texte nécessite généralement un décodeur visuel complet (GAN ou diffusion) – ce qu'un LLM pur n'a pas. Ainsi, avec les architectures actuelles, la *« lecture d'images » s'aligne plus étroitement avec les capacités de base d'un LLM (interpréter les entrées et produire du texte)*, tandis que la *« génération d'images » exige une architecture qui modélise les pixels*. En effet, les premiers LLM multimodaux s'appuyaient sur des modules séparés pour produire des images. Ce n'est que récemment (par exemple, ANOLE (Source: bohrium.dp.tech) que des efforts open-source ont intégré le décodeur d'images dans le même modèle, illustrant comment les tâches de génération continuent de stimuler l'innovation architecturale.

- **Données d'Entraînement :** Les tâches de lecture peuvent exploiter de grands ensembles de données étiquetées (ImageNet, COCO avec légendes étiquetées, ensembles VQA). Les modèles linguistiques peuvent même amorcer des ensembles de données d'images par légendage (auto-instruction). Une étude note la construction d'un ensemble de données de 100 000 à 1,2 million de « légendes » en sollicitant GPT-4V sur des images (Source: academic.oup.com). Les tâches génératives utilisent souvent de *vastes collections d'images non étiquetées* (LAION, images web) ainsi que du texte apparié (par exemple, des légendes) pour superviser le conditionnement textuel. L'échelle est immense : les modèles de diffusion s'entraînent sur des milliards d'images. En bref, la compréhension d'images utilise des *étiquettes organisées* ou des paires image-texte ; la génération utilise des *images brutes* à très grande échelle. Les exigences différentes en matière de données reflètent leurs objectifs distincts.

- **Performance et Maturité :** Comme le montrent les Tableaux 1 et 2, les tâches de lecture disposent de **références objectives** où la précision est très élevée et les améliorations saturent (la précision d'ImageNet stagne, les modèles auto-supervisés égalent les modèles supervisés). Les tâches génératives ont des **mesures de qualité ouvertes**, mais les progrès ont été explosifs : l'état de l'art produit désormais des résultats quasi-humains sur de nombreux critères subjectifs. Cependant, la génération est toujours confrontée à des problèmes de correction subtile, tandis que les tâches de lecture hallucinent rarement des détails. Les LLM multimodaux actuels atteignent une compétence à peu près égale à celle des systèmes spécialisés pour les tâches de lecture (surtout après affinage), mais aucun générateur d'images purement basé sur des LLM n'atteint un niveau de sortie expert sans composants supplémentaires.

- **Cas d'utilisation et Impact :** Dans les applications réelles, la *compréhension d'images* est essentielle (diagnostics médicaux, détection de dangers, OCR pour les documents). Les exigences ici sont la précision et la fiabilité ; les solutions actuelles (CNN + éventuellement explication par LLM) répondent à ces besoins. La *génération d'images*, quant à elle, est principalement utilisée pour l'assistance créative (images marketing, visualisation de concepts) ou la simulation (augmentation des données d'entraînement). Comme l'indiquent les enquêtes d'adoption, le déploiement de l'IA générative monte en flèche : environ 33 % des organisations utilisent l'IA générative et 40 % prévoient d'investir davantage (Source: www.edge-ai-vision.com), signalant une forte confiance de l'industrie dans les outils de génération. Parallèlement, les outils de vision traditionnels sont déjà des produits matures dans l'industrie (tâches de vision par ordinateur dans la fabrication, la surveillance, etc.). Le constat est que les **industries créatives adoptent la génération**, tandis que les **industries objectives s'appuient sur la lecture**.

Dans l'ensemble, les preuves suggèrent qu'à l'heure actuelle, les **LLM multimodaux sont effectivement comparativement plus performants pour « lire » des images que pour les « générer »**. Lorsqu'on lui pose des questions factuelles sur une image, un modèle de type GPT-4V peut souvent répondre correctement. Mais si on lui demande de créer une nouvelle image, le même modèle doit passer le relais à un moteur de diffusion ; il n'a pas de « vocabulaire » interne d'images en pixels. Les modèles génératifs spécialisés surpassent les simples pipelines de LLM pour la création d'images. Cela dit, les deux mondes convergent – la recherche sur les modèles unifiés (modèles à jetons latents, décodeurs intégrés) est active (Source: bohrium.dp.tech). Il est possible que les futurs LLM fassent les deux de manière transparente. Pour l'instant, cependant, nous constatons que chaque domaine a ses leaders et ses limites.

Études de Cas et Applications

Cas 1 – Véhicules Autonomes : Dans les voitures autonomes, la **compréhension d'images** est primordiale. La détection d'objets en temps réel (voitures, piétons) et la reconnaissance de voies doivent être extrêmement fiables et rapides. Les systèmes actuels utilisent des réseaux de vision optimisés (YOLO, ResNet, EfficientDet) avec des latences de l'ordre de la milliseconde sur du matériel spécialisé (Source: medium.com). Les LLM ne font pas partie de la boucle de perception critique pour la sécurité. Ils *pourraient* être utilisés pour des descriptions de plus haut niveau (« rapport de trafic »), mais les décisions fondamentales reposent sur des modèles de vision. L'**IA générative** a une utilisation directe limitée ici, bien que la simulation de scénarios de conduite (via la génération d'images synthétiques ou des mondes virtuels) devienne importante pour l'entraînement. NVIDIA, par exemple, utilise des GAN pour générer des scénarios de cas limites rares. Cela montre que la *lecture et la génération d'images* répondent à des besoins différents : les véhicules privilégient la compréhension en arrière-plan, tandis que les pipelines de contenu automatisés peuvent utiliser la génération pour l'augmentation de données.

Cas 2 – Contenu Créatif et Marketing : Dans la publicité, la génération d'images par l'IA a explosé. Les marques utilisent désormais couramment Midjourney ou DALL·E pour l'art conceptuel initial, les publications sur les réseaux sociaux, voire les maquettes de produits. La qualité est suffisamment élevée pour que la production entre souvent dans les campagnes avec des modifications minimales. Parallèlement, les tâches de compréhension comme l'identification de logos de marque dans les images sont également automatisées, mais celles-ci sont relativement résolues par les outils de vision par ordinateur traditionnels. Ici, la **capacité générative** est au premier plan : selon les rapports de l'industrie, 39 % des spécialistes du marketing américains utilisent désormais l'IA pour la création d'images (Source: quantumailabs.net). Cela reflète les tableaux ci-dessus : les modèles génératifs (coût par image faible, qualité élevée) sont intégrés rapidement, tandis que les modèles de lecture opèrent en coulisses pour l'analyse.

Cas 3 – Imagerie Médicale : La radiologie et la pathologie reposent fortement sur l'analyse d'images (lecture) : détection de tumeurs dans les scanners, classification des tissus. Les outils modernes (CNN, réseaux segmentés) ont atteint une sensibilité diagnostique rivalisant avec celle des experts humains dans certains cas. Les modèles génératifs d'IA apparaissent également ici : les réseaux antagonistes génératifs sont utilisés pour synthétiser des images médicales (IRM, échographie) afin d'augmenter les données d'entraînement rares ou d'anonymiser les scanners de patients. Des outils comme la diffusion générative ont été utilisés pour « halluciner » des présentations tumorales rares afin d'élargir les ensembles de données. Néanmoins, il s'agit de stades de recherche ; les cliniciens font davantage confiance à l'interprétation (lecture) qu'à la génération synthétique d'images, qui reste une ressource auxiliaire.

Ces exemples illustrent que les **utilisations réelles de la lecture par rapport à la génération diffèrent**. Les modèles de lecture font partie intégrante des systèmes d'IA fondamentaux (sécurité, analyse), tandis que les modèles génératifs améliorent actuellement la créativité et la simulation de données. Les deux contribuent, mais de différentes manières.

Implications et Orientations Futures

Perspectives Technologiques : L'écart entre la lecture et la génération pourrait se réduire. Des recherches comme ANOLE (Source: bohrium.dp.tech) suggèrent que des modèles véritablement unifiés sont possibles. Nous pourrions voir les LLM produire directement des jetons d'image (par exemple, un embedding discrétisé) à l'avenir. Certains travaux utilisent déjà les LLM pour coordonner plusieurs modules spécialisés (voir OpenLEAF (Source: bohrium.dp.tech), qui intercale la génération de texte et d'images). Les architectures de transformeurs sont étendues aux images dans des espaces latents. Il est plausible que « GPT-5 » ou des modèles similaires puissent nativement répondre *et* créer des images. Cependant, des défis subsistent pour faire évoluer de tels modèles et garantir qu'ils ne perdent pas leur fidélité perceptuelle.

Métriques et Alignement : Pour les tâches de vision, l'évaluation robuste continue de s'améliorer. Par exemple, des **métriques multimodales** (comme l'utilisation de GPT pour juger les légendes d'images (Source: academic.oup.com) émergent. L'évaluation de l'alignement génératif (l'image reflète-t-elle fidèlement l'invite nuancée) nécessite des tests spécialisés. La communauté explore des métriques telles que les perturbations d'invite, les invites contradictoires et la notation avec intervention humaine (Source: huggingface.co) (Source: huggingface.co).

Éthique et Société : La lecture et la génération soulèvent des questions critiques. Les systèmes de lecture peuvent mal étiqueter ou véhiculer des biais sociaux (controverses sur la reconnaissance faciale). Les modèles génératifs risquent les *deepfakes* et la désinformation. Les mêmes architectures favorisant la créativité peuvent être utilisées à mauvais escient. En effet, des études mettent en évidence des cas d'hallucination (désinformation) et de biais (comme l'erreur des « Pères Fondateurs » de Gemini (Source: www.edge-ai-vision.com). Comme le note [22], l'IA en 2024 est au « sommet des attentes exagérées » (Source: www.edge-ai-vision.com). Un développement et une évaluation responsables (tests de résistance, tests d'équité) sont essentiels dans les deux domaines.

Applications Futures : Les applications futures potentielles combinent les deux compétences. Par exemple, un *assistant interactif basé sur l'image* pourrait inspecter votre pièce via une caméra (lecture), puis modifier une photo pour montrer une rénovation proposée (génération). OU, dans la créativité, un LLM pourrait critiquer ou affiner itérativement des images générées en les « lisant ». Nous voyons déjà les premiers directeurs artistiques être remplacés par des paires d'IA. Dans l'éducation et la recherche, les tuteurs IA pourraient interpréter les dessins des étudiants et générer de nouveaux diagrammes à la demande.

Défis de Recherche : Deux grands défis émergent : Premièrement, l'**évolutivité et les données**. La compréhension d'images peut nécessiter une annotation plus nuancée (par exemple, géométrie 3D, physique dans les images), ce qui met à rude épreuve les ensembles de données actuels. Les modèles génératifs ont besoin de plus de contrôle et de diversité (par exemple, générer des vidéos haute résolution, des scènes 3D). Deuxièmement, la **cohérence avec le langage**. Les LLM multimodaux doivent intégrer profondément la logique textuelle à la vision : par exemple, comprendre des diagrammes ou combiner de longs documents avec des images. Des benchmarks comme PCA-Bench (Source: huggingface.co) et les tests de contexte long (MileBench) soulignent que les modèles ont encore du mal avec le raisonnement complexe image+texte sur plusieurs étapes. Combler cet écart impliquera probablement une meilleure mémoire et des innovations architecturales.

Conclusion

En conclusion, les **systèmes multimodaux basés sur les LLM démontrent aujourd'hui une forte capacité à *lire* des images**, tirant parti d'un pré-entraînement à grande échelle pour classer, légender et raisonner sur des entrées visuelles avec une précision proche de l'état de l'art. Ils ont apporté un nouveau niveau de flexibilité aux tâches de vision, permettant des questions-réponses ouvertes et des explications que les systèmes de vision par ordinateur traditionnels ne pouvaient pas offrir. Parallèlement, la **génération d'images à la pointe de la technologie** est actuellement dominée par des modèles spécialisés (diffusion, transformeurs) qui produisent des

images photoréalistes à partir d'invites textuelles, souvent avec une qualité prête pour l'industrie. Ces deux domaines répondent à des besoins différents : les tâches de lecture privilégient l'exactitude factuelle et l'interprétation, tandis que les tâches génératives privilégient la créativité et la qualité perceptuelle. Les deux ont fait des progrès remarquables, mais ils restent largement des écosystèmes distincts.

Les preuves de ce rapport indiquent qu'en 2025, les *LLM multimodaux sont relativement meilleurs pour interpréter les images que pour les générer*, tandis que les modèles génératifs dédiés mènent la course à la synthèse d'images. Cependant, les frontières s'estompent. Les modèles émergents (par exemple, les réseaux unifiés basés sur des jetons) et les pipelines intelligents permettent désormais aux LLM de participer à la génération (et vice versa). L'avenir verra probablement des systèmes plus intégrés capables de passer fluidement de la vision à la création. Pour l'instant, les praticiens devraient choisir l'outil adapté à la tâche : utiliser les modèles vision-langage pour l'analyse et la compréhension, utiliser les modèles génératifs d'images pour la production créative, ou combiner les deux dans des architectures hybrides pour tirer le meilleur parti des deux mondes.

Toutes les affirmations et données de ce rapport sont étayées par des recherches actuelles, des benchmarks et des analyses d'experts (Source: research.aimultiple.com) (Source: www.researchgate.net) (Source: bohrium.dp.tech) (Source: www.edge-ai-vision.com). Nous encourageons les lecteurs à consulter les sources citées pour plus de détails.

Références

- Dilmegani, C., & Şipi, N. « Vision Language Models Compared to Image Recognition » (2025) – Benchmark de modèles vision-langage (CNN vs GPT-4V) (Source: [research.aimultiple.com](https://research.aimultiple.com/vision-language-models/#:~:text=Traditional%20image%20recognition%20models%2C%20such,This)) (Source: [research.aimultiple.com](https://research.aimultiple.com/vision-language-models/#:~:text=For%20vision%20language%20models%2C%20including,60%20mAP)). - OpenAI, *CLIP: Connecting Text and Images* (2021) – Modèle image-texte conjoint, 400M données (Source: [towardsai.net](https://towardsai.net/p/l/notes-on-clip-connecting-text-and-images#:~:text=The%20authors%20propose%20a%20pre,specific%20training)). - Jamal *et al.*, « Perception and evaluation of text-to-image generative models... » (2024) – Comparaison DALL·E, Imagen, SD, métriques FID/SSIM (Source: [www.researchgate.net](https://www.researchgate.net/publication/385290574_Perception_and_evaluation_of_text-to-image_generative_AI_models_a_comparative_study_of_DALL-E_Google_Imagen_GROK_and_Stable_Diffusion#:~:text=and%20similarity%20to%20real%20images,exhibited%20the%20most%20promising%20performance)) (Source: [www.researchgate.net](https://www.researchgate.net/publication/385290574_Perception_and_evaluation_of_text-to-image_generative_AI_models_a_comparative_study_of_DALL-E_Google_Imagen_GROK_and_Stable_Diffusion#:~:text=mathematical%20evaluation%2C%20DALL,perceived%20realism%20compared%20to%20Stable)). - Borade, K., *From Pixels to Prompts* (2025) – Enquête sur ML vs LLM pour les tâches de vision (Source: [medium.com](https://medium.com/@kanchanborade/from-pixels-to-prompts-choosing-between-ml-llm-for-image-tasks-46aedab918d5#:~:text=Modern%20LLMs%20like%20OpenAI%E2%80%99s%20GPT,support%20vision%20%2B%20language%20tasks)) (Source: [medium.com](https://medium.com/@kanchanborade/from-pixels-to-prompts-choosing-between%20ml-llm-for-image-tasks-46aedab918d5#:~:text=Use%20ML%20when%3A)). - N-shot, *Text-to-Image Generation: State-of-the-Art* (Décembre 2023) – Aperçu historique de la révolution de la diffusion (Source: [n-shot.com](https://n-shot.com/text-to-image-generation-from-evaluation-metrics-to-state-of-the-art-models/#:~:text=%2A%202018,image%20quality%20and%20prompt%20adherence)). - Edge AI & Vision Alliance (Tenyks), *Evaluating GenAI Vision Models* (2025) – Tendances de l'industrie, notes de politique (Source: [www.edge-ai-vision.com](https://www.edge-ai-vision.com/2025/01/dall-e-vs-gemini-vs-stability-genai-evaluations/#:~:text=%E2%80%8DA%20recent%20survey%20conducted%20by,2)) (Source: [www.edge-ai-vision.com](https://www.edge-ai-vision.com/2025/01/dall-e-vs-gemini-vs-stability-genai-evaluations/#:~:text=%E2%80%8DEven%20Google%E2%80%99s%20new%20AI%20image,5)). - Yu *et al.*, *ANOLE: Autoregressive Large Multi-modal Model* (2024) – Génération image-texte tout-en-un (Source: [bohrium.dp.tech](https://bohrium.dp.tech/paper/arxiv/2407.06135?s=pr#:~:text=reliance%20on%20additional%20diffusion%20models,experimentation%20for%20researchers%20at%20different)). - Yin *et al.*, *A Survey on Multimodal Large Language Models* (2024, NSR) – Enquête complète sur les MLLM (Source: [academic.oup.com](https://academic.oup.com/nsr/article/11/12/nwae403/7896414#:~:text=Since%20the%20benchmark%20evaluation%20is,spite%20of%20different%20response%20styles)). - Heusel *et al.*, « GANs trained by a two time-scale update rule converge to a Nash equilibrium » (2017) – Métrique FID (contexte).

Toutes les citations utilisent le format [source†L..] tel qu'énuméré ci-dessus.

Étiquettes: ia-multimodale, generation-images, comprehension-images, vision-par-ordinateur, texte-vers-image, modeles-diffusion, llm-vision, gpt-4v

---