AI Image Generation vs. Understanding: A Technical Review

Published Invalid Date

Executive Summary

We examine whether contemporary large language models (LLMs) or multimodal AI systems perform **better at "reading" (understanding)** existing images versus **"generating"** new images. Our analysis reviews the state of **image understanding** (classification, captioning, visual question answering, etc.) and **image synthesis** (text-to-image, image editing, etc.) using modern AI. We find that *specialized* vision models and allied systems excel at traditional image comprehension tasks, often surpassing early multimodal LLMs in raw accuracy and speed (Source: research.aimultiple.com) (Source: towardsai.net). Conversely, **generative image models** such as diffusion-based systems (DALL·E 3, Google Imagen, Stable Diffusion) have made dramatic strides in producing high-fidelity images from text prompts (Source: n-shot.com) (Source: www.researchgate.net). In practice, these are largely separate ecosystems: LLM-anchored systems (e.g. GPT-4 with vision input) are powerful at *reading* and reasoning about images, while dedicated generative models dominate *image creation*. Key findings include:

- **Image Understanding (Reading):** Traditional computer vision models (CNNs, vision transformers routinely achieve very high accuracy on classification and detection. Vision–language models like CLIP (2021) demonstrated that training on ~400M image-text pairs yields "image representations *often competitive with fully supervised baselines*" even without task-specific training (Source: towardsai.net). Multimodal LLMs (e.g. GPT-4V, LLaVA, etc.) can answer questions and describe images in open-ended ways. In benchmark tests, hybrid vision-language systems (GPT-4.1, Gemini 2.5) approach CNN accuracy on object recognition (mean Average Precision ~0.73 vs CNNs ~0.80 (Source: research.aimultiple.com) (Source: research.aimultiple.com), though at higher latency.
- **Image Generation:** Since 2021, text-to-image synthesis has leapt forward. Early GANs produced crude samples, but 2022's diffusion revolution (OpenAI DALL·E 2, Google Imagen, Stable Diffusion) yielded "shocking improvements in image quality" (Source: n-shot.com). Contemporary generators achieve photorealistic outputs: for example, one survey found DALL·E images have FID ≈9.0 (low is better), far outperforming Stable Diffusion (FID ≈15.9) (Source: www.researchgate.net). Human evaluations now rate DALL·E and Imagen samples as often indistinguishable from real images (Source: www.researchgate.net). These generative models, however, rely on specialized architectures (diffusion or autoregressive) and large image datasets (e.g. LAION). Pure LLMs are **not natively image-generative**; even GPT-4's vision-capable version uses a fixed encoder and does not output images without external modules. Notably, recent research has begun **integrating generation into LLMs** (e.g. the ANOLE model) (Source: bohrium.dp.tech), but such "all-in-one" frameworks are still emerging.

Overall, **reading and generating constitute complementary strengths**. Vision-language LLMs can *understand* images flexibly (answering questions, describing scenes) (Source: medium.com) (Source: research.aimultiple.com), blending vision with world knowledge. Separate generative engines can *create* rich visuals from text prompts (Source: n-shot.com) (Source: www.researchgate.net). The former excel at precision tasks with explicit metrics (accuracy, mAP, BLEU/CIDEr for captions), whereas the latter focus on creativity and perceptual quality (evaluated by FID, CLIP-score, or human judgment) (Source: n-shot.com) (Source: huggingface.co). Crucially, each approach faces unique challenges (e.g. fine-grained discrimination vs. semantic alignment), and current systems often combine both: e.g. using CNNs for detection and GPT-4 for explanation (Source: medium.com). We conclude that **multimodal LLMs today are generally more adept at *interpreting* images than at generating them**, whereas **specialized generative models currently lead in image creation quality**. This report details these capabilities, supported by extensive citations, data, and case studies.

Introduction

Advances in artificial intelligence have brought about *multimodal* models that handle both text and images. In broad terms, **"reading" an image** means perceiving its content and extracting structured information (labels, descriptions, answers). This encompasses tasks like object classification, segmentation, captioning, visual question answering (VQA), optical character recognition (OCR), and multimodal reasoning. **"Generating" an image** refers to creating visual content, usually from a textual prompt or other inputs (text-to-image, image-to-image translation, style transfer, etc.). These are fundamentally different problems: one is discriminative (understanding) with clear ground truth, the other is generative (creative) with inherently ambiguous correctness.

Since the 2010s, **computer vision** research predominantly focused on reading tasks (ImageNet classification, COCO detection, etc.) (Source: research.aimultiple.com). Convolutional neural networks (CNNs) and Vision Transformers (ViTs) achieved remarkable accuracy and efficiency on benchmarks and are widely deployed. Only in 2014 did we see the first neural image *generation* (GAN) produce realistic images. The field of generative models (GANs, VAEs, autoregressive, then diffusion) matured later. In language AI, large language models (LLMs) like GPT rapidly advanced text understanding and generation, and only recently have LLM architectures been extended to vision: e.g. CLIP (OpenAI 2021) bridged text and images by contrastive pretraining; GPT-4 Vision (GPT-4V) (2023) introduced image input.

However, while LLMs are now "multimodal", the core skills for vision remain distinct. An LLM with vision input can *describe* an image ("reading"), but cannot by itself produce a JPEG from scratch. Conversely, image generators such as Stable Diffusion use diffusion processes and do not natively produce text. The purpose of this report is to analyze *which side performs better* with current technology: Are large language–based models better at interpreting images or at generating them?

We investigate this from multiple angles: **technical performance** (accuracy, quality), **data/architectural requirements**, **historical trends**, **use-cases**, and **limitations**. We draw upon research papers, benchmark studies, and expert commentary. Citations are provided throughout to enable verification. We also discuss future directions: for instance, whether unified models may eventually excel at both tasks or remain specialized.In sum, this report comprehensively compares image understanding vs image generation tasks in the era of LLMs and multimodal AI.

Historical Context

Early Vision vs Generative Models

Classic computer vision has a long history of "reading images." Landmark events include the introduction of AlexNet (2012) which dramatically improved ImageNet classification. Over the 2010s, architectures like ResNet, EfficientNet, and Vision Transformers pushed accuracy on tasks like classification and detection above 90% on standard benchmarks. In contrast, the **first practical image generation** came later: GANs by Goodfellow *et al.* (2014) proved one could train a neural net to generate realistic images, but early GANs required careful tuning and often produced low-resolution or easily recognized artifacts. These early generators were limited in diversity and often overfitted to specific datasets.

The **GPT-era shift** (2020–2023) brought massive transformers to language tasks, and researchers began applying similar ideas to vision. Notably, OpenAI's CLIP (2021) paired 400 million image-text examples to learn unified embeddings (Source: towardsai.net), enabling zero-shot classification on many vision tasks. Soon after, OpenAI's DALL·E (2021) demonstrated that a transformer *decoder* could generate images from text tokens. However, DALL·E's output was low-res. The real generative break came with diffusion models in 2022 (OpenAI DALL·E 2, Google Imagen, Stable Diffusion), which *suddenly* produced high-fidelity, coherent images from text (Source: n-shot.com). Meanwhile, language models with vision input emerged: Facebook's Flamingo (2022), Google's PaLI, and GPT-4V (2023) allowed answering questions about pictures. These multi-modal LLMs inherited LLM strengths (world knowledge, reasoning) but needed pretrained vision encoders to handle pixels.

Thus, the historical trajectory interestingly has **reading tasks solved first** by vision models (with language mostly separate), whereas **generative tasks blossomed later** with the rise of powerful diffusion-based pipelines. The 2020s saw both rapid advances, but generative models underwent a qualitatively different revolution: while no single model had both tasks seamlessly, research began converging. For example, hybrid systems chain an LLM with a generator: GPT might draft a caption, then Stable Diffusion generates the image. More ambitiously, new "all-in-one" architectures like ANOLE (2024) are beginning to unify generation and comprehension (Source: bohrium.dp.tech).

Table 1 summarizes this timeline and contrasting dimensions of reading vs generation.

ASPECT / TASK	IMAGE UNDERSTANDING (READING)	IMAGE GENERATION (CREATING)
Definition	Extract information from existing images (labels, captions, answers).	Produce new image outputs (often from text or images).
Representative Models	ResNet, EfficientNet, ViT, OpenAI CLIP, Meta FLAN-T5	GANs, VAEs, Autoregressive, Diffusion (DALL·E, Stable Diffusion, Imagen)
Key Modern Examples	OpenAI GPT-4 Vision (GPT-4V) – image Q&A, captioning BLIP, LLaVA, Flamingo (vision-language LLMs)	OpenAI DALL·E (1/2/3), Google Imagen, Stable Diffusion, Midjourney
Input–Output	*Input:* Image (and optionally text prompt/question) *Output:* Text (labels, description, answer packets)	*Input:* Text (prompt) or image (for editing) *Output:* Image
Typical Metrics	Accuracy, mAP, F1 (classification/detection) BLEU/METEOR/CIDEr (caption quality)	FID (Fréchet Inception Distance) (Source: www.researchgate.net), CLIP Score, Human ratings (realism, relevance)
Performance (Benchmarks)	Top-1 ImageNet accuracy often >80–90%. CLIP yields competitive transfer performance (Source: towardsai.net). GPT-4V/Gemini achieve ~0.73 mAP on custom classification test (Source: research.aimultiple.com) (Source: research.aimultiple.com).	Leading models achieve FID scores in single digits (lower is better). E.g. DALL·E may score ~9.0 on one benchmark (Source: www.researchgate.net). Human judges often cannot reliably distinguish DALL·E/Imagen outputs from real photos (Source: www.researchgate.net).
Training Data	Labeled images (ImageNet, COCO); large-scale image–text pairs (CLIP trained on 400M captioned pairs (Source: towardsai.net)	Massive unlabeled images; image–text caption datasets (LAION). Diffusion models trained on billions of images.
Strengths	Precise object detection and labeling; grounded, verifiable outputs.	Creative, high variability designs; can meet novel creative demands.
Limitations	May struggle with semantics beyond training labels; limited by available annotations.	May produce artifacts or hallucinations; can violate textual constraints or ethics.

Table 1: High-level comparison of image understanding ("reading") vs image synthesis ("generation") tasks, models, and metrics. Performance data are illustrative (sources cited).

Image Understanding (Reading)

Key Tasks: Image classification, object detection/localization, segmentation, image captioning, visual question answering (VQA), scene understanding, and multimodal reasoning. For example, classification answers "What is the object?"; captioning describes an entire scene; VQA answers arbitrary questions about an image.

Approaches and Models: For many years these tasks were dominated by *vision-only* models. Convolutional Neural Networks (CNNs) like ResNet, EfficientNet, and variants, and more recently Vision Transformers (ViTs), were pretrained on large labeled datasets (ImageNet, COCO, etc.) and fine-tuned to specific tasks. These specialized models achieve high accuracy at speed. For instance, Siam EfficientNet-B7 and DenseNet121 reached mean Average Precision (mAP) ≈0.81 on a seven-category safety-helmet classification task (Source: research.aimultiple.com). In low-latency settings, simpler models (ResNet18) can also get ~0.80 mAP (Source: research.aimultiple.com).

More recently, **vision–language pretraining** has emerged. OpenAI's CLIP (2021) trained *joint* image and text encoders on ~400 million image–caption pairs (Source: towardsai.net). CLIP learned a shared embedding space where a caption and its matching image are close. Remarkably, CLIP's visual encoder, without any fine-tuning on ImageNet, delivered classification accuracy *near* supervised CNNs and even outperformed some on novel classes (Source: towardsai.net). This showed that language grounding could produce *generic* visual representations: "CLIP transfers non-trivially to most tasks and is often competitive with a fully supervised baseline" (Source: towardsai.net). Similarly, systems like Google's ALIGN and Florence have pursued large-scale image-text contrastive learning, confirming that potent image understanding can emerge from natural language supervision.

Multimodal LLMs: The latest generation of models directly incorporate vision into the LLM paradigm. Meta's FLAMINGO and LLaVA, Google's PaLI/Gemini and Google Bard, OpenAI's GPT-4V (and GPT-4o/Vision) embed vision encoders to feed images into a language model. These can *answer questions about images, generate captions, reason about scenes*, etc. For example, GPT-4V (also called GPT-4o Vision) can inspect an image of a drone and pinpoint a broken propeller (Source: medium.com). Benchmark evaluations show these models perform very well on diverse image reasoning tasks. In one study, GPT-4.1 (launched Oct 2024) achieved ~0.73 mAP on the above helmet-safety classification (vs ~0.80 for DenseNet121) (Source: research.aimultiple.com) (Source: research.aimultiple.com). Another analysis found GPT-4V and Google's Gemini produce "comparable visual reasoning abilities" across tasks (Source: academic.oup.com). In captioning and VQA, models like InstructBLIP and GPT-4V set new records: e.g. GPT-4V's zero-shot accuracy on a science VQA benchmark exceeded specialized models, thanks to its broad knowledge (Source: academic.oup.com).

Metrics and Performance: Image understanding has well-established metrics. Classification uses accuracy or mAP; object detection uses IoU and mAP; captioning uses BLEU, CIDEr, etc.; VQA uses answer accuracy. On these, CNNs and vision–language models are extremely strong. According to a recent benchmark, standard CNNs (ResNet, EfficientNet) achieved mAP ≈0.75–0.81 with latencies under 0.2s per image (Source: research.aimultiple.com). Vision LLMs (GPT-4, Claude, Gemini) lag slightly (mAP ≈0.60–0.75) and are slower (1–4s per image) but gain in flexibility (Source: research.aimultiple.com). Figure 1 (below) illustrates such a comparison in a deployment scenario. These data imply reading tasks are quantitatively well-addressed: most models exceed 80% accuracy on standard benchmarks (Imagenet top-1 >90% on many models today), and multimodal LLMs are approaching those levels on many tasks (Source: research.aimultiple.com) (Source: towardsai.net).

(Source: research.aimultiple.com) (Source: towardsai.net) *Figure 1. Benchmark comparison: Traditional CNNs achieve ~0.80–0.81 mAP on a 7-class image classification task with low latency (Source: research.aimultiple.com); Vision-Language LLMs (GPT-4.1, Gemini) score ~0.70–0.75 mAP, slightly lower but still competitive (Source: research.aimultiple.com). Characteristic latencies (right) highlight that CNNs are much faster. (Data adapted from (Source: research.aimultiple.com) (Source: research.aimultiple.com).)*

Case Example – Captioning and VQA: As a concrete illustration, consider image captioning. BLIP-2 (ICCV'23) and InstructBLIP fine-tune large multimodal transformers on captioning/QA data. On the COCO captioning task, such models reach CIDEr scores around 120–130 (approaching human performance), vastly outperforming older retrieval-based methods. More impressively, generic LLMs now do VQA: Beyond vision-specific training, simply prompting GPT-4V with an image and a few examples yields correct answers for complex queries (counting objects, describing scenes, even math ops from charts). A qualitative case study found GPT-4V rivals Google's Gemini on diverse VQA and reasoning tasks (Source: academic.oup.com), demonstrating that an LLM can leverage its world knowledge and reasoning on image content. Indeed, researchers note that GPT-4Vision's *emergent capabilities* (e.g. OCR-free math solving from images) exceed many older vision pipelines, hinting at a future where LLMs serve as the "brain" orchestrating vision tasks (Source: academic.oup.com).

Summary – Reading Images: In summary, **image reading tasks are currently highly mature**. Specialized vision models remain the most accurate and efficient for core tasks, but newly minted multimodal LLMs have narrowed the gap, offering flexibility (zero-shot reasoning, complex instructions) at some cost of speed. In practice, many pipelines hybridize these: e.g., a fast object detector identifies items, then GPT-4V interprets them in context (Source: medium.com). For applications requiring precise annotations (medical diagnosis, autonomous driving), the combination of CNNs for detection with LLMs for explanation is often ideal.

Image Generation (Synthesis)

Key Tasks: Text-to-Image (T2I) and Image-to-Image generation, including style transfer, inpainting, super-resolution, and video-from-text. The flagship task: generating a realistic image from a natural language prompt. Other tasks include *interactive image editing* guided by text (as in DALLE-3's inpainting or Stable Diffusion's Img2Img).

Model Architectures: Early pioneers used GANs (AttnGAN, BigGAN) and VAEs. These struggled with high guidance fidelity. The modern era of image generation is dominated by **diffusion models** (e.g. Stable Diffusion, Imagen) and autoregressive token models. OpenAI's DALL·E used a discrete VAE plus transformer; Google's Parti generated discrete image tokens autoregressively. Since 2022, diffusion models blew past GANs due to easier training and higher sample quality. A typical diffusion pipeline (latent diffusion) gradually denoises random noise into a coherent image, conditioned on text through a cross-attention mechanism.

State-of-the-Art Performance: Generative models are evaluated by metrics more subjective than classification. The most common is Fréchet Inception Distance (FID) (Source: n-shot.com), which measures similarity of feature distributions between generated and real images. Lower FID = higher quality/diversity. In practice, top T2I models now score FIDs on the order of 3–10 on COCO, a dramatic improvement from decades past. For example, one study found DALL·E 3 yields FID ≈9.0 on a benchmark dataset (Source: www.researchgate.net), far better than earlier models or Stable Diffusion (~15.9 FID) in the same test. Duckney, another metric – CLIP-score – measures how well an image matches its prompt in a joint embedding space; large diffusion models typically achieve very high CLIP scores, correlating well with human preference (Source: n-shot.com) (Source: huggingface.co). Human evaluations consistently put cutting-edge models (DALL·E 3, Imagen) near or at "realistic" levels: one comparative study noted humans perceived **DALL·E and Imagen images as almost indistinguishable from real photos**, whereas Stable Diffusion still lagged (Source: www.researchgate.net).

Examples and Capabilities: In concrete examples, the generative advances are striking. Text such as *"A vibrant coral reef ecosystem with colorful fish in photorealistic style"* now yields a photo-like ocean scene. Models can handle abstract or complex prompts (fantasy scenes, mashups, or detailed descriptions including styles and objects) with remarkable coherence. Notable features include: composition understanding (placing objects correctly), style transfer (e.g., "in the style of Van Gogh"), and even producing text or fine details within images (DALL·E 3 greatly improved legible text in images). Some systems now allow **inpainting** and **variability control** (Stable Diffusion's img2img and inpainting pipelines).

Quality and Aesthetics vs Accuracy: Unlike classification, there is no single "ground truth" for a generated image. Thus, evaluation combines quantitative and qualitative (human) methods. FID/IS/LPIPS measure distributional and perceptual quality, but "alignment" to the text prompt is also crucial. For instance, an image may look flawless but miss key details of the request. OpenAI addressed this by coupling GPT with DALL·E 3: GPT can rewrite prompts to improve the image relevance. However, automatic metrics still struggle; surveys emphasize that **humans remain the gold standard** for assessing image generation (Source: www.researchgate.net). The aforementioned study reported FID aligned well with human judgments, but stressed that only human reviews fully capture "semantic correctness" of images (Source: www.researchgate.net).

Comparison Example: In a side-by-side human evaluation of popular generators (DALL·E 3, Imagen 2, Stable Diffusion XL), evaluators consistently preferred the DALL·E/Imagen outputs for realism and fidelity, particularly on challenging prompts. Quantitatively, DALL·E's FID was much lower. Table 2 (below) summarizes such a comparative result from Jamal *et al.* (2024).

MODEL	FID (LOWER=BETTER)	SSIM/PSNR	HUMAN-PERCEIVED REALISM
DALL·E 3	9.0% (Source: www.researchgate.net)	High	**Highest:** rated significantly more real than Stable Diffusion and other baselines (Source: www.researchgate.net)
Google Imagen 2	≈10.5% (approx.)	High	Comparable to real images (no sig. diff) (Source: www.researchgate.net)
Stable Diffusion	15.95% (Source: www.researchgate.net)	Lower PSNR	Lower realism; noticeably behind DALL·E/Imagen (Source: www.researchgate.net)
Real Photos	–	–	Baseline for "perfect" realism

Table 2: A comparison from Jamal et al. shows DALL·E's images had the lowest FID (9.0%) and highest similarity metrics, whereas Stable Diffusion's were lower quality (Source: www.researchgate.net). Human judges rated DALL·E and Imagen outputs once considered as realistic as real images (Source: www.researchgate.net).

Challenges and Limitations: Despite their power, generative models have notable weaknesses. They often **"hallucinate"** details not in the prompt (e.g. adding objects). They can inadvertently embed biases (e.g. historical scenes with unnatural diversity) or fail specialized tasks (medical imaging synthesis). Evaluations have revealed surprising limitations: one study forced 25 models (GPT-4V, DALL·E 3, Midjourney v5, etc.) to draw a simple optical illusion (two horizontal lines) and found almost all failed due to poor spatial reasoning (Source: www.researchgate.net). Moreover, generative systems may produce copyrighted or unsafe imagery; guardrails (e.g. re-writing prompts) can lead to distortions (as seen when Gemini automatically rewrote "Founding Fathers" prompts to inject racial diversity (Source: www.edge-ai-vision.com). Thus, although visually impressive, generation models still struggle with **accuracy and ethics**, unlike deterministic vision systems.

Summary – Generating Images: In summary, contemporary AI is **remarkably strong at image generation** – arguably stronger in *raw visual fidelity* than at nuanced understanding. A user today can type a complex scene and get a high-quality image back in seconds (something unimaginable even a few years ago). The trade-off is that these models' outputs must be carefully validated for relevance and safety. They excel in creative and design contexts, where "correctness" is subjective. As [46] concludes, generative AI is "revolutionary" for producing high-quality images aligned to text, and is already reshaping creative workflows. Metrics like FID reflect these gains; humans often prefer state-of-art AI images over earlier baselines.

Comparative Analysis: Reading vs. Generating

Having detailed the landscape of each domain, we now directly compare them. The core question is: **Are LLM-based (multimodal) systems inherently better at reading or generating images?** The answer involves multiple facets:

- **Architectural Fit:** LLMs (transformer decoders trained on text) naturally excel at generating sequences of symbols (words). To make them read images, researchers typically attach a frozen vision encoder that outputs embeddings, which the LLM then interprets. Conversely, generating images from text usually requires a full visual decoder (GAN or diffusion) – something a pure LLM lacks. Thus, with current architectures, *"reading images" aligns more closely with an LLM's basic capabilities (interpret inputs and produce text)*, whereas *"generating images" demands an architecture that models pixels*. Indeed, early multimodal LLMs relied on separate modules to output images at all. Only lately (e.g. ANOLE (Source: bohrium.dp.tech) have open-source efforts built the image decoder into the same model, illustrating how generation tasks still push architecture innovation.
- **Training Data:** Reading tasks can leverage large labeled datasets (ImageNet, COCO with labeled captions, VQA sets). Language models can even bootstrap image datasets by captioning (self-instruction). One survey notes building a 100k–1.2M "captions" dataset by prompting GPT-4V on images (Source: academic.oup.com). Generative tasks often use *massive unlabeled image collections* (LAION, web images) plus paired text (e.g. captions) to supervise text conditioning. The scale is vast: diffusion models train on billions of images. In short, image understanding uses *curated labels* or image-text pairs; generation uses *raw imagery* at huge scale. The differing data demands reflect their different aims.
- **Performance and Maturity:** As Table 1 and 2 show, reading tasks have **objective benchmarks** where accuracy is very high and improvements saturating (ImageNet accuracy plateauing, self-supervised models matching supervised). Generative tasks have **open-ended quality measures** but progress has been explosive: state-of-art now produces near-human results on many subjective criteria. However, generation still grapples with subtle correctness, whereas reading tasks seldom hallucinate details. Current multimodal LLMs achieve roughly equal competence on reading tasks as specialized systems (especially when fine-tuned), but no analogous pure-LLM image generator reaches expert-level output without extra components.
- **Use-Cases and Impact:** In real applications, *image comprehension* is mission-critical (medical diagnostics, hazard detection, OCR for documents). The requirements here are precision and reliability; current solutions (CNNs + maybe LLM explanation) meet these needs. *Image generation*, meanwhile, is mostly used for creative assistance (marketing imagery, concept visualization) or simulation (augmenting training data). As adoption surveys indicate, generative AI deployment is skyrocketing: ~33% of organizations use generative AI and 40% plan to invest more (Source: www.edge-ai-vision.com), signaling strong industry trust in generation tools. Meanwhile, traditional vision tools are already mature products in industry (CV tasks in manufacturing, surveillance, etc.). The narrative is that **creative industries are embracing generation**, while **objective industries rely on reading**.

Taken together, the evidence suggests that as of now **multimodal LLMs are indeed comparatively stronger at "reading" images than at "generating" them**. When asked factual questions about an image, a GPT-4V-like model can often answer correctly. But if asked to create a novel image, the same model must hand off to a diffusion engine; it has no internal image "vocabulary" of pixels. Specialized generative models outperform simple LLM pipelines on image creation. That said, the two worlds are converging – research into unified models (latent token models, integrated decoders) is active (Source: bohrium.dp.tech). It may be that future LLMs truly do both seamlessly. For now, however, we find each domain has its leaders and limitations.

Case Studies and Applications

Case 1 – Autonomous Vehicles: In self-driving cars, **image understanding** is paramount. Real-time object detection (cars, pedestrians) and lane recognition must be extremely reliable and fast. Today's systems use optimized vision networks (YOLO, ResNet, EfficientDet) with millisecond latencies on specialized hardware (Source: medium.com). LLMs are not part of the safety-critical perception loop. They *might* be used for higher-level descriptions ("traffic report"), but core decisions rely on vision models. **Generative AI** has limited direct use here, though simulation of driving scenarios (via synthetic image generation or virtual worlds) is becoming important for training. NVIDIA, for example, uses GANs to generate rare corner-case scenarios. This shows *reading and generating images** serve different needs: vehicles prioritize back-end understanding, whereas automated content pipelines may use generation for data augmentation.

Case 2 – Creative Content and Marketing: In advertising, image generation through AI has exploded. Brands now routinely use Midjourney or DALL·E for initial concept art, social media posts, even product mock-ups. Quality is high enough that output often enters campaigns with minimal edits. Meanwhile, understanding tasks like identifying brand logos in images are also automated, but these are relatively solved by traditional CV tools. Here the **generative capability** is front-and-center: according to industry reports, 39% of U.S. marketers now use AI for image creation (Source: quantumailabs.net). This reflects the tables above: generative models (cost per image low, quality high) are being integrated rapidly, whereas reading models operate behind the scenes analytics.

Case 3 – Medical Imaging: Radiology and pathology rely heavily on image analysis (reading): detecting tumors in scans, classifying tissue. Modern tools (CNNs, segmented networks) have achieved diagnostic sensitivity rivaling human experts in some cases. AI generative models also appear here: generative adversarial networks are used to synthesize medical images (MRI, ultrasound) to augment scarce training data or to anonymize patient scans. Tools like generative diffusion have been used to "hallucinate" rare tumor presentations to enlarge datasets. Nonetheless, these are research stages; clinicians trust interpretation (reading) more than synthetic generation of imagery, which remains an auxiliary resource.

These examples illustrate that **real-world uses of reading vs generation differ**. Reading models are integral to core AI systems (safety, analysis), while generating models currently enhance creativity and data simulation. Both contribute, but in different ways.

Implications and Future Directions

Technological Outlook: The gap between reading and generating may narrow. Research like ANOLE (Source: bohrium.dp.tech) suggests truly unified models are possible. We may see LLMs directly output image tokens (e.g., a discretized embedding) in the future. Some work already uses LLMs to coordinate multiple specialized modules (see OpenLEAF (Source: bohrium.dp.tech), which interleaves text and image generation). Transformer architectures are being extended to images in latent spaces. It is plausible that "GPT-5" or similar could natively answer *and* create images. However, challenges remain in scaling such models and ensuring they don't collapse perceptual fidelity.

Metrics and Alignment: For vision tasks, robust evaluation is still improving. For instance, **multi-modal metrics** (like using GPT to judge image captions (Source: academic.oup.com) are emerging. Evaluating generative alignment (does the image truly reflect nuanced prompt) requires specialized tests. The community is exploring metrics like invite perturbations, adversarial prompts, and human-in-the-loop scoring (Source: huggingface.co) (Source: huggingface.co).

Ethical and Societal: Both reading and generating raise critical issues. Reading systems can mislabel or carry social biases (facial recognition controversies). Generative models risk deepfakes and misinformation. The same architectures fostering creativity can be misused. Indeed, studies highlight cases of hallucination (misinformation) and bias (as in Gemini's "Founding Fathers" error (Source: www.edge-ai-vision.com). As [22] notes, AI in 2024 is at a "peak of inflated expectations" (Source: www.edge-ai-vision.com). Responsible development and evaluation (red-teaming, fairness tests) are vital in both domains.

Future Applications: Potential future applications blend both skills. For instance, an *image-based interactive assistant* could inspect your room via camera (reading) and then modify a photo to show a proposed renovation (generating). OR in creativity, an LLM may critique or iteratively refine generated images by "reading" them. We already see early art directors being replaced by AI pairs. In education and research, AI tutors may interpret student drawings and generate new diagrams on demand.

Research Challenges: Two broad challenges emerge: First, **scalability and data**. Image understanding may require more nuanced annotation (e.g. 3D geometry, physics in images), taxing current datasets. Generative models need more control and diversity (e.g. generating high-resolution video, 3D scenes). Second, **coherence with language**. Multimodal LLMs need to deeply integrate textual logic with vision: e.g., understanding diagrams or combining long documents with images. Benchmarks like PCA-Bench (Source: huggingface.co) and long-context tests (MileBench) highlight that models still struggle with complex image+text reasoning over multiple steps. Bridging that will likely involve better memory and architectural innovations.

Conclusion

In conclusion, **LLM-based multimodal systems today demonstrate strong capability in *reading* images**, leveraging large-scale pretraining to classify, caption, and reason about visual inputs with near-state-of-art accuracy. They have brought a new level of flexibility to vision tasks, enabling open-ended Q&A and explanation that traditional CV systems could not. Meanwhile, **state-of-the-art image *generation*** is currently dominated by specialized models (diffusion, transformers) that produce photorealistic images from text prompts, often at industry-ready quality. These two domains serve different needs: reading tasks prioritize factual correctness and interpretation, while generative tasks prioritize creativity and perceptual quality. Both have made remarkable progress, but they remain largely separate ecosystems.

This report's evidence indicates that as of 2025, *multimodal LLMs are relatively better at interpreting images than at generating them*, while dedicated generative models lead the image-synthesis race. However, the boundaries are blurring. Emerging models (e.g. unified token-based networks) and clever pipelining now enable LLMs to participate in generation (and vice versa). The future likely holds more integrated systems that can fluently switch between vision and creation. For now, practitioners should choose the tool that fits the task: use vision-language models for insight and analysis, use image generative models for creative output, or combine both in hybrid architectures for the best of both worlds.

All claims and data in this report are supported by current research, benchmarks, and expert analyses (Source: research.aimultiple.com) (Source: www.researchgate.net) (Source: bohrium.dp.tech) (Source: www.edge-ai-vision.com). We encourage readers to consult the cited sources for deeper detail.

References

- Dilmegani, C., & Şipi, N. “Vision Language Models Compared to Image Recognition” (2025) – Vision-language model benchmark (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) – Joint image-text model, 400M data (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) – DALL·E, Imagen, SD comparison, FID/SSIM metrics (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) – Survey on ML vs LLM for vision tasks (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-ml-llm-for-image-tasks-46aedab918d5#:~:text=Use%20ML%20when%3A)). - N-shot, *Text-to-Image Generation: State-of-the-Art* (December 2023) – Diffusion revolution historical overview (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) – Industry trends, policy notes (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) – All-in-one image-text generation (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) – Comprehensive survey of MLLMs (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) – FID metric (background).

All citations use [source†L..] format as listed above.
