Survey Article on Recent Advancements in Text-to-Image and Video Generation: AI Systems, Platforms, and Methods

The problem statement arose with the need to understand image and video generation from text. There are multiple steps of doing this. The first step involves understanding image and video generation capabilities of different GPTs and how they operate, at a higher level. The next would involve creating a similar model from scratch but that is out of scope for now. AI has come a long way since the 1960s. Not only can it take a short description and turn it into realistic artwork, animation images, videos, or even a teaching aid but also connect language with visuals. This characteristic is currently being applied to everything from education to marketing, creative design, advertising and content creation. Some come built into chat apps, others are code-based, and a few are specialized design tools. But there is a major pressing challenge if left unchecked. And that is hallucination. Hallucinations often occur through misspelled text or made-up details. This survey first explores how leading AI tools handle text-to-image and text-to-video generation, how they perform, the common issues they face, how those are fixed, and what alternatives exist if someone weren't to rely on the tools.

AI-Powered Text-to-Image Synthesis in Prominent Language Models

Modern large language models or even small language models come with their own image generation features. They use diffusion models that interpret user-defined prompts and build detailed visuals. Diffusion models add Gaussian noise (forward process) and then learn to remove it (reverse process). In latent diffusion, a VAE encoder compresses images into a smaller latent space, a denoising U-Net operates in that latent space, and a VAE decoder reconstructs pixels from the latent. This is a high level understanding of diffusion. Its critical to understand diffusion as they form the foundation of image and video generation. And since they involve generation of images, audio, video; they come under generative Models. Now, the level of integration of these LLMs (many closed- and open-source) with the image generation models vary. Some are native, while others call external APIs. This choice affects both output quality and flexibility.

Perplexity AI's Methodology

Users can create images from text prompt using Perplexity AI without additional steps. What perplexity does under the hood is using image generation models from numerous providers like OpenAI, Google, etc to generate visuals part by part. At present the imge generation models in Perplexity include GPT Image 1 (OpenAI's advanced model), Dall-E by Open AI, Gemini 2.5 flash by Google, FLUX1 (From black forest labs), Seedreem 4.0 by Bytedance. It allows users to select image generation model via settings. In default cases, Perplexity automatically picks the best model based on the user's query. All of these information has been found from YouTube and from Perplexity itself. It uses a unique pixel representation. It often produces simple yet creative results and works best in case of less text provided for infographics. It has been observed that for infographic-like content, sometimes, it adjusts layouts or skips minor text details. Its chat-based process is similar to other AI tools, but there is an additional factor involved. Perplexity uses real-time search engine, which improves factual accuracy of the retrieved results. Recent updates have made image generation process faster and more detailed, allowing quick switches between realistic and artistic styles in seconds.

ChatGPT with DALL-E Fusion

ChatGPT by OpenAI, at its core, uses OpenAI's DALL-E, which uses transformer-based architecture to treat text and images as part of a single stream. Chatgpt takes user's prompts and gives infographic friendly prompts. This, in turn, allows the Dall-E models to generate visuals directly from prompts. For complex requests, the prompts are refined automatically, improving detail and accuracy compared to earlier versions. You can add text into images, though sometimes unexpected elements appear. DALL-E 3 is available through GPT-4. In this case, prompts can be adjusted collaboratively for better results. Its biggest strength lies in contextual understanding, making it ideal for educational visuals and storytelling.

ChatGPT has shown higher accuracy results and better performance as compared to Gemini and Perplexity in terms of generation of complex infographics and accuracy but takes more time as compared to both.

Google Gemini alongside Imagen

Google's Gemini uses Google's Deep Mind image generation model, Imagen, a diffusion-based model that creates realistic images from scratch. Imagen model blends the text with visuals seemlessly, by taking prompts from Google's Gemini and generating image from scratch. It has been observed to perform well for story-driven imagery, provided the prompts are comprehensible. At present, API is available for Imagen in languages like Python, that enables developers to build custom applications locally or in Google Cloud. Gemini also supports combination of text with audio, video for better outputs.

Grok AI

Grok AI, which has been built on xAI technology supports native image generation via xAI's image model and image-to-video with Imagine but it provides outputs in formats like Markdown, SVG, or HTML and guides users for image generation, a trait not seen in other models like ChatGPT and Gemini, unless explicitly queried through prompts. Imagine handles complex prompts and can turn a single input image into short, stylized videos; current flows typically process one image per prompt. Its integration with live data and quick responses keeps it competitive, even though some technical details remain under wraps.

Claude AI

Claude AI does not create images directly; it crafts detailed prompts or HTML/CSS and can analyze images. It can produce runnable code, but there is no supported "30% objective" execution metric.

It analyzes uploaded images or convert handwriting into digital text, completing part of the visual pipeline. Through integrations with platforms like Hugging Face, Claude can connect to tools such as Krea for image generation. This indirect approach works well for storytelling, where Claude finetunes description for other generators.

Unified Design Environments: Canva's AI Capabilities for Visual and Motion Content

Outside of chat-based models, tools like Canva use AI to simplify image and animation creation — especially for beginners. Canva's Magic Studio, which includes Magic Media, uses systems like DALL·E and Imagen to turn text into editable visuals. You simply type a description, pick a style (like watercolor or realistic), and insert it directly into your project. You can edit, re-style, or upscale the result afterward. This mixed approach uses public image libraries for ideas, but precise prompts might be needed to skip generic outputs. Canva's AI creation is hybrid, making images as starting points that people edit by hand, mixing AI with human tweaks for accuracy and personal touch.

Compared to tools focused on language models like ChatGPT's DALL-E, Canva shines in being easy for newbies, blending smoothly with templates and edit tools, and speeding up design tasks. It does great with after-creation changes, cutting need for other editors, and gets kudos for handling complex requests with less delay than chat-based AIs. While DALL-E might win in raw creativity for real-looking stuff, Canva's setup makes it better for real-world uses like ad materials.

Moving to moving content, Canva's AI builds animated clips through features like Create a Video Clip, powered by Google's Veo-3, turning text hints into eye-catching segments with ease.

Users can also experiment with HeyGen for animated characters or use effects like "AI Dance" and "AI Fight" from text or image inputs. This allows beginners to make quick video snippets for online content — combining AI automation with manual editing flexibility.

AI Video Generation in Large Language Models

Beyond still images, some large language models can now generate videos directly from text or images.

ChatGPT uses Sora for video generation, creating clips for upto a minute. though results depend heavily on prompt quality. Claude doesn't have native video generation, sticking to text and structured outputs. Perplexity gives video creation with sound for Pro users, making 8-second clips but not handling videos natively. Grok's Imagine lets you make videos from prompts and images, with sound, dealing with one image at a time for up to 15-second clips, including a "spicy" mode for adult content. Gemini uses Veo for video making, supporting sound and mixed inputs.

Among tools that make videos, Sora is top for super real sound mixing and personal touches like cameos, good for movie-like animations. Veo stands out with built-in audio, custom parts based on elements, and solid animation controls. Grok Imagine has sound and personalization, including adult stuff, with animation from pictures. Canva's Magic Studio, through Veo-3, offers sound, editable personal options, and animation tools.

Effectiveness Assessments and Prevalent Obstacles

The performance from least to best has been observed to be of the order: Gemini < Perplexity < ChatGPT. Although there is no such standardized publicly available peer-reviewed benchmarks that ranks Gemini, Perplexity and ChatGPT universally. More research is needed with higher text corpus and additional benchmarks.

User reviews show ChatGPT leading in creativity, usability, and accuracy, followed by Perplexity. Gemini performs well but can lag in speed. Canva consistently ranks high for ease of use and practical design capabilities. ChatGPT's DALL-E link provides refined, detailed results, while Perplexity prioritizes factual consistency, sometimes at the cost of visual richness. Gemini handles multimodal inputs well but takes longer to process.

Across all systems, hallucinations such as incorrect or missing details, misspelled text, or mismatched visuals remain a common issue. These errors mostly come from biased training data or limitations in pattern prediction.

Approaches to Alleviate Hallucinations

To minimize such issues, a few strategies help. Using diverse and balanced training data reduces bias. Writing clear prompts with specific rules (like "ensure correct spelling") guides the model better. Retrieval-augmented generation (RAG) improves factual accuracy by pulling verified data. OCR-based post-processing can detect and fix spelling errors in generated images. Fine-tuning models on specialized datasets and including human quality checks further improve reliability. Transparency about model limits and iterative prompt refinement also help maintain accuracy.

Comparative Table – Text to Image and/or Video Generation

System / Model	Modality	How you access it	Notable strengths	Common caveats	Text-in-image reliability	Notes
ChatGPT + DALL·E 3	Image	ChatGPT UI/API (Images), prompt rewrite under the hood	Strong instruction following; solid typography vs prior gens; easy iteration	Still occasional artifacts/unwanted elements	Better than older gens; not perfect	Paper describes caption-driven gains; Cookbook notes prompt rewriting. (cdn.openai.com)
Gemini + Imagen 3	Image	Gemini / AI Studio / APIs	High realism; good text blending; fast turnarounds	Access tiers vary; guardrails stricter	Reported strong; depends on prompt clarity	Imagen 3 tech report + arXiv. (Google Cloud Storage)
Perplexity (in-chat image gen)	Image	Perplexity (best on web; Pro plan)	One-shot image gen inside search/chat; fast; grounded prompts	Providers/details opaque; layout fidelity for dense infographics can vary	Mixed; better for light text	Official help confirms feature; vendors not disclosed. (Perplexity AI)
Grok (Aurora) + Grok Imagine	Image, Image→Video	X app / Grok; Imagine for animation	Native image gen; image-to-video with multiple styles; “Spicy” mode	Text-to-video not directly supported; policy concerns	Image text varies; sparse public evals	xAI release + coverage. (x.ai)
Claude 3/3.5/4 family	Assistive (no native T2I)	Claude UI/API	Excellent prompt engineering; HTML/CSS layouts; vision analysis	No direct image synthesis	N/A	Official help center says no image generation. (Claude Help Center)
Canva Magic Studio	Image, Video	Canva app + apps marketplace	Beginner-friendly; integrates DALL·E/Imagen	Generic outputs if prompts are vague; quality varies by app	Decent with style controls; dense	Canva pages list DALL·E and Imagen apps;

System / Model	Modality	How you access it	Notable strengths	Common caveats	Text-in-image reliability	Notes
(Magic Media)			apps; end-to-end editing		text still tricky	Magic Media help. (Canva)
Sora (OpenAI)	Video	Not broadly public; partner content; OpenAI showcases	High realism, temporal consistency; "up to a minute"	Limited public access; safety gating	N/A	OpenAI Sora page. (OpenAI)
Veo 3 / 3.1 (Google)	Video (8s variants; audio)	Gemini app; AI Studio / Vertex AI	Fast 8s clips incl. audio; social-ready; improving formats (vertical)	Length limits; pricing/tiers	N/A	Gemini page; recent coverage on 9:16 & pricing. (Gemini)
Runway Gen-2	Video	Runway app/API	Mature editor + effects; text/image→video	Some artifacts; length limits	N/A	Runway pages. (Wikipedia)
Midjourney Video v1	Video	Midjourney (beta tools)	Artistic motion; stylized control	Early-stage; changing features	N/A	Midjourney updates/docs. (Perplexity AI)
Pika (2.x)	Video	Pika app	Quick stylized clips; "Pikaframes"	Varies by prompt; short duration	N/A	Pika site / tutorials. (glbgpt.com)

Substitute Coding-Oriented Methods

Developers can also build text-to-image workflows manually for better control.

These are some of the steps to understand text to image generation:

1- Python libraries
2- Stable Diffuser
3- Web development tools

Python Utilities: Matplotlib and Pillow

Pillow supports adding text to images with custom fonts, though it needs careful placement adjustments. Matplotlib is better for annotated data visuals but requires more code. A lot of tweaks are required and often the image generated is not at par with the image generation capabilities of the GPTs.

Stable Diffusion Framework

Stable Diffusion uses pre-trained diffusion weights to generate images from text. Initial outputs can be messy, but fine-tuning on relevant datasets significantly improves quality. Techniques like self-play fine-tuning help adapt the model for personalized or domain-specific needs. I tried with pretrained self-diffusion models like Stable Diffusion but a lot of hallucination was observed (approximately 90 percent) hence fine-tuning is required.

Web Development tools

HTML and CSS can render text-based visuals like web banners, later converted to formats like PNG through APIs. While code-heavy, they offer precise design control and flexibility for dynamic visuals.

Future Scope

Looking ahead, creative tools are moving fast toward video for animations. In animation, companies are looking to automate the animation process using AI by generating short movies, episodes when prompts (scripts) and images are given as inputs. Midjourney is testing text-to-video features in artistic styles. Runway specializes in text-to-video and image-to-video transitions with lifelike motion. Pika offers text-to-video with features like “Pikaframes” for animation. These tools are expected to bring longer, higher-quality, and more customizable outputs soon.

Final Thoughts

Text-to-image and video generation have evolved rapidly. ChatGPT currently leads the field, while platforms like Canva offer the most practical, all-in-one creative solutions. Despite ongoing challenges like hallucinations, combining AI outputs with correction strategies and hybrid workflows can produce more reliable results. The future looks set for sharper accuracy and richer creativity.

References

1. Improving Image Generation with Better Captions - OpenAI (https://cdn.openai.com/papers/dall-e-3.pdf)

2. Paper Summary #12 - Image Recaptioning in DALL-E 3 (https://shreyansh26.github.io/post/2024-02-18_dalle3_image_recaptioner/)

3. Technical Details of DALL-E 3? : r/singularity - Reddit (https://www.reddit.com/r/singularity/comments/16xj668/technical_details_of_dalle_3/)

4. Video generation models as world simulators | OpenAI (https://openai.com/index/video-generation-models-as-world-simulators/)

5. Collection of Dall-E 3 prompting tips, issues and bugs (https://community.openai.com/t/collection-of-dall-e-3-prompting-tips-issues-and-bugs/889278)

6. [2408.07009] Imagen 3 - arXiv (https://arxiv.org/abs/2408.07009)

7. Paper page - Imagen 3 - Hugging Face (https://huggingface.co/papers/2408.07009)

8. [PDF] Imagen 3 - Googleapis.com (https://storage.googleapis.com/deepmind-media/imagen/imagen_3_tech_report_update_dec2024_v3.pdf)

9. Imagen - Google DeepMind (https://deepmind.google/models/imagen/)

10. Google's new Imagen 3 compared to other leading AI Image models (https://www.reddit.com/r/google/comments/1hfqdvs/googles_new_imagen_3_compared_to_other_leading_ai/)

11. Meet Magic Studio | Canva's AI Tools (https://www.canva.com/magic/)

12. Introducing Magic Studio: the power of AI, all in one place - Canva (https://www.canva.com/newsroom/news/magic-studio/)

13. Your all-in-one AI assistant - Canva AI (https://www.canva.com/ai-assistant/)

14. Meet MAGIC STUDIO - Canva's BEST AI Features - LaShonda Brown (https://www.lashondabrown.com/blog/meet-magic-studio)

15. How to Use Canva AI & Magic Studio | Full Tutorial for Beginners (https://www.youtube.com/watch?v=6EIf2Hmia60)

16. Claude Skills: Customize AI for your workflows - Anthropic (https://www.anthropic.com/news/skills)

17. Claude.ai (https://claude.ai/)

18. Introducing Claude 4 - Anthropic (https://www.anthropic.com/news/claude-4)

19. What is Claude AI, and how does it compare to ChatGPT? - Pluralsight (https://www.pluralsight.com/resources/blog/ai-and-data/what-is-claude-ai)

20. What Is Claude AI? - IBM (https://www.ibm.com/think/topics/claude-ai)

21. Grok Image Generation Release | xAI (https://x.ai/news/grok-image-generation-release)

22. Image Generations - xAI Docs (https://docs.x.ai/docs/guides/image-generations)

23. Grok Image Generator: My Hands-On Guide to xAI's Visual Revolution (https://skywork.ai/skypage/en/Grok-Image-Generator:-My-Hands-On-Guide-to-xAI%27s-Visual-Revolution/1976187330244571136)

24. Beginner's notes on Grok Imagine: tips, limits, and what actually works (https://www.reddit.com/r/grok/comments/1ml5yv3/beginners_notes_on_grok_imagine_tips_limits_and/)

25. X Image Generator - Grok AI (https://ximagegenerator.com/)

26. When AI Gets It Wrong: Addressing AI Hallucinations and Bias (https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/)

27. Understanding and Mitigating AI Hallucination - DigitalOcean (https://www.digitalocean.com/resources/articles/ai-hallucination)

28. AI Strategies Series: 7 Ways to Overcome Hallucinations (https://insight.factset.com/ai-strategies-series-7-ways-to-overcome-hallucinations)

29. LLM Hallucinations in 2025: How to Understand and Tackle AI's ... (https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models)

30. What Are AI Hallucinations? - IBM (https://www.ibm.com/think/topics/ai-hallucinations)

31. High-Resolution Image Synthesis with Latent Diffusion Models - arXiv (https://arxiv.org/abs/2112.10752)

32. Stable Diffusion Paper : r/StableDiffusion - Reddit (https://www.reddit.com/r/StableDiffusion/comments/18cu6vz/stable_diffusion_paper/)

33. SDXL: Improving Latent Diffusion Models for High-Resolution Image ... (https://arxiv.org/abs/2307.01952)

34. Stable Diffusion 3: Research Paper - Stability AI (https://stability.ai/news/stable-diffusion-3-research-paper)

35. CompVis/stable-diffusion: A latent text-to-image diffusion model (https://github.com/CompVis/stable-diffusion)

36. Video - Midjourney (https://docs.midjourney.com/hc/en-us/articles/37460773864589-Video)

37. Complete Guide to Midjourney Video Generator (https://www.youtube.com/watch?v=Z1bIGyXnh6U)

38. Correct me if I'm wrong, but does the new Midjourney video gen ... (https://www.reddit.com/r/midjourney/comments/1lgbg7j/correct_me_if_im_wrong_but_does_the_new/)

39. Introducing Our V1 Video Model - Midjourney (https://updates.midjourney.com/introducing-our-v1-video-model/)

40. Create Cinematic AI Video Using Midjourney (https://www.youtube.com/watch?v=cpr-Q2aoAjo)

41. Gen-2: Generate novel videos with text, images or video clips (https://runwayml.com/research/gen-2)

42. Runway | AI Image and Video Generator (https://runwayml.com/)

43. Generative Video from Text with RunwayML (https://www.youtube.com/watch?v=klpyL9HlwFA)

44. Text to video : r/runwayml - Reddit (https://www.reddit.com/r/runwayml/comments/1ig6735/text_to_video/)

45. Runway AI Video Generator: Create Videos from Text/Image Free (https://www.easemate.ai/runway-ai-video-generator)

46. Pika (https://pika.art/)

47. Pika AI [Free Trial] - Monica (https://monica.im/ai-models/pika-ai)

48. Pika AI Free: Try Pika Art AI Video Generator (Pika Labs) - Pollo AI (https://pollo.ai/m/pika-ai)

49. Pika Labs: Introducing Pika 1.0 (AI Video Generator) - Reddit (https://www.reddit.com/r/singularity/comments/185yapi/pika_labs_introducing_pika_10_ai_video_generator/)

50. Create AMAZING Videos & Ai VFX | PIKA Ai 2.2 Tutorial (https://www.youtube.com/watch?v=x1z3Ypd49hs)

51. Video generation models as world simulators | OpenAI (https://openai.com/index/video-generation-models-as-world-simulators/)

52. Sora: Creating video from text - OpenAI (https://openai.com/index/sora/)

53. Sora: A Review on Background, Technology, Limitations, and ... - arXiv (https://arxiv.org/abs/2402.17177)

54. OpenAI Sora's Technical Review - Jianing Qi (https://j-qi.medium.com/openai-soras-technical-review-a8f85b44cb7f)

55. [PDF] SORA AI: The Future of Video Generation - TechRxiv (https://www.techrxiv.org/users/909988/articles/1283935/master/file/data/SORA_AI_Paper%5B1%5D/SORA_AI_Paper%5B1%5D.pdf)

56. Veo - Google DeepMind (https://deepmind.google/models/veo/)

57. Veo 3 | Google AI Studio (https://aistudio.google.com/models/veo-3)

58. Gemini AI video generator powered by Veo 3.1 (https://gemini.google/overview/video-generation/)

59. Generate videos with Veo on Vertex AI in Vertex AI (https://docs.cloud.google.com/vertex-ai/generative-ai/docs/video/overview)

60. Veo (text-to-video model) - Wikipedia (https://en.wikipedia.org/wiki/Veo_%28text-to-video_model%29)