

An Analysis of Text-to-image Models of OpenAI, Stability AI, and Google

Zijie Li¹, Mutian Yang^{2,*}, Shuo Yang³

¹College of Artificial Intelligence, Beijing Information Technology College, Beijing, China

²College of letter and science, University of California, Santa Barbara (UCSB), Santa Barbara, California, United States of America

³College of Information and Mechatronic Engineering, Shanghai Normal University (shnu), ShangHai, China

*Corresponding author: mutian_yang@ucsb.edu

Abstract. Text-to-image generation has rapidly evolved through a series of significant models since it was first introduced in 2015. This paper examines the development of leading models: OpenAI's DALL·E 1–3 and GPT-4o, Stability AI's Stable Diffusion series (v1.5, XL, 3.0), and Google's Imagen 1–3. These models have an astonishing overlap in the time points when they are updated and iterated. Also, they have similar technological focuses and development trajectories at these time knots, although various methods have been applied, ranging from Transformer-based autoregressive designs to latent diffusion with CLIP conditioning. By setting each year from 2021 to 2024 as one of the time knots, this paper compared the techniques they used at these nodes horizontally, identifying the convergence of their emphases and uses of technologies. Estimated the progress of the models vertically, this paper has verified the necessity and effectiveness of model's iterations. This study also identified some existing issues with the model, with some possible solutions.

Keywords: Text-to-image generation; Multimodal artificial intelligence; Transformer; Latent Diffusion Model (LDM).

1. Introduction

Text-to-image generation refers to the process of creating images from text descriptions. It is a key task in the field of multimodal artificial intelligence and is closely linked to natural language processing and computer vision. In 2015, AlignDRAW—one of the earliest recognized deep learning text-to-image models—was launched. Since then, this innovative task has gained attention, which in turn has driven more research into generative tasks [1]. As a result, the quality of model-generated outputs has improved: the resolution of generated images has gradually increased from blurry results, and more details in text descriptions can be captured. By 2021, with the wider adoption and integration of technologies such as Transformers and diffusion models, text-to-image models had undergone more frequent upgrades and reached unprecedented levels of performance. Under this trend, many companies—including OpenAI (which released the DALL·E series and GPT-4o), Stability AI (which released the Stable Diffusion/SD series), and Google (which released the Imagen series)—have rolled out their own text-to-image models [2]-[4]. These models have gone through multiple iterations since their initial release and now lead the field. Notably, their iteration timelines coincide: they tend to update at similar time points, and their technological enhancement directions share commonalities. This paper treats each year from 2021 to 2024 as one of four key "time knots" for model changes and discusses the status of the models at these knots. These models adopt diverse technical methods. DALL·E models use Transformer-based designs to map text to images [2], [5]-[7]; Stable Diffusion models employ latent diffusion with CLIP-based conditioning [3], [8]-[9]; Imagen models combine text embeddings derived from large language models (LLMs) with diffusion. These methods are detailed in the following section [4], [10].

2. Method

This section introduces the methods adopted by each model in chronological order and focuses on conducting horizontal comparisons of the methods used by different models at the same time.

2.1. Time Knot 1: January 2021

At this first key time knot, OpenAI developed DALL-E 1, while Google's Imagen and Stability AI's Stable Diffusion remained unreleased. As Aditya Ramesh et al. described [2], DALL-E 1 generates images in four steps: first, convert the input prompt to a BPE-encoded text token and add a "start-of-image token"; second, feed the prompt token into a trained autoregressive Transformer to generate an image token (repeated 21,024 times until all 32×32 tokens are obtained); third, recover these tokens into a 256×256 RGB image via a discrete variational autoencoder (dVAE); finally, sample candidate images from the Transformer and output the one with the highest CLIP contrastive model score. Text and image tokens are treated as the same sequence, enabling zero-shot text-to-image generation. Ramesh et al. further noted that the model's training has two main stages [2]. Stage 1 trains the dVAE to compress 256×256 RGB images into 32×32 discrete tokens (ensuring reversible compression). Each token has 8,192 possible values, shortening the Transformer-bound token sequence by $\sim 192 \times$ while preserving key visual structures. Both the dVAE's encoder and decoder are convolutional ResNets, outputting $32 \times 32 \times 8192$ logits (corresponding to token classification at each position). Stage 2 splices text and image tokens, then feeds them into a 12-billion-parameter, 64-layer decoder-only Transformer. This Transformer uses sparse attention patterns (row, column, convolutional masks) and is trained autoregressively to predict the next token (text or image) in the sequence. These trainings let DALL-E 1 combine multiple concepts, and its zero-shot capability delivers high-quality generation even without training on Microsoft Common Objects in Context (MS-COCO) dataset captions.

2.2. Time Knot 2: April-October 2022

In 2022, Stability AI and Google entered the field with SD v1.5 and Imagen 1 respectively, while DALL-E 1 evolved into DALL-E 2. As Ramesh et al. introduced [5], DALL-E 2 generates images in two stages. Stage 1 (hierarchical generation) converts text to a text embedding via a CLIP text encoder, feeds it into an autoregressive Transformer or diffusion model for a low-resolution CLIP image embedding (capturing key objects and layout), sends latent vectors to a refiner for texture/details/local structures, and then performs progressive upscaling (each level's output as the next input) to retain details and ensure consistency. Stage 2 uses VQGAN and other decoders (trained to minimize pixel-level/perceptual-level reconstruction errors) to restore high-resolution latent vectors to pixel images. Per Chitwan Saharia et al. [4], Imagen 1 also uses a diffusion model: it converts input text to a latent representation and outputs an image after multiple denoising steps. When generating, it first uses a pre-trained large language model to create a text embedding (capturing prompt semantics), then a base diffusion model takes the embedding and random noise to generate a 64×64 image via iterative denoising. A cascaded super-resolution diffusion model then upscales it to 256×256 (first model) and 1024×1024 (second model) [11]. Coincidentally, SD v1.5 relies on a diffusion model too [3]. Like the other two models, it first converts text prompts to token embeddings, prepares a Gaussian noise-filled latent tensor (matching VAE's latent space shape), performs iterative denoising (step-by-step in a time loop), and then passes the clean latent to a pre-trained autoencoder's decoder D for a high-resolution RGB image. At this time, all three models targeted high-resolution image generation, used diffusion models as aids, and independently adopted similar hierarchical approaches to improve output quality.

2.3. Time Knot 3: July-December 2023

In 2023, significant advancements were made across the board as OpenAI released DALL-E 3, Google launched Imagen 2, and Stability AI updated SD from v1.5 to SDXL. OpenAI has not

disclosed the architecture or training methods of DALL-E 3 due to its proprietary nature. However, information from official sources and related research indicates it is built on DALL-E 2 and integrated with ChatGPT [12], [13]. This integration allows for iterative user-dialogue-based updates, reducing traditional workflow burdens. Its Transformer-based architecture, trained on large-scale paired data, excels at capturing sequential context and mapping text to visuals, emphasizing prompt adherence. SDXL, like SD v1.5, is based on the Latent Diffusion Model framework [8]. The initialization phase involves obtaining text embeddings from two encoders and a Gaussian noise representation. Denoising is split into two stages: the first uses a larger U-Net (three times the size of previous versions with more attention blocks for cross-attention expansion), and the second employs a refiner for enhanced visual details. The refined representation is then decoded into an RGB image via a VAE. SDXL has a more compact yet larger-scale structure than previous versions, enhancing efficiency, and uses an advanced attention-control module for better semantic alignment. Imagen 2 continues to use Latent Space technology [4]. It maps images to a low-dimensional space to cut computational costs and boost generation efficiency. Its core process remains denoising latent representations and reconstructing real-world images. Although exact details are scarce, official reports suggest it integrates techniques such as multi-scale CNNs for detail improvement and advanced regularization for diversity and accuracy [14]. After iterative denoising, it generates highly accurate prompt-following images. At this stage, all three models centered on improving the alignment between generated images and text descriptions. Their efforts to expand model scales and enhance training data quality also reflect their pursuit of better generation outcomes.

2.4. Time Knot 4: May-August 2024

In 2024, SD v3.0 was released, Imagen evolved to its third generation, and OpenAI launched GPT-4o—a unified text-image model. GPT-4o integrates text and image processing in one model [15]. For image generation, its language understanding component first processes text prompts to create rich semantic representations, directly reducing noise in original prompts [7]. Though its detailed architecture is undisclosed, benchmark tests like GenEval show it outperforms DALL-E 3 in prompt-image alignment [16], detail rendering, and attribute consistency, proving these areas remain key for text-to-image model breakthroughs. SD v3.0 follows SDXL’s overall strategy [9]. In initialization, text prompts are tokenized and embedded via two CLIP encoders, with a latent space representation prepared for later denoising. Denoising steps are similar to the previous generation, but it uses a larger U-Net backbone (with more residual blocks and self-attention layers) and integrates Transformer attention modules (self-attention + cross-attention) to capture more details. Finally, the denoised latent is reconstructed into a final RGB image via a VAE decoder. Imagen 3 retains Google’s focus on diffusion strategies while upgrading architecture and mechanisms [10]. In initialization, input text prompts are encoded via a T5-XXL encoder to generate rich language embeddings. An initial low-resolution latent representation is created and refined progressively via cascaded diffusion models to boost resolution. Each stage uses diffusion Transformers with cross-attention layers (to integrate text embeddings and maintain semantic accuracy) and self-attention (to capture details). The highest-resolution latent is decoded into an output image via a VAE, resulting in highly photorealistic images with strong prompt adherence. Notably, Imagen 3’s architecture is now highly similar to successive Stable Diffusion versions: it adopts a Transformer-based backbone and integrates self-attention/cross-attention modules to enhance prompt adherence and detail rendering. Though GPT-4o’s exact architecture is unpublished, it pursues similar goals: improving prompt fidelity, compositional accuracy, and clear visual detail rendering.

3. Comparative Analysis

In the section above, a structured understanding of the techniques and methods of models at various time points is given. Building on this foundation, this article now applies these insights to quantitative evaluations.

3.1. Evaluative Indicators

When testing and evaluating the performance of these models, multiple quantitative indicators were employed.

Fréchet Inception Distance (FID). As introduced by Martin Heusel et al. in [17], FID is a metric used to measure the quality of images produced by generative models. It compares the distribution of generated images against that of a set of real images, and the distance between them is computed. A significant model is expected to have a low FID score.

$$FID^2 = ||\mu - \mu'||^2 + tr\left(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{\frac{1}{2}}\right) \quad (1)$$

In this part, this article focuses on the Zero-shot FID scores of models on the Microsoft Common Objects in Context (MS-COCO) 256*256 dataset [18], meaning models have not been fine-tuned on the image-text pairs from MS-COCO.

MS-COCO Caption CLIP score. The CLIP score is an automatic evaluation metric depending on a pretrained CLIP model [19]. A higher score implies a better capture of the inputted description by the model.

GenEval. According to Dhruva Ghosh et al. in [16], GenEval describes compositional image properties and evaluates models from multiple aspects, including color, count, position, appearance of multiple objects, etc. Range from 0 to 1. The model with a score closer to 1 is implied to be a better model.

Win rate (%). Win rate is a quantified measurement of human preference in pairwise comparisons. This is usually done by asking reviewers to choose one image among two of which were generated by two different models with the same input caption. The model with a higher win rate is expected to be a better model.

Elo score. The Elo rating system is a rating system that is based on pairwise comparisons and focuses on human preference [20]. When being employed in model comparison, the Elo rating system effectively reflects the human preference for models and is adaptable to dynamic updates.

Data and evaluations on these models were obtained from a range of papers, the score variation under the same metric can reflect the image generation capabilities and trend of these variations.

3.2. OpenAI

Table 1. Performance comparison of DALL·E series and GPT-4o (data from [5], [7], [21])

	Zero-shot FID	MSCOCO Caption CLIP score	GenEval(overall)
DALL-E1	28		
DALL-E2(diffusion)	10.39	31.4	0.52
DALL-E3		32.0	0.67
GPT-4o			0.84

As shown in Table 1, the performance of OpenAI’s text-to-image models has continuously improved from DALL-E 1 to GPT-4o. In terms of zero-shot FID, FID of DALL-E 2 is significantly lower than that of DALL-E 1, indicating a significant improvement in the accuracy of capturing descriptions. Regarding the MSCOCO Caption CLIP score, DALL-E 3 has improved compared to DALL-E 2. This reflects the better capability of DALL-E 3 in aligning the alignment between input text and the generated image. For the GenEval overall score, a continuous increase from DALL-E 2 to GPT-4o. These results demonstrate the stronger overall performance of GPT-4o on various

aspects. Fig. 1 shows the performance comparison of OpenAI’s text-to-image models across Zero-shot FID, MSCOCO Caption CLIP score, and GenEval overall score.

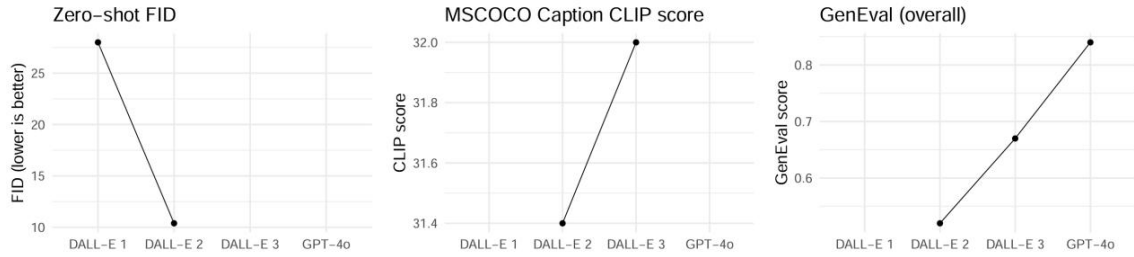


Fig. 1 Performance comparison of OpenAI’s text-to-image models across Zero-shot FID, MSCOCO Caption CLIP score, and GenEval overall score (Data from: [5], [7], [21])

3.3. Stability AI

Table 2. Performance comparison of Stable Diffusion series [8, 9]

	Win rate (%)	GenEval (overall)
SD v1.5	7.91	0.43
SDXL	36.93	0.55
SD v3.0		0.74

The results in Table 2 demonstrate the performance of the Stability AI model across different generations. SD v1.5 has the lowest Win rate and GenEval total score. Compared with later versions, it indicates that its performance in image generation tasks is relatively limited. With the release of SDXL, both indicators have increased. The latest model SD v3.0 achieved the highest total score in GenEval, showing significant improvement compared to the previous two generations. This stable improvement indicates that during the iteration process of the Stability AI model, the image quality, input capture, and overall generation ability have all been improved. Fig. 2. shows the performance comparison of Stable Diffusion models across generations in terms of Win Rate and GenEval overall score.

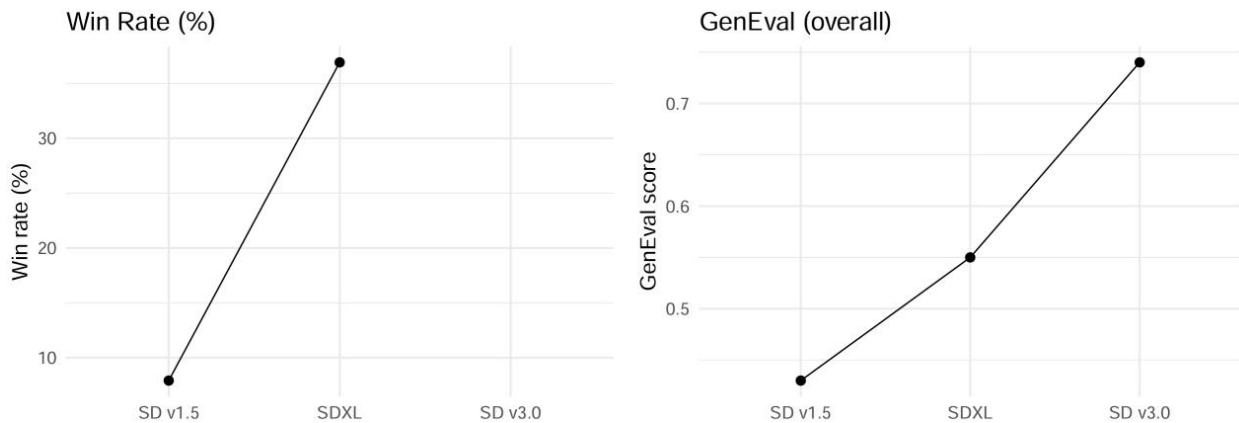


Fig. 2. Performance comparison of Stable Diffusion models across generations in terms of Win Rate and GenEval overall score (Data from: [8, 9])

3.4. Google

Table 3. Elo score comparison of Imagen series across three benchmark evaluations [10]

	Imagen 2	Imagen 3
GenAI-Bench [22]	941	1089
DrawBench [23]	929	1068
DALL-E 3 Eval [21]	874	1079

Some qualitative reports on Imagen 2 indicate that the images generated by Imagen 2 are brighter, more accurate, have more diverse styles, and have better input capture [14]. These observations collectively indicate a significant improvement in performance from Imagen 1 to its next-generation. Then, Table 3 above provides a quantitative comparison between Imagen 2 and Imagen 3, showing the ELO scores of these two models on three different benchmark datasets: GenAI Bench focuses on cue following and image quality, DrawBench tests the breadth of the model, and Dall-E 3 Eval tests the performance of the model when competing with Dall-E 3. Under each benchmark dataset, the ELO scores differ by at least 1000 points, and Imagen 3 is significantly better than Imagen 2 in human preference-based ELO evaluation. These results indicate that from Image 1 to Image 3, Google has achieved continuous improvements in model architecture, training data, and fast understanding from text to image generation. Fig. 3. Shows the elo score comparison between Imagen 2 and Imagen 3 across three benchmark datasets: GenAI-Bench, DrawBench, and DALL·E 3 Eval.



Fig. 3. Elo score comparison between Imagen 2 and Imagen 3 across three benchmark datasets: GenAI-Bench, DrawBench, and DALL·E 3 Eval (Data from: [10])

4. Discussion

From the section above, a clear upward trend in the generative capability of the models and be observed. However, there are still some limitations and problems with the latest models: prompts' semantic capture, lack of ability to generate specific content, challenges with high resolution and complex scenes.

4.1. Consistency of Input Text and Image

Across all three models, maintaining the accuracy of description captures remains imperfect in complex scenarios. GPT-4o has shown strong ability in Inferring requirements from prompts, but still has problems with consistency, [7]. A similar problem also appears in Image 3; partially misplaced output will be produced in some complicated scenes [24]. Although SD v3.0 has optimized the architecture to prevent these problems, failure in following the prompt still happens occasionally [25].

One of the possible solutions for partial painting errors is to redraw the problematic sections. By masking out the part that is easy to make mistakes and draw again with high resolution, problems can be alleviated to some extent. Then, to solve the insufficient functionality when creating an image with high resolution, making good use of the super-resolution technique of AI is a good choice. When generating an image with relatively lower resolution and then enhancing it to the target resolution, obvious structural generation errors can be avoided.

4.2. Limited Generations in a Specific Domain and Culture

GPT-4o struggles with generating Chinese, Japanese, and Arabic symbol elements [7]. Meanwhile, SD v3.0 may falter on prompts involving certain human body poses [26].

One idea for this is to use an attention-based formulation of Generative Semantic Nursing (GSN), which intervenes in the generative process and forces the model to allocate attention to all subject tokens. This method relies on a precise understanding of the prompts of the text encoder; if multiple

words are merged or split inappropriately, this method will lose its significance and even exacerbate the situation.

5. Conclusion

This article reviewed text-to-image generative models from OpenAI, Stability AI, and Google, compared methods and results of their successive generations, and identified consistent upward trends through all metrics, providing strong evidence for the significance and necessity of method changes. Noticeable coincidences of models in the selection of phased goals and strategies are also identified, implying shared common technological opportunities and challenges. Yet current models still face problems on prompt capture, limited generation, and some ethnic concerns, highlighting the need for continued method innovation and fine-tuning of details.

6. Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] Mansimov E, Parisotto E, Ba JL, et al. Generating images from captions with attention. arXiv preprint arXiv:1511.02793, 2015.
- [2] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. International Conference on Machine Learning, 2021: 8821-8831.
- [3] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684-10695.
- [4] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 2022, 35: 36479-36494.
- [5] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.
- [6] Analytics Vidhya. DALL-E 3: Key features and improvements. Analytics Vidhya, 2024. [Online]. Available: <https://www.analyticsvidhya.com/blog/2024/07/dall-e3/>
- [7] Yan Z, Ye J, Li W, et al. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. arXiv preprint arXiv:2504.02782, 2025.
- [8] Podell D, English Z, Lacey K, et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023.
- [9] Esser P, Kulal S, Blattmann A, et al. Scaling rectified flow transformers for high-resolution image synthesis. International Conference on Machine Learning, 2024.
- [10] Baldridge J, et al. Imagen 3. arXiv preprint arXiv:2408.07009, 2024.
- [11] Ho J, Saharia C, Chan W, et al. Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research, 2022, 23(47): 1-33.
- [12] OpenAI. DALL·E 3. OpenAI, 2025. [Online]. Available: <https://openai.com/dall-e-3/>
- [13] OpenAI. DALL·E 3 System Card. OpenAI, 2023. [Online]. Available: <https://openai.com/index/dall-e-3-system-card/>
- [14] Tirumalashetty V. Imagen 2 on Vertex AI is now generally available. Google Cloud Blog, 2023. [Online]. Available: <https://cloud.google.com/blog/products/ai-machine-learning/imagen-2-on-vertex-ai-is-now-generally-available>
- [15] OpenAI. GPT-4o System Card. OpenAI, 2024. [Online]. Available: <https://openai.com/index/gpt-4o-system-card/>
- [16] Ghosh D, Hajishirzi H, Schmidt L. Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems, 2023, 36: 52132-52152.

- [17] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 2017, 30.
- [18] Lin TY, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. *European Conference on Computer Vision*, 2014: 740-755.
- [19] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021: 8748-8763.
- [20] Elo AE, Sloan S. The rating of chessplayers: Past and present. 1978.
- [21] Shi Z, Zhou X, Qiu X, et al. Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*, 2020.
- [22] Li B, Lin Z, Pathak D, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.
- [23] Bakr EM, Sun P, Shen X, et al. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. *IEEE/CVF International Conference on Computer Vision*, 2023: 20041-20053.
- [24] Analytics Vidhya. Is Google's Imagen 3 the future of AI image creation. Analytics Vidhya, 2024. [Online]. Available: <https://www.analyticsvidhya.com/blog/2024/09/google-imagen-3/>
- [25] Analytics Vidhya. Stable Diffusion 3: Guide to the latest text-to-image model by Stability AI. Analytics Vidhya, 2024. [Online]. Available: <https://www.analyticsvidhya.com/blog/2024/06/stable-diffusion-3/>
- [26] Xu S. Stable Diffusion 3: Text master, prone problems. BentoML Blog, 2024. [Online]. Available: <https://www.bentoml.com/blog/stable-diffusion-3-text-master-prone-problems>