JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text-to-Image Generators Explored: A Comprehensive Review on Generative AI

[1]Rushabh Dhamne, [2]Akash Chaudhari, [3]Prashik Gawai, [4]Prajkta Khaire

[1]Student, [2]Student, [3]Student, [4]Assistant Professor
[1]Department of Information Technology,
[1]Shivajirao S. Jondhale College of Engineering, Dombivli, India

Abstract: Text-to-image generation is the process of converting input prompts into high-quality images using some of the stable diffusion models that are mostly used in the industry. Industries are Highly growing towards the AI generation that helps generate more productivity and Text-to-Image generation is one of them. Diffusion models described in several research papers highlight the progress made in text-to-image generation through diffusion models, specifically focusing on the Stable Diffusion model. While these models produce high-quality and creative results, their multi-step sampling process is slow, often requiring many iterations for satisfactory outputs. Attempts to speed up this process through distillation have not been successful in creating a functional one-step model. The papers also mention other text-to-image generation models but suggest the Stable Diffusion model as a preferred option for enhanced productivity.

IndexTerms **– Diffusion Models, ControlNet Integration, Shifted Diffusion Model, Generative Frameworks, Stable Diffusion, T2I models, Diffusion Denoising Probabilistic Models, FID scores.**

I. INTRODUCTION

In recent years, the realm of artificial intelligence (AI) has seen remarkable progress in generating visual content from textual descriptions through Text-to-Image Generators. These systems, fueled by generative AI, can translate written words into lifelike images. This paper presents a comprehensive exploration of Text-to-Image Generators, examining various models introduced in recent years to bridge the gap between text and images.

Text-to-Image Generators mark a significant convergence of natural language processing and computer vision, offering exciting possibilities across multiple domains. For instance, consider a scenario where a user inputs a description like "a red apple on a wooden table." Through the power of Text-to-Image Generation, the system can produce a corresponding image depicting precisely that scene. Such capabilities have sparked immense interest among researchers and practitioners, eager to harness this technology's potential in design, entertainment, and online retail.

This review delves into the mechanisms and advancements of Text-to-Image Generators, exploring various models introduced in recent years. By examining these models, we aim to understand their capabilities and limitations in generating realistic images from text inputs. However, challenges such as maintaining image quality and ensuring diversity in generated outputs persist. Overcoming these obstacles requires continuous innovation and collaboration within the AI community. Through this review, we aim to provide insights into the current landscape of Text-to-Image Generation and inspire further advancements in this field.

II. REVIEW OF LITERATURE

In 2023, The paper "VectorFusion: Text-to-SVG by Abstracting Pixel-Based Diffusion Models" [1] introduces a novel approach for generating Scalable Vector Graphics (SVGs) from text captions. It addresses the challenge of creating vector representations from pixel-based diffusion models, offering a solution for digital art and design workflows.

Graphic designers and artists often express concepts abstractly, using shapes and lines to evoke scenes. A declarative format for this kind of expression is provided by Scalable Vector Graphics (SVGs). Despite the difficulty in designing vector graphics, recent advancements like diffusion models open possibilities for generating high-quality SVGs from text. This work introduces VectorFusion, a method bridging text-to-image diffusion models with SVG synthesis, enhancing coherence and quality in generated graphics.

Diffusion Models: The paper discusses the use of diffusion models for text-to-image synthesis, highlighting their ability to generate raster images from captioned datasets. It also emphasizes the limitations of existing diffusion models in directly generating vector graphics, such as SVGs, and the challenges in adapting them for this purpose.

Method: VectorFusion: The proposed VectorFusion method leverages a pretrained diffusion model to generate abstract SVG representations from text captions. It employs a differentiable vector graphics renderer and a score distillation sampling (SDS) loss to refine shape parameters and optimize the generated SVGs to be consistent with the input text captions. The method demonstrates the generation of diverse styles, including iconography, pixel art, and line drawings.

Result: VectorFusion achieves greater quality than CLIP-based approaches in generating coherent and aesthetically pleasing vector graphics from text descriptions. It provides control over the style of art generated, enabling the synthesis of diverse and meaningful vector graphics. The method showcases potential for various applications in digital art and design workflows.

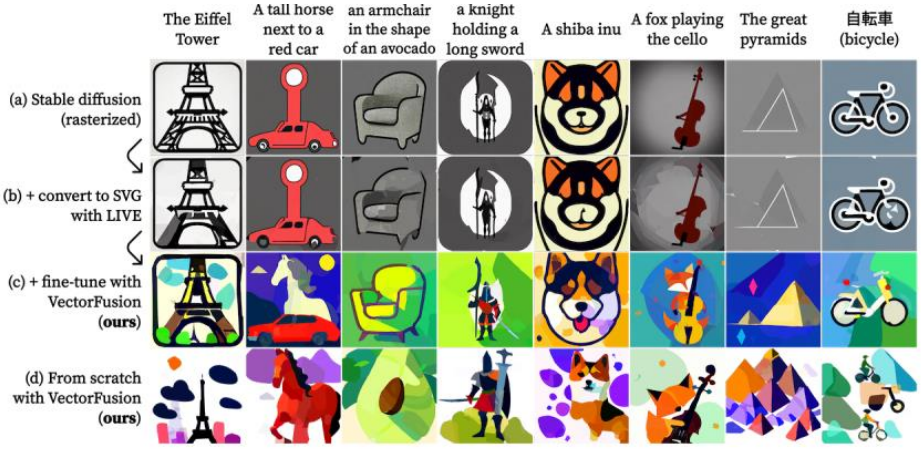


figure 1. text-to-svg with vectorfusion [1]

In 2023, a research paper titled “ReCo: Region-Controlled Text-to-Image Generation” [2] introduces a technique for regional control in text-to-image (T2I) generation, addressing the limited controllability of large-scale T2I models in precisely specifying content in specific regions with free-form text descriptions. It proposes augmenting T2I models' inputs with position tokens to enable precise region control for arbitrary objects described by open-ended regional texts.

ReCo Model: The ReCo (Region-Controlled T2I) model extends T2I models with the ability to understand coordinate inputs, allowing for accurate and open-ended regional control. By combining text and position tokens in the input query, ReCo achieves the best of both worlds in text-to-image and layout-to-image, enabling free-form description and precise position control.

Region-Controlled T2I Generation Results: Empirical results demonstrate that ReCo significantly improves region control accuracy and image generation quality across a wide range of datasets and designed prompts. It achieves better image quality, improved region control accuracy, and competitive performance in generating realistic images from text prompts.

Limitations: The limitations of traditional T2I models include limited controllability in precisely specifying content in specific regions with free-form text descriptions. The naive use of position-related text words often results in ambiguous and verbose input queries, making region control difficult and prompting the need for "prompt engineering."

Conclusion: The ReCo model presents a significant advancement in T2I generation by addressing the limitations of traditional models and achieving better controllability and image generation quality. It offers a more flexible input interface, alleviating controllability issues and providing a more effective and interaction-friendly conditioning signal while preserving the pre-trained T2I capability.

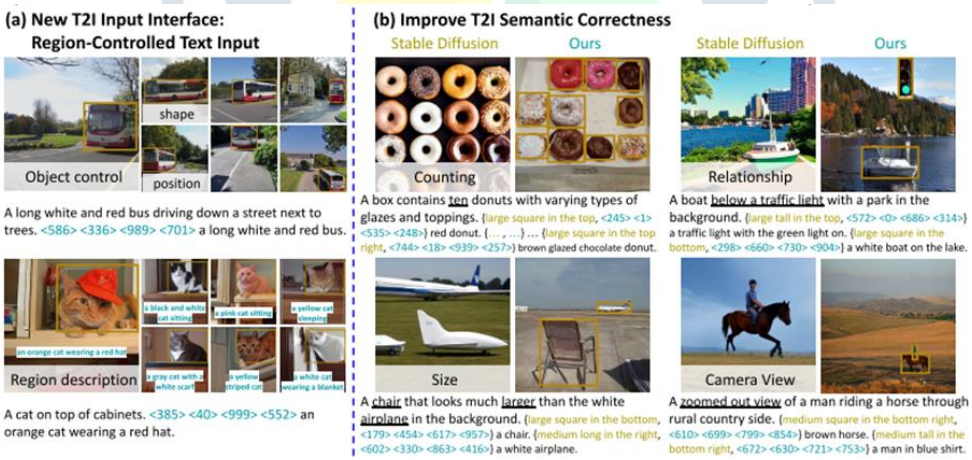


figure 2. enhancing text-to-image models with position tokens for improved control over object properties and semantic correctness in complex scenes.

In 2022, The paper "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models" [3] explores the application of guided diffusion models for text-conditional image synthesis. It aims to empower humans to create diverse visual content with ease and refine images using natural language prompts, critical for real-world applications.

Diffusion Models: GLIDE (Guided Language to Image Diffusion for Generation and Editing) [3] a 3.5 billion parameter text-conditional diffusion model that uses a text encoder to condition on natural language descriptions is introduced. It compares two techniques for guiding diffusion models towards text prompts: CLIP guidance and classifier-free guidance, finding that the latter yields higher-quality images.

Image Inpainting: The paper discusses the training of a 3.5 billion parameter text-conditional diffusion model and a 1.5 billion parameter upsampling diffusion model to increase resolution. It also fine-tunes the model to perform image inpainting, enabling powerful text-driven image editing. The model explicitly fine-tunes to perform inpainting, resulting in better image quality.

Limitations: Acknowledging the potential misuse of the model for creating disinformation or Deepfakes and releases a smaller diffusion model and a noised CLIP model trained on filtered datasets to safeguard against these use scenarios and supporting upcoming investigations.

Results: The results show that the classifier-free guidance technique yields higher-quality images, preferred by human evaluators for both photorealism and caption similarity. The model achieves high aesthetic quality and competitive performance in generating realistic images from text prompts, demonstrating its potential for diverse image generation and editing.

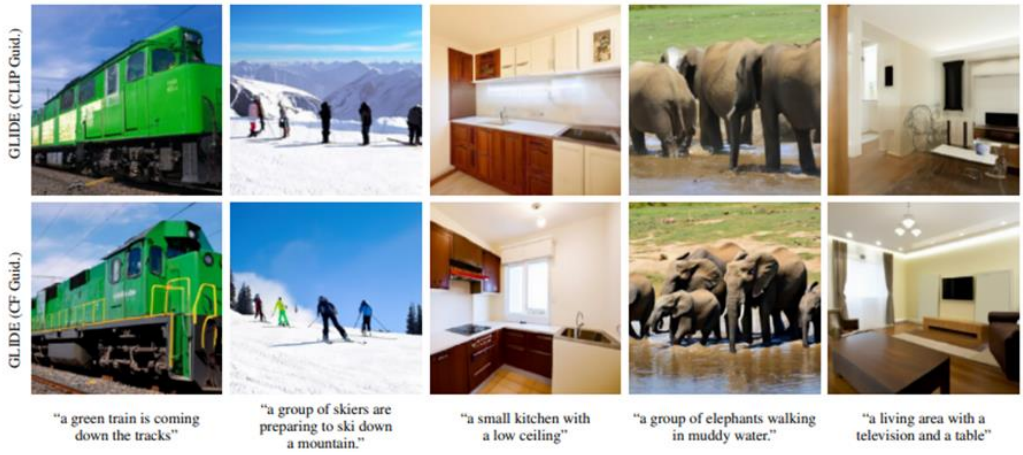


figure 3. images generated by glide

The 2023 exploration paper on "RAPHAEL Text-to-Image Generation via Large Admixture of Diffusion Paths" [4] by Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo describes us that there are different prolixity models like stable prolixity model [5] and just like that so this paper reviews us RAPHAEL prolixity model is a textbook- to- image creator that surpasses former prolixity model by using prolixity model with admixture- of- experts approach. Unlike models, RAPHAEL precisely aligns textbook generalities with image regions, performing with superior image dedication. Figure 1 represents the Comparisons of RAPHAEL with recent representative creators, Stable prolixity XL, Deep- Floyd, DALL- E 2, and ERNIE- ViLG2.0 [4].

The crucial donation and findings of this work include: Performance on COCO Dataset RAPHAEL achieves a state-of-the-art FID- 30k score of 6.61 on the COCO dataset, surpassing other prominent image creators like Stable prolixity [5], Imagen [4], ERNIE- ViLG2.0 [4] and DALL- E 2 [4]. mortal Evaluations Using the ViLG-300 standards, RAPHAEL outperforms challengers in terms of both image-textbook alignment and image quality in stoner studies. mortal artists, ignorant of the model generating the images, constantly rate RAPHAEL advanced.

Extensions with LoRA, ControlNet, and SR- GAN RAPHAEL can be extended with ways like LoRA, ControlNet, and SR-GAN for further improvement and robustness. From reviewing the excursus of RAPHAEL, RAPHAEL demonstrates superior performance against overfitting compared to Stable prolixit [5] and achieves enhanced image resolution using SR- GAN.

figure 4. comparisons of raphael with recent representative generators, stable diffusion xl, deep-floyd, dall-e 2, and ernie-vilg 2.0

In the year 2023, the paper "Adding Conditional Control to Text-to-Image Diffusion Models" by Lvmin Zhang, Anyi Rao, and Maneesh Agrawala helps to describe the delve into the advancements in text-to-image models and the challenges faced in controlling spatial composition within generated images. The introduction of text-to-image diffusion models has enabled the creation of visually stunning images through text prompts, yet limitations exist in precise spatial control. To address this, researchers have explored the use of additional images to specify desired image composition, such as edge maps, human pose skeletons, and segmentation maps, as conditioning inputs in the image generation process.

ControlNet emerges as a novel solution, offering an end-to-end neural network architecture that learns conditional controls for large pre-trained text-to-image diffusion models [5]. By locking the parameters of the large model and creating a trainable copy of its encoding layers, ControlNet facilitates efficient finetuning with spatially localized input conditions. This approach has been tested successfully with various conditioning inputs, demonstrating robust and scalable training outcomes, even rivaling industrial model performance on a single GPU.

Image Diffusion Models: Introduced by Sohl-Dickstein et al. [5], these models have been applied to image generation, with Latent Diffusion Models performing diffusion steps in latent image space to reduce computation cost.

Text-to-Image Diffusion: State-of-the-art results are achieved by encoding text inputs into latent vectors, with Stable Diffusion being a large-scale implementation of latent diffusion using a U-Net structure.

ControlNet Integration: ControlNet enhances large pre-trained text-to-image diffusion models by injecting additional conditions into neural network blocks for personalized image generation.

Training Stability: Stable Diffusion uses latent images for training stability and pre-processing methods similar to VQ-GAN, ensuring a stable training process.

Efficient Integration: The integration of ControlNet with Stable Diffusion improves training efficiency, saves GPU memory, and speeds up the training process.

figure 5. controlling stable diffusion with learned conditions.

The 2023 research paper on "Shifted Diffusion for Text-to-Image Generation" by Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu discusses the advancements in AI-generated content, focusing on high-fidelity text-aligned image synthesis [6]. Various models like DALL-E, Latent Diffusion Model, GLIDE [3], DALL-E 2 [3], Imagen [4], and Parti have been developed to enhance text-to-image generation quality and efficiency. In contrast, the Corgi model is introduced as a novel diffusion model that aims to improve the diffusion process itself for more effective text-to-image generation. Corgi bridges the image-text modality gap and data availability gap, enabling semi-supervised and language-free text-to-image generation. The model incorporates prior knowledge from pre-trained models like CLIP and achieves promising results with minimal captioned images in the training dataset.

Framework Overview: The text-to-image generation framework consists of three key components: a pre-trained image encoder, a decoder for image generation, and a prior model for generating image embeddings from text captions.

Flexibility and Adaptability: The framework allows for various generation tasks such as text-to-image, image-to-image, and conditional generation based on both image and text inputs. It supports semi-supervised training, enabling a mix of image-text pairs and pure images for training.

Focus on Prior Model: The paper focuses on enhancing the prior model, which is less explored in previous works. The proposed shifted diffusion model leverages prior knowledge from the pre-trained CLIP image encoder to improve the generation process by considering the effective output space of the CLIP image encoder.

Shifted Diffusion Model: The shifted diffusion model introduces a parametric noise distribution instead of the standard Gaussian distribution for improved approximation of target image embeddings, aiming to enhance generation quality and efficiency.

figure 6: we propose corgi, a novel diffusion model designed for flexible text-to-image generation which can "bridge the gap".

The 2023 research paper on "Diffusion Priors for Text-to-Image Generation" by Jingwen Chen, Yingwei Pan, Ting Yao and Tao Mei, discusses To streamline the training of our ControlStyle in an unpaired setting, we introduce two diffusion regularizations aimed at harmonizing the stylized image with both the content and style images in terms of structure and aesthetics. In these diffusion regularizations, features from the upsample blocks within the stable diffusion auto-encoder are leveraged to assess the disparity between the generated image and the content/style image throughout the training process.

Diffusion Models: - Recently, diffusion denoising probabilistic models (DDPM) have revolutionized computer vision, especially in image synthesis. DDPM operates through diffusion and reverse processes, gradually adding and removing Gaussian noise from data. Although computationally demanding, improvements like latent diffusion models (LDM) have made DDPM training more efficient, yielding stunning results. DDPM is now a leading approach in text-to-image synthesis, 3D, and video generation. For instance, stable diffusion combines pre-trained text-image embeddings with DDPM for fast, high-quality image generation. ControlNet further enhances DDPM's capabilities for nuanced text-to-image synthesis.

Analysis and Discussion: - We develop two diffusion regularizations to enhance ControlStyle training in unpaired settings, aligning stylized images with content and style images. By analyzing upsample block features, we identify UpBlock_3 as most informative for structure and utilize a subset of blocks for style transformation to avoid overlearning. Comparing with perceptual loss, our diffusion regularizations produce smoother images with fewer artifacts, benefiting from image priors learned in stable diffusion auto-encoder training.

Result: - This paper introduces a novel task: text-driven stylized image generation, aiming to align images with input text prompts and style images without needing a content image. Our model, ControlStyle, enhances diffusion models for content creation by stylizing pre-trained text-to-image models with a trainable modulation network. We devise two diffusion regularizations to train ControlStyle effectively in unpaired settings. Extensive experiments demonstrate ControlStyle's superiority over existing methods, and we further enhance it with additional controls without retraining, showcasing its practical potential.

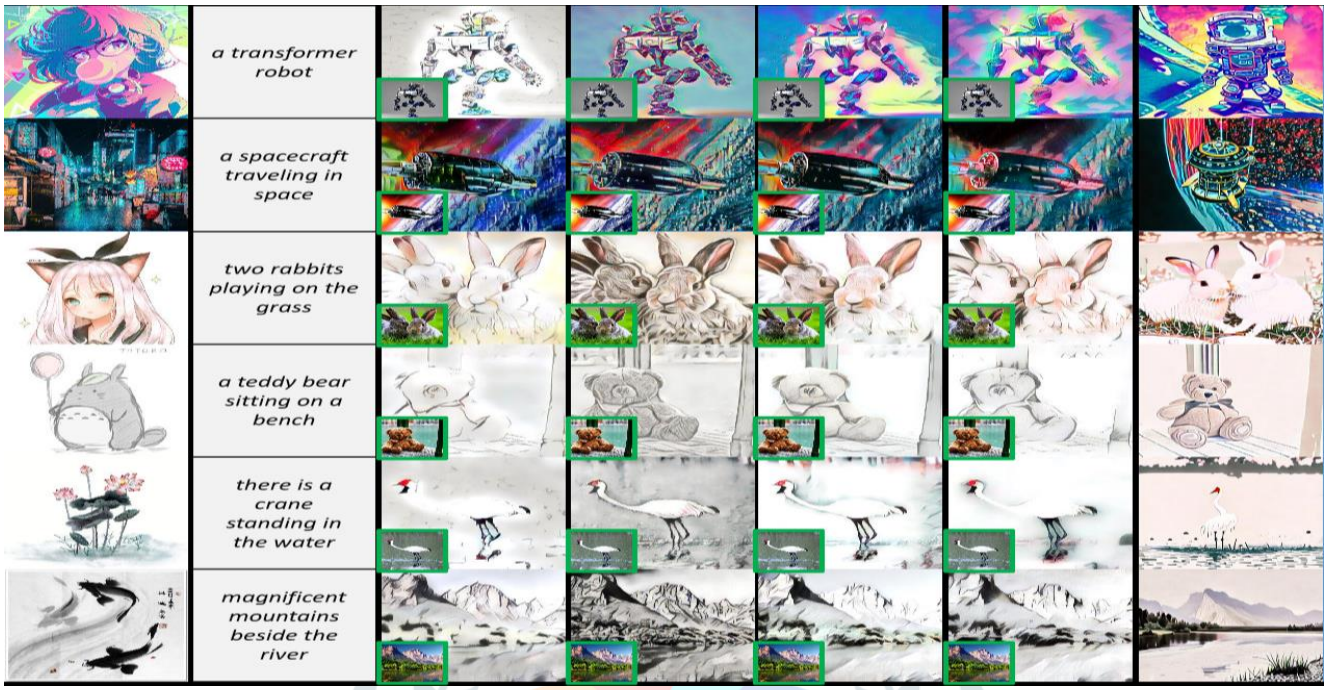


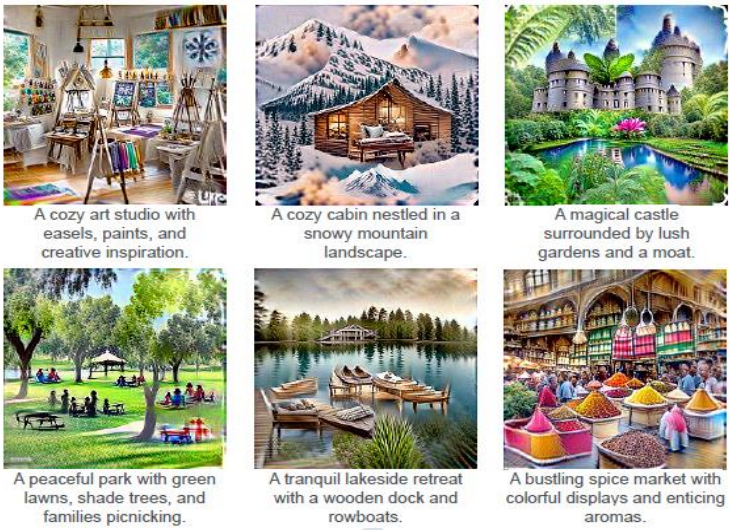
figure 7: examples generated of controlstyle and adain in three styles unseen in the training data (i.e., wikiart): cyberpunk (row 1-2), anime (row 3-4), and chinese ink style (row 5-6).

In 2023, The paper "CLIPAG: Towards Generator-Free Text-to-Image Generation Perceptually Aligned Gradients (PAG) are observed in robust image classification models, where input gradients align with human perception. [8] This phenomenon, previously studied in unimodal vision-only architectures, is extended to Vision-Language architectures in this work. Through adversarial robustification finetuning of CLIP, robust Vision-Language models exhibit PAG compared to vanilla counterparts. CLIP with PAG (CLIPAG) improves various vision-language generative tasks, and its "plug-n-play" integration yields substantial performance gains. Additionally, CLIPAG enables text-to-image generation without large generative models, leveraging its PAG property

Generative Frameworks: - CLIP, known for its strong alignment between vision and language, serves as a crucial component in various text-to-image generative tasks like text-based editing and generation. In this section, we show how CLIPAG, its robust counterpart, can seamlessly replace vanilla CLIP in existing applications. By integrating CLIPAG into frameworks like CLIPDraw and VQGAN+CLIP, we explore its effects on generative capabilities. CLIPDraw, particularly, offers simplicity and lacks a generative model, making it ideal for comparing CLIPAG with standard CLIP. We also demonstrate in Appendix B how CLIPAG enhances explainability beyond generative tasks.

Generative Model: - In this section, we investigate the advantages of CLIPAG (CLIP with Perceptually Aligned Gradients) in different creative tasks, focusing on two primary scenarios: CLIP-driven generative frameworks and text-to-image generation without traditional generators.

Result: -This paper explores Perceptually Aligned Gradients (PAG) in Vision-Language architectures using CLIP. We establish PAG's presence in CLIP through adversarial finetuning, extending its applicability to multimodal models. Integration of CLIPAG



into existing text-to-image frameworks leads to significant improvements, and CLIPAG enables generator-free text-to-image synthesis. These findings underscore the practical implications of leveraging PAG in real-world Vision-Language applications, inspiring further advancements in multimodal research.

figure 8: clipag generator-free text-to-image generation

In the year 2023, the paper "GLIGEN: Open-Set Grounded Text-to-Image Generation" by Yuheng Li1§, Haotian Liu1§, Qingyang Wu2, Fangzhou Mu1, Jianwei Yang3 and Jianfeng Gao3, Chunyuan Li3, Yong Jae Lee1 helps to describe the delve into the advancements in text-to-image models and the challenges faced in controlling spatial composition within generated images. In recent years, there have been significant breakthroughs in the field of generating images from text descriptions. Until recently, Generative Adversarial Networks (GANs) were considered the cutting-edge approach. They were well-known for their ability to manipulate images through a latent space and conditional inputs. Additionally, they were proficient in generating images based on specific text inputs. However, in the past couple of years, text conditional autoregressive and diffusion models have emerged as formidable contenders. These models have showcased remarkable image quality and coverage of various concepts. This success can be attributed to their more stable learning objectives and extensive training on datasets containing pairs of web images and corresponding text descriptions.

Diffusion Models: - Diffusion-based methods are highly effective for text-to-image tasks, with the latent diffusion model (LDM) and Stable Diffusion standing out as powerful models in the research community. LDM reduces computational costs by training in two stages: first, learning a bidirectional mapping network to obtain the latent representation of images, and second, training a diffusion model on the latent space. Focusing on the simplicity of the latent generation space of LDM, we overlook the details of the initial bidirectional mapping.

Results: - Our approach exhibits superior image synthesis quality compared to state-of-the-art baselines, as measured by FID, owing to rich visual knowledge learned during pretraining. Despite comparable FID scores to the LDM* baseline, our model trained with detection annotation instructions (COCO2014D) achieves the best overall performance. However, when evaluated with COCO2014CD instructions, its performance decreases, possibly due to limited understanding of real captions. The model trained with GLIP grounding instructions (COCO2014G) also shows slightly worse FID and low YOLO score, likely because of learning from pseudo-labels. Nonetheless, all grounding instruction types are valuable, suggesting that combining their data can yield complementary benefits. Overall, our model successfully incorporates additional conditions like boxes without compromising image generation quality.

figure 9: gligen enables versatile grounding capabilities for a frozen text-to-image generation model, by feeding different grounding conditions. gligen supports (a) text entity + box, (b) image entity + box, (c) image style and text + box, (d) keypoints, (e) depth map, (f) edge map, (g) normal map, and (h) semantic map.

Table 1: Table summarizing the comparison of the aforementioned research studies

Paper Title	Primary Observations	Methods Utilized	Outcome
VectorFusion: Text-to-SVG by Abstracting Pixel-Based Diffusion Models [1]	- The approach involves optimizing a differentiable vector graphics rasterizer and using a score distillation sampling (SDS) loss to refine shape parameters, resulting in improved fidelity and coherence with the caption.	VectorFusion	VectorFusion is a text-to-vector generative model that utilizes a pretrained diffusion model to generate Scalable vector graphics from text captions, achieving greater quality than CLIP-based approaches and supporting diverse styles such as flat polygonal vector icons, abstract line drawings, and pixel art.
ReCo: Region-Controlled Text-to-Image Generation [2]	- ReCo's implementation is based on Stable Diffusion and extends the text tokens with an extra vocabulary specialized for spatial coordinate referring. - The model significantly improves image generation quality and controllability, demonstrating its potential in enhancing image generation controllability and quality over a wide range of datasets and designed prompts.	ReCo	The ReCo model extends pre-trained T2I models to understand spatial coordinate inputs, allowing precise regional control for arbitrary objects described by open-ended regional texts, significantly improving image generation quality and controllability.
GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models [3]	- The model is capable of performing image inpainting, enabling realistic edits to existing images using natural language prompts. - The model achieves high aesthetic quality and competitive performance in generating realistic images from text prompts, demonstrating its potential for diverse image generation and editing.	GLIDE	The model, GLIDE, demonstrates high performance in generating photorealistic images and editing them based on text prompts. It achieves competitive results in human evaluations, produces diverse and realistic images, and effectively performs image inpainting, showcasing its potential for advanced image generation and editing capabilities.
RAPHAEL Text-to-Image Generation via Large Admixture of Diffusion Paths [4]	- RAPHAEL prolixity model surpasses former models with admixture-of-experts approach - Outperforms Stable prolixity XL, Deep-Floyd, DALL-E 2, and ERNIE-ViLG2.0 - Achieves state-of-the-art FID-30k score of 6.61 on COCO dataset	RAPHAEL	RAPHAEL prolixity model stands out with its admixture-of-experts approach. Achieves state-of-the-art FID-30k score of 6.61 on COCO dataset. Outperforms competitors in image-text alignment and quality based on ViLG-300 standards
Adding Conditional Control to Text-to-Image Diffusion Models [5]	-Focuses on adding conditional control to text-to-image diffusion models to address challenges in spatial composition control. - ControlNet, an end-to-end neural network architecture, facilitates efficient fine-tuning of large pre-trained text-to-image diffusion models by learning conditional controls. - ControlNet's approach enhances spatial composition	Conditional Control	ControlNet provides efficient fine-tuning of large pre-trained models, showcasing potential for revolutionizing image generation processes. Stable Diffusion and ControlNet integration improve training stability, efficiency, and enable personalized image generation.

	control in text-to-image models, showcasing potential for revolutionizing image generation processes.		
Shifted Diffusion for Text-to-Image Generation [6]	- Corgi bridges the image-text modality gap and data availability gap, enabling semi-supervised and language-free text-to-image generation. - The framework discussed includes a pre-trained image encoder, image generation decoder, and a prior model for generating image embeddings from text captions.	Shifted Diffusion	The shifted diffusion model leverages prior knowledge from CLIP to enhance generation efficiency and quality. Emphasis is placed on the prior model's enhancement, enabling semi-supervised training and supporting various generation tasks effectively.
Text-Driven Stylized Image Generation Using Diffusion Priors [7]	-A novel approach to text-driven stylized image generation. ControlStyle uses a diffusion model with a trainable modulation network to produce high-quality stylized images that are both semantically relevant to the input text prompt and aligned with a given style image, surpassing existing techniques.	Diffusion Priors	-A novel method for text-driven stylized image generation. It combines input text prompts and style images to create visually appealing images that are both semantically relevant and aligned with the specified style. ControlStyle utilizes diffusion models and a trainable modulation network to achieve high-quality results, surpassing traditional methods.
CLIPAG: Towards Generator-Free Text-to-Image Generation [8]	-The Perceptually Aligned Gradients (PAG) phenomenon from robust image classification models to Vision-Language architectures. The study shows that robust Vision-Language models, such as CLIPAG, exhibit PAG, leading to substantial improvements in vision-language generative tasks and enabling text-to-image generation without large generative models.	CLIP	-Perceptually Aligned Gradients (PAG) from image classification to Vision-Language architectures. By robustifying CLIP, it's shown that robust Vision-Language models exhibit PAG, enhancing semantic alignment of input gradients. This CLIP with PAG (CLIPAG) extension improves vision-language generative tasks and enables text-to-image generation without large generative models.
GLIGEN: Open-Set Grounded Text-to-Image Generation [9]	-GLIGEN is a novel approach for text-to-image generation that enhances existing large-scale models by allowing them to be conditioned on grounding inputs alongside text. By injecting grounding information into new trainable layers, GLIGEN achieves open-world grounded text-to-image generation with strong zero-shot performance on datasets like COCO and LVIS, outperforming existing baselines.	GLIGEN	-A novel approach for text-to-image generation. GLIGEN enhances existing models by incorporating grounding inputs alongside text inputs, achieved through a gated mechanism. This enables open-world grounded text-to-image generation with strong zero-shot performance, outperforming existing baselines on datasets like COCO and LVIS.

III. CONCLUSION

Each of the mentioned papers contributes uniquely to the field of text-to-image generation, offering distinct approaches and innovations. "VectorFusion" stands out for its emphasis on converting text descriptions into scalable vector graphics (SVG), providing advantages in scalability, resolution independence, and interpretability. In contrast, "ReCo" focuses on region-controlled text-to-image generation, offering fine-grained control over specific regions of generated images, enabling users to dictate desired details or attributes within those regions.

"GLIDE" distinguishes itself by aiming for photorealistic image generation and editing guided by text descriptions, achieving highly detailed and realistic results suitable for applications requiring high-fidelity visual content. Meanwhile, "RAPHAEL" introduces text-to-image generation via a large mixture of diffusion paths, enhancing diversity and richness in generated images, offering versatility in the generated image corpus.

"Adding Conditional Control to Text-to-Image Diffusion Models" extends text-to-image diffusion models with conditional control, allowing users to influence specific attributes or characteristics of generated images with fine-grained precision. "Shifted Diffusion for Text-to-image Generation" improves image quality and diversity by leveraging shifted diffusion models, leading to visually appealing and diverse results compared to traditional diffusion models.

"ControlStyle" enables text-driven stylized image generation using diffusion priors, providing users with control over stylistic aspects of generated images, facilitating the creation of visually appealing and stylized content based on textual descriptions. "CLIPAG" introduces a generator-free approach to text-to-image generation, leveraging CLIP with Perceptually Aligned Gradients (PAG) to generate images directly from text descriptions, offering simplicity and efficiency in the generation process.

Lastly, "GLIGEN" focuses on open-set grounded text-to-image generation, enabling the generation of images from text descriptions while incorporating grounding conditions, thus offering versatility and robustness in image generation tasks. Each of these papers collectively advances the state-of-the-art in text-to-image synthesis, catering to a wide range of applications and requirements within the field.

REFERENCES

[1] Jain A, Xie A, Abbeel P. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 1911-1920).

[2] Yang Z, Wang J, Gan Z, Li L, Lin K, Wu C, Duan N, Liu Z, Liu C, Zeng M, Wang L. Reco: Region-controlled text-to-image generation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 14246-14255).

[3] Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741. 2021 Dec 20.

[4] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, Ping Luo. RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths.37th Conference on Neural Information Processing Systems (NeurIPS 2023)

[5] Lvmin Zhang, Anyi Rao, Maneesh Agrawala; Adding Conditional Control to Text-to-Image Diffusion Models. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3836-3847

[6] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, Jinhui Xu; Shifted Diffusion for Text-to-Image Generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10157-10166 https://openaccess.thecvf.com/content/CVPR2023/papers/Zhou_Shifted_Diffusion_for_Text-to-Image_Generation_CVPR_2023_paper.pdf

[7] Chen J, Pan Y, Yao T, Mei T. Controlstyle: Text-driven stylized image generation using diffusion priors. InProceedings of the 31st ACM International Conference on Multimedia 2023 Oct 26 (pp. 7540-7548).

[8] Ganz R, Elad M. Clipag: Towards generator-free text-to-image generation. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2024 (pp. 3843-3853).

[9] Li Y, Liu H, Wu Q, Mu F, Yang J, Gao J, Li C, Lee YJ. Gligen: Open-set grounded text-to-image generation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 22511-22521).