Unlocking Creativity: Exploring Advanced AI Techniques For Visual Content Generation In Collaborative Platforms

Guide: Prof. Dr.Swapna Bhavsar
Swapnil Devkate
Karan Suryawanshi
Omkar Thete
Pratik Pawar
Department of Information Technology
PES Modern College of Engineering Pune

Abstract:
Text-to-Image and Text-to-Video AI generation models represent groundbreaking technologies that leverage deep learning and natural language processing (NLP) to produce images and videos from textual descriptions. This paper delves into the most advanced methods in the realms of Text-to-Image and Text-to-Video AI generation. It offers a comprehensive survey of the current literature and analyzes various approaches applied in multiple studies. This includes data preprocessing techniques, types of neural networks, and the evaluation metrics utilized in this field. Moreover, the paper explores the challenges and limitations associated with Text-to-Image and Text-to-Video AI generation and outlines potential directions for future research. These models hold significant promise for applications across diverse domains such as video production, content creation, and digital marketing.

Keywords:
artificial intelligence, deep learning, natural language processing (NLP), large language model, AI text-to-image generation, AI text-to-video generation, DALL-E, CogView, Imagen, NUWA, Phenaki, GODIVA.

I. INTRODUCTION

Recent advancements in deep learning and natural language processing (NLP) have led to the emergence of AI text-to-image and AI text-to-video generators, which are powerful tools for generating images and videos from text descriptions. These AI generators utilize advanced and intricate techniques such as attention-based Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), and transformers to analyze textual data and produce corresponding high-quality visuals.

The impetus behind developing AI text-to-image and AI text-to-video generators is the need to automate content creation, making it more efficient and cost-effective. These systems have potential applications across various fields, including marketing, education, and entertainment. For instance, in marketing, AI text-to-image generators can create product designs, catalogs, and user manuals. In education, AI text-to-video generators can develop instructional videos and animations to enhance the learning experience. In the entertainment industry, these generators can produce movie promotional videos, teasers, and more, thereby enhancing user engagement and improving user experiences.

However, the rapid development of AI text-to-image and AI text-to-video generators presents several challenges. One major challenge is the need for extensive, high-quality training datasets, which can be difficult to obtain and label. Another challenge is the lack of interpretability in the generated outputs, making it hard to understand the reasoning behind the visual content. Additionally, these systems may not always align with the intended message, leading to errors and conflicts in the generated output. There is also a trade-off between visual quality and processing time, as generating high-quality images and videos can be computationally expensive and slow. Moreover, the content produced may not always adhere to social or public norms, leading to misinterpretation or misrepresentation of the intended message. These limitations and challenges must be carefully considered when implementing AI text-to-image and text-to-video generators.

This paper aims to provide an overview of the current state-of-the-art techniques in AI text-to-image and text-to-video generation. It examines the underlying technologies, such as data preprocessing techniques, types of neural networks, and the evaluation metrics

used. The structure of the paper is as follows: Section II reviews AI text-to-image generators, including popular techniques and their capabilities. Section III explores popular AI text-to-video generators and compares their capabilities. Section IV analyzes the current state-of-the-art in both AI text-to-image and AI text-to-video generation. Finally, Section V concludes the paper.

II. AI TEXT-TO-IMAGE GENERATORS

AI text-to-image generators are powerful tools that integrate natural language processing and computer vision to create images from text descriptions. Here, we discuss some prominent state-of-the-art text-to-image generators, such as CogView2, DALL-E 2, and Imagen. Table 1 provides a comparison of these systems.

A.CogView2

CogView2 is an AI text-to-image generator that employs a hierarchical transformer-based approach to produce images from text descriptions. It utilizes the Cross-Modal General Language Model (CogLM), a pre-trained 6B-parameter transformer that performs self-supervised tasks to mask and predict various token types in a text and image token sequence. The hierarchical design of CogView2 allows for rapid and efficient generation of high-resolution images by initially generating low-resolution images and refining them through an iterative super-resolution module that uses local parallel autoregressive generation. CogView2 is ten times faster than its predecessor, CogView, which used sliding-window super-resolution for generating images of similar resolution and quality. Additionally, CogView2 supports interactive text-guided image editing.

TABLE I. AI TEXT-TO-IMAGE GENERATORS AND THEIR COMPARISON

	Architecture	Training Data	Image Quality	Computational Requirements	Interpretability	Novelty
CogView2	Hierarchical transformer based	Large-scale text only datasets	High-resolution, better quality	Relatively efficient	Difficult to interpret	Fast and efficient generation
DALL-E 2	Large-scale transformer language model with StyleGAN2 architecture	Large dataset of image-text pairs	High-quality and diverse	High computational cost	Easy to interpret	Complex and diverse textual prompts
Imagen	Large-scale frozen T5-XXL encoder and diffusion models	Large text-only corpora	High-quality and photorealistic	Relatively efficient	Easy to interpret	Leveraging existing language models for image generation

B.DALL-E2

DALL-E 2, developed by OpenAI, builds on the success of the original DALL-E model. It generates high-resolution (1024x1024) images from textual input by training a large transformer model with 175B parameters, making it the largest language model trained to date. Unlike the original DALL-E, which used a simple VQVAE architecture, DALL-E 2 employs a more powerful StyleGAN2 architecture capable of producing realistic and diverse images. It can handle complex and diverse textual prompts and generate a wide range of objects and scenes. DALL-E 2's training involves a multi-stage process that combines pretraining on a large text corpus with fine-tuning on an image-text dataset. During inference, given a textual prompt, DALL-E 2 generates a sequence of image tokens autoregressively, with each token representing a patch of the final image. These tokens are then processed through the StyleGAN2 generator to produce the final high-resolution image.

C.Imagen

Google's Imagen combines the strengths of large transformer language models and diffusion models to generate high-quality images. It uses a large frozen T5-XXL encoder to encode the input text into embeddings, which a conditional diffusion model then maps into a 64x64 image. Imagen employs text-conditional super-resolution diffusion models to upscale the image from 64x64 to 256x256 and finally to 1024x1024.

Imagen achieves state-of-the-art results in terms of FID score and image-text alignment on the COCO dataset, outperforming recent methods in side-by-side comparisons on the comprehensive DrawBench benchmark for text-to-image models.

In summary, popular AI text-to-image generators like CogView2, DALL-E 2, and Imagen employ various methods to produce images from text input. CogView2's hierarchical transformer-based method facilitates fast

and efficient generation of high-resolution images. DALL-E 2 uses a large transformer language model and a robust StyleGAN2 architecture to produce a wide range of lifelike visuals. Imagen leverages the power of large transformer language models and diffusion models to

create high-quality images. All three models excel in generating diverse, high-quality images closely aligned with the input text, showcasing impressive results.

Video Generator	Model Type	Advantages	Limitations
Make-A-Video	2I (Text-to-Image) models and unsupervised learning on unlabelled video data	Accelerated training, Unsupervised learning, Inheritance of image generation models.	Cannot learn associations between text and certain phenomena in videos.
Imagen Video	cascade of video diffusion models	High fidelity, Diverse video generation, 3D objects understanding, Text animations	Trained on problematic data [18][19][20], social biases, and stereotypes.
Phenaki	Encoder-decoder model with a transformer	Good performance on video prediction, can generate long videos conditioned on text and starting frame	Trained on biased datasets.
GODIVA	Text-to-video pretrained model with three-dimensional sparse attention mechanism	Reduced computation cost, Good zero-shot capability	Challenge to generate long videos with high resolution, evaluating text-to-video generation task remains a challenge.
CogVideo	Inherits pretrained text-to-image model CogView2	Efficiently leverages image generation capacity, better understanding of text-video relations	Restriction on input sequence length, large scale model and limitation of GPU memory.
NUWA	Multimodal pretrained model with 3D transformer encoder-decoder framework	Reduces computational complexity, Good zero-shot capabilities	Poor text-video alignment in frames.

III. AI TEXT-TO-VIDEO GENERATORS

AI text-to-video generators have garnered significant interest due to their potential to revolutionize the video production industry. These generators enable users to create highly personalized and engaging video content quickly and easily. They utilize advancements in deep learning and natural language processing to generate videos from text descriptions. While early AI text-to-video generators were limited in the quality and variety of videos they could produce, recent advancements have shown promising results in creating highly realistic videos. However, challenges such as maintaining video coherence and the high computational requirements remain.

This section discusses state-of-the-art AI text-to-video generators, including Make-A-Video, Imagen Video, Phenaki, GODIVA, and CogVideo, highlighting their strengths, limitations, and potential applications. Table 2 provides a comparison of these models.

A.Make-A-Video

Make-A-Video is an innovative approach that extends a diffusion-based text-to-image model to text-to-video generation through a spatiotemporally factorized diffusion model. By leveraging joint text-image priors, this method eliminates the need for paired text-video data, allowing it to scale to larger video datasets. Make-A-Video introduces super-resolution strategies in both spatial and temporal dimensions, generating high-definition, high frame-rate videos from user-provided textual input. It is thoroughly evaluated against existing T2V systems, demonstrating state-of-the-art results in both quantitative and qualitative measures.

B.ImagenVideo

Imagen Video uses a frozen T5 text encoder, a base video diffusion model, and interleaved spatial and temporal super-resolution diffusion models to produce high-quality videos. The system can generate 128-frame, 1280x768 high-definition videos at 24 frames per second. It

offers high controllability and world knowledge, enabling the generation of diverse videos and text animations in various artistic styles with 3D object understanding. The system's design decisions, such as the use of fully-convolutional temporal and spatial super-resolution models and the v-parameterization of diffusion models, contribute to its successful performance.

C.Phenaki

Phenaki is a lightweight model from Google that generates videos from short text inputs. It is limited to simple actions and movements and lacks fine-grained details. Phenaki's encoder-decoder architecture, called C-ViViT, compresses videos to discrete embeddings (tokens) and exploits temporal redundancy to improve reconstruction quality while reducing the number of video tokens. The model uses a transformer to translate text embeddings generated by a pretrained language model (T5X) into video tokens. Phenaki demonstrates the ability to generalize beyond what is available in

video datasets, generating long, temporally coherent, and diverse videos conditioned on open-domain prompts or sequences of prompts that tell a story.

D.GODIVA

GODIVA utilizes a transformer encoder-decoder architecture, with the encoder transforming text into text embeddings and the decoder combining text embeddings with visual tokens to autoregressively generate visual tokens. These tokens represent consecutive video frames, which are decoded into individual frames. Each frame is processed into final video frames by a pretrained VQVAE model. GODIVA achieves high-quality videos and can handle various types of textual input. However, the model requires high computational resources and is sensitive to the quality of text descriptions.

E.CogVideo

CogVideo, developed by Tsinghua University, is a transformer model capable of generating videos from textual input. It utilizes a pre-trained VQVAE model to encode video frames into latent representations and a transformer model to generate these representations from text input. CogVideo's ability to generate high-quality videos with consistent frame transitions and detailed visuals makes it a powerful tool for video creation. However, the model's reliance on large amounts of computational power and the challenge of generating videos that accurately match complex textual descriptions remain areas for improvement.

In conclusion, AI text-to-video generators such as Make-A-Video, Imagen Video, Phenaki, GODIVA, and CogVideo represent significant advancements in this field. These generators leverage sophisticated techniques, including diffusion models, super-resolution strategies, and transformer architectures, to create high-quality, diverse, and temporally coherent videos from textual descriptions. Despite the challenges, such as maintaining video coherence and high computational requirements, these models demonstrate the potential of AI in revolutionizing the video production industry by enabling the rapid and efficient creation of personalized and engaging video content.

IV. STATE-OF-THE-ART IN AI TEXT-TO-IMAGE AND TEXT-TO-VIDEO GENERATION

AI text-to-image and text-to-video generation are two closely related fields that have witnessed significant advancements in recent years. The state-of-the-art techniques in these fields leverage deep learning, natural language processing, and computer vision to produce high-quality visual content from textual descriptions.

A. State-of-the-Art in AI Text-to-Image Generation

The current state-of-the-art in AI text-to-image generation is characterized by the use of powerful neural networks, such as transformers and GANs, to generate high-resolution and diverse images from textual input. Techniques like hierarchical transformers (used by CogView2), StyleGAN2 (used by DALL-E 2), and diffusion models (used by Imagen) have proven to be highly effective in creating high-quality images. These models utilize large-scale datasets and complex architectures to understand the nuances of textual descriptions and produce images that closely match the input text. Additionally, these models are evaluated using metrics such as FID score, image-text alignment, and user satisfaction to ensure the quality and relevance of the generated images.

B. State-of-the-Art in AI Text-to-Video Generation

The field of AI text-to-video generation has also seen remarkable progress, with models like Make-A-Video, Imagen Video, Phenaki, GODIVA, and CogVideo pushing the boundaries of what is possible. These models employ advanced techniques, such as spatiotemporal diffusion models, super-resolution strategies, and transformer architectures, to generate high-quality, coherent, and engaging videos from textual descriptions. The evaluation of these models involves quantitative metrics like FID score and qualitative assessments of video coherence, quality, and user satisfaction. The ability to create videos with complex scenes, diverse actions, and fine-grained details is a testament to the advancements in this field.

V. CONCLUSION

AI text-to-image and text-to-video generators represent significant advancements in the field of artificial intelligence, with the potential to revolutionize content creation across various industries. These generators leverage advanced techniques in deep learning, natural language processing, and computer vision to produce high-quality visual content from textual descriptions.

Despite the challenges and limitations, such as the need for large-scale datasets, high computational requirements, and the complexity of aligning generated content with intended messages, these models demonstrate the potential to automate and enhance content creation processes.

The state-of-the-art techniques in AI text-to-image generation, such as hierarchical transformers, StyleGAN2, and diffusion models, enable the creation of high-resolution and diverse images from textual input. Similarly, the advanced methods in AI text-to-video generation, including spatiotemporal diffusion models, super-resolution strategies, and transformer architectures, allow for the production of high-quality, coherent, and engaging videos. The evaluation of these models using quantitative and qualitative metrics ensures the quality and relevance of the generated visual content.

Future research in this field should focus on addressing the challenges and limitations of AI text-to-image and text-to-video generation, such as improving the interpretability of generated outputs, reducing computational requirements, and ensuring the alignment of generated content with intended messages. Additionally, exploring new applications and use cases for these technologies can further enhance their impact and utility across various industries.

REFERENCES:

[1] T. Zia, S. Arif, S. Murtaza, and M. A. [2] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016, pp. 1060-1069

[3] N. A. Fotedar and J. H.

[4] H. Chang, H. Zhang, J. Barber, A. J. Maschinot, J. Lezama, L. Jiang, M. -H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, and D. Krishnan, "Muse: Text-To-Image Generation via Masked Generative Transformers," arXiv preprint arXiv:2301.00704, 2023.

[5] A. Haleem, M. Javaid, M. A. Qadri, R. P. Singh, and R. Suman, "Artificial intelligence (AI) applications for marketing: A literaturebased study," International Journal of Intelligent Networks, vol. 3, pp. 119-132, 2022. doi: 10.1016/j.ijin.2022.08.005.

. [9] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, and J.

[10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," in arXiv preprint arXiv:2202.10775, 202

[11] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. Seyed Ghasemipour, B. Karagol Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M.

[12] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv:2209.14792 [cs.CV], Sep. 2022.

[13] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. 2022. [Online]. Available: https://arxiv.org/abs/2210.02303. [14] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D

[15] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "CogVideo: Largescale Pretraining for Text-to-Video Generation via Transformers," arXiv:2205.15868 [cs.CV], May 2022.

[17] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N.

[18] B. Bordia and S. R. Bowman, "Identifying and Reducing Gender Bias in Word-Level Language Models," arXiv:1904.03035 [cs.CL], 2019.

[20] E. M. Bender, T. Gebru, A. McMillan-Major, and S.