International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 3, pp 67-77, March-2025

https://doi.org/10.47001/IRJIET/2025.903009

Text-to-Image Datasets: Characteristics, Challenges, and Opportunities

[1]Haitham ALHAJI, [2]Alaa Yaseen Taqa

[1]Computer Science Department, College of Computer Science and Mathematics, University of Mosul, Nineveh, Iraq
[2]Computer Science Department, College of Education for Pure Science, University of Mosul, Nineveh, Iraq

***Abstract* - Text-to-image synthesis is an intriguing field of study that seeks to create visuals from textual descriptions. The primary objective of this domain is to provide visuals that align with the provided written description for both semantic coherence and visual reality. Despite significant advancements in text-to-image synthesis in recent years, it continues to encounter numerous hurdles, primarily concerning picture realism and semantic coherence. To address these challenges, selecting diverse datasets with comprehensive annotations will markedly improve model performance in addressing these difficulties. Datasets with varied visual material and comprehensive textual descriptions aid models in understanding intricate links between text and images, enhancing both semantic coherence and image authenticity. This review paper examines 20 datasets available for text-to-image synthesis, categorizing them by scope, variety, and application domains. The meticulous selection and curation of datasets are crucial for enhancing text-to-image synthesis technology. Ultimately, the careful selection and curation of datasets play a pivotal role in advancing the state-of-the-art in text-to-image synthesis.**

Keywords: Text-to-Image Datasets, Dataset Diversity, Dataset Limitations, Scene Complexity, Generative AI Datasets.

I. INTRODUCTION

Text-to-image synthesis is an emerging field that seeks to generate images from textual descriptions [1]. Text-to-image generation (T2I) in computing combines natural language processing and computer vision, signifying the production of realistic images from textual descriptions via specialized models [2]. T2I has garnered considerable attention due to its potential to transform content creation across multiple areas. This technology connects language and imagery by transforming verbal descriptions into visual content, creating new opportunities for art, design, and multimedia applications. Furthermore, T2I is pivotal in AI-generated content and signifies a significant milestone in pursuing general artificial intelligence [3], with modern methods such as [4].

Advancements in deep learning [5] have enabled T2I models to produce increasingly sophisticated images that accurately correspond to textual descriptions [6]. The advent of Generative Adversarial Networks (GANs) [7] facilitated substantial advancements in Text-to-Image (T2I) technology [8]. Motivated by the conditional GAN (cGAN) [9], they created the GAN-CLS and GNA-INT models, which were the first implementations of GAN in text-to-image (T2I) tasks, demonstrating the advantages of GAN in generating high-quality, detail-rich images.

OpenAI introduced the DALL-E model [10] utilizing the Transformer architecture [11] from natural language processing. This model was the inaugural implementation of an autoregressive (AR) [12] approach for image synthesis, leveraging extensive datasets to create varied visuals. Despite the AR technique exhibiting superior generation capabilities, its considerable computational expense restricts its applicability in specific application contexts [13].

In recent years, diffusion models (DM) [14] derived from nonequilibrium thermodynamics have progressively emerged as the most sophisticated methodology in the T2I domain. GLIDE [15] was the inaugural project to implement diffusion models for T2I tasks, showcasing remarkable generative proficiency by functioning within pixel space. The advent of the Latent Diffusion Model (LDM) [14] underscored the importance of latent space in diffusion models significantly improves the quality of generated images. With the progression of study, diffusion models have demonstrated unparalleled effectiveness, positioning them as a primary emphasis in T2I investigations [16].

The continuous advancement of T2I technology has sparked intense discussions within the community [17]. Numerous research teams regularly disseminate new publications, and the swift advancement of technology presents considerable obstacles for newcomers attempting to initiate and maintain relevance [18]. Recent research has investigated the advancement of GANs in text-to-image (T2I) applications, the evolution of diffusion models, and pertinent studies on controlled generation modules [3]. Nonetheless, thorough evaluations of the most recent trends in T2I are still scarce [17].

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Issue 3, pp 67-77, March-2025
https://doi.org/10.47001/IRJIET/2025.903009

Table 1: Comprehensive Overview of Datasets Used in Text-to-Image

Year	Dataset	Category	Images	Resolution	Annotations	Attributes	Public
2008	Oxford-102 Flowers [19]	Flower	8,189	variety	10	-	✓
2011	CUB-200-2011 [20]	Bird	11,788 (-)	variety	10	-	✓
2012	MNIST [21]	Numbers	60,000	28x28	1		✓
2014	MS-COCO2014 [22]	Iconic Objects	200k labeled, 330k non	variety	5	-	✓
2017	visual genome [23]	Objects	108,077	variety	~5	18	✓
2019	SCU-Text2face [24]	Face	1,000	256×256	5	-	×
2020	Multi-Modal CelebA-HQ [25]	Face	30,000	512×512	10	38	✓
2021	FFHQ-Text [26]	Face	760	1024×1024	9	162	✓
2021	M2C-Fashion [27]	Clothing	10,855,753	256×256	1	-	×
2021	CelebA-Dialog [28]	Face	202,599	178×218	~5	5	✓
2021	Faces a la Carte [29]	Face	202,599	178×218	~10	40	×
2021	LAION-400M [30]	Random, Crawled	400M	variety	1	-	✓
2022	Face2Text [31]	Face	~11,000	variety	1~	-	✓
2022	Bento800 [32]	Food	800	600×600	9	-	✓
2022	LAION-5B [33]	Random Crawled	5.85B	variety	1	-	✓
2022	DiffusionDB [34]	Synthetic Images	14M	variety	1	-	✓
2022	COYO-700M [35]	Random Crawled	747M	variety	1	16	✓
2022	DeepFashion[36]	Full Body	44,096	750×1101	1	-	✓
2023	ANNA [37]	News	29,625	256×256	1	-	✓
2023	DreamBooth [38]	Objects & Pets	158	variety	25	-	✓

This work aims to elucidate the fundamental principles of GAN, AR, and DM models while comprehensively evaluating their recent advancements in T2I. This paper will thoroughly examine innovative research trajectories in T2I, offering researchers a definitive roadmap and essential references for further investigation.

The efficacy of text-to-image generative models depends fundamentally on the presence of high-quality, diversified datasets that capture the complex interplay between written descriptions and visual information [2]. These datasets provide the basis for training models to produce images that are both visually realistic and semantically congruent with their respective descriptions [39]. The selection of datasets, ranging from detailed collections like CUB-200-2011 and Oxford-102 Flowers to extensive, general-purpose corpora such as MS COCO and Flickr30k, significantly impacts model performance, affecting its capacity to manage complexity, diversity, and domain-specific subtleties. The advancement of the field necessitates the creation of specialized datasets that tackle emerging challenges, including ethical considerations, multilingual support, and under-represented domains, thereby highlighting the crucial importance of datasets in fostering innovation and ensuring inclusivity in text-to-image synthesis systems [40]. Most models are based on the moderation of new software maintenance [41].

II. TYPES OF DATASETS

A dataset is the fundamental component of every machine learning task. This section offers a summary of the conventional text-image datasets commonly employed to evaluate text-to-image synthesis techniques. The datasets are essential for assessing the models' capacity to generate realistic pictures that correspond with the textual descriptions. NLP [42] models learn primarily from textual data; in contrast, CV models learn from information based on images. Using both text and images, vision-language pre-trained models combine the capabilities of NLP and CV.

Human experts must precisely annotate the data to train complex models. Whenever large datasets or domain-specific data are handled, this process uses many resources and money. Table 1 provides an extensive list of significant datasets used in text-to-image generation research, ranging from small to large. Table 2 delineates essential datasets and their corresponding features.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 3, pp 67-77, March-2025

https://doi.org/10.47001/IRJIET/2025.903009

Table 2: Feature Overview of Text-to-Image Datasets

Feature	Dataset
Bounding Box	CUB-200-2011 [20], MS-COCO2014 [22], visual genome [23], FFHQ-Text [26], Bento800 [32]
Segmentation	CUB-200-2011 [20], MS-COCO2014 [22], Bento800 [32], LAION-5B [33]
Keypoint	CUB-200-2011 [20], DeepFashion-MultiModal [36]
Identity Label	CelebA-Dialog [28],
URL	LAION-5B [33], COYO-700M [35]
Masks	Multi-Modal CelebA-HQ [25]
Sketches	Multi-Modal CelebA-HQ [25]
Label	Bento800 [32]

The Oxford-102 Flowers dataset [19] offers a compelling challenge for researchers exploring text-to-image generation. This dataset features 102 distinct flower categories, each containing various images showcasing variations in scale, pose, and lighting conditions, as illustrated in Figure 1. Notably, the included flowers are commonly found in the UK, potentially introducing cultural and regional biases that necessitate careful consideration during model development. Additionally, the dataset harbors inherent ambiguity within specific categories and presents several visually similar flower types. These characteristics push text-to-image models beyond essential object generation, demanding a deeper understanding of subtle visual nuances and the ability to translate textual descriptions into accurate and visually distinct floral representations. Figure 1 shows.

Figure 1: The Oxford-102 Flowers dataset

The CUB-200-2011 dataset [20], known as Caltech-UCSD Birds-200-2011, reigns supreme in fine-grained visual categorization tasks. This goldmine holds 11,788 images meticulously categorized into 200 distinct bird subcategories, as illustrated in Figure 2. Each image boasts rich annotations, including a subcategory label, 15-part locations (think beak, wing, tail), 312 binary attributes (e.g., migratory, ground-nesting), and a bounding box. But that's not all! The dataset also features textual descriptions for each image, offering a valuable bridge between visual and semantic information. This comprehensive nature makes CUB-200-2011 a popular choice for training and evaluating image classification models, particularly those tackling distinguishing subtle differences within specific categories, like differentiating between various sparrow species.

Figure 2: The CUB-200-2011 dataset

The MS-COCO2014 dataset [21] is a goldmine for computer vision and image visualizing researchers. Over 330,000 images, more than 200,000 meticulously labeled, capture the essence of everyday scenes, as illustrated in Figure 3.

Figure 3: The MS-COCO2014 dataset

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 3, pp 67-77, March-2025

https://doi.org/10.47001/IRJIET/2025.903009

Each image is a snapshot of reality, bursting with information about the objects it holds. Not just their presence but their locations and attributes, a detailed picture painted for the algorithms to decipher. This treasure trove fuels the development of cutting-edge object detection algorithms, pushing the boundaries of computer vision. But MS-COCO2014 goes beyond mere detection. It delves into the relationships between objects, weaving a tapestry of understanding that machines can learn from.

The Visual Genome [23] is a large-scale dataset of images with rich annotations, including object bounding boxes, attributes, and relationships, as illustrated in Figure 4. It was created by a team of researchers at Stanford University and the University of California, Berkeley. The dataset comprises more than 100,000 photos, each tagged with an average of 21 items, 18 attributes, and 18 pairwise associations among objects. The annotations were gathered using a crowd sourcing tool and meticulously filtered to guarantee quality.

Figure 4: The Visual Genome dataset

The SCU-Text2face dataset comprises 1,000 images from the CelebA dataset, each paired with five distinct human-written descriptions, as illustrated in Figure 5; SCU-Text2face captures diverse facial attributes, expressions, and even emotional nuances [24]. This comprehensive annotation allows researchers to evaluate the effectiveness of text-to-face models in capturing the intricate details of human appearance as described in natural language. Therefore, SCU-Text2face plays a vital role in advancing the field of text-to-face generation, offering a standardized and well-annotated dataset for training and evaluating the performance of new models.

Figure 5: The SCU-Text2face dataset

Expanding upon the high-resolution images of CelebA-HQ, Multi-Modal CelebA-HQ offers a rich 30,000-image dataset for facial research [25]. Each image is accompanied by a segmentation mask, as used in many other applications such as [43], for precise feature analysis, a sketch for structure understanding, a descriptive text capturing attributes and emotions, and a transparent background for flexible manipulation. This diverse data empowers researchers in tasks like text-to-image generation, as illustrated in Figure 6, text-guided manipulation, sketch-to-image generation, image captioning, and visual question answering, solidifying its position as a valuable tool for advancing research in understanding and manipulating human faces.

Figure 6: Multi-Modal CelebA-HQ Dataset

The M2C-Fashion dataset is a valuable resource for researchers and developers in computer vision and artificial intelligence, particularly those interested in fashion image generation and editing [27]. This large-scale dataset provides a rich collection of clothing images categorized according to various attributes, as illustrated in Figure 7. Each image is

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Issue 3, pp 67-77, March-2025
https://doi.org/10.47001/IRJIET/2025.903009

accompanied by detailed annotations, including style codes, allowing for fine-grained analysis and manipulation.

Figure 7: The M2C-Fashion dataset

The CelebA-Dialog dataset [28], with over 200,000 richly annotated facial images, empowers researchers in fine-grained facial editing, as illustrated in Figure 8. Each image features detailed labels, captions, and user editing requests, enabling models to translate text descriptions into corresponding facial edits. This unique dataset bridges the gap between NLP, which is used in many fields such as [44], and CV, offering exciting possibilities for photo editing tools, personalized avatars, and realistic image generation.

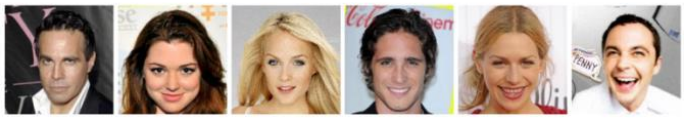

Figure 8: The CelebA-Dialog dataset

The Faces a la Carte dataset tackles the challenge of text-to-face generation. Unlike traditional single-label datasets, it features faces paired with detailed multi-label textual descriptions encompassing diverse facial attributes [29]. This empowers researchers to train models that generate faces based on individual features and capture the nuances of human appearance described in the text, as illustrated in Figure 9. These hold promise for applications like developing facial composites, creating personalized avatars, and improving facial recognition.

Figure 9: The Faces a la Carte dataset

The LAION-400M dataset is a treasure trove for researchers in multimodal learning, particularly those exploring the connection between vision and language [30]. This massive dataset holds a staggering 400 million image-text pairs, each featuring an image and its corresponding English caption, as illustrated in Figure 10.

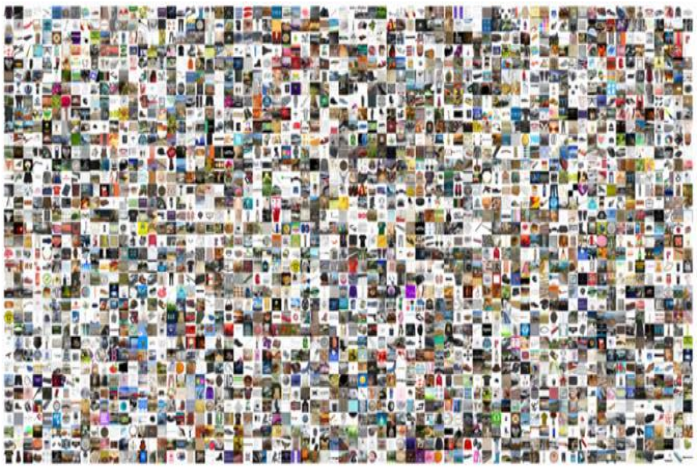

Figure 10: The LAION-400M dataset

Face2Text started as a research project with a unique dataset that automatically generates detailed descriptions of human faces in images [31]. Launched in 2018, it aimed to bridge the gap between computer vision and natural language processing. The project trained a model to analyze facial features and translate them into text. This involved creating the "Face2Text" dataset, containing over 5,600 images from CelebA, each with human-written descriptions capturing both visual attributes (e.g., glasses, smile) and even emotions, as illustrated in Figure 11. Interestingly, the research found that using generic image features rather than specialized face detection models led to better descriptions, showcasing the challenge of capturing the intricacies of human faces. The Face2Text project and its dataset became valuable resources for researchers working on automated image captioning and bridging the gap between visual and textual understanding. Moreover, a second version of the project was released in 2022, containing an expanded dataset of 10,559 images with 17,022 descriptions.

Figure 11: The Face2Text Dataset

The Bento800 dataset caters to the niche world of aesthetic box lunch design [32]. This first-of-its-kind resource offers 800 manually annotated images of diverse bento boxes, meticulously crafted to represent various lunch presentation styles. Categorized by food group, these images provide a rich training ground for AI models, as illustrated in Figure 12.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 3, pp 67-77, March-2025

https://doi.org/10.47001/IRJIET/2025.903009

Bento800 empowers researchers to delve into computational aesthetics, specifically focusing on generating visually appealing and culturally relevant bento box designs. This unique dataset holds potential for personalized suggestions, automated design tools, and even exploring cultural nuances of food presentation through AI analysis.

Figure 12: The Bento800 dataset

LAION-5B revolutionizes multimodal learning research [33]. This massive dataset boasts 5.85 billion image-text pairs in over 100 languages, empowering researchers to explore the connection between vision and language, as illustrated in Figure 13. Offering various access formats and functionalities, LAION-5B fuels advancements in image captioning, text-based image retrieval, and understanding the link between visual and linguistic representations. This dataset holds immense potential to accelerate progress across diverse fields, from AI creativity to human-computer interaction.

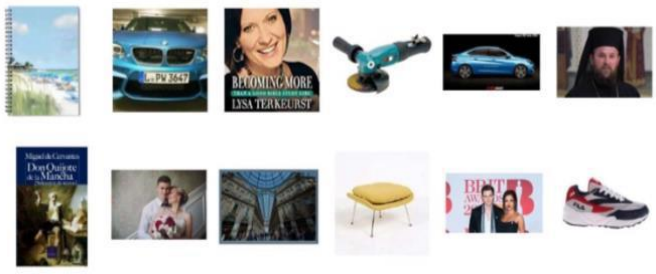

Figure 13: The LAION-5B Dataset

The DiffusionDB dataset caters to researchers exploring text-to-image generation with large language models (LLMs) [34], as illustrated in Figure 14. This unique resource offers text-code pairs, where each entry includes a textual description and the corresponding Python code for generating the desired image using a diffusion model. This empowers researchers to train LLMs that translate natural language descriptions into executable code, paving the way for advancements in text-driven image generation, automating image creation tasks, and potentially even exploring creative image generation through natural language prompts.

Figure 14: The DiffusionDB dataset

The COYO-700M dataset is a large-scale dataset containing over 747 million image-text pairs [35], as illustrated in Figure 15. It goes beyond just images and text, however, as it also includes various "meta-attributes" to enhance its usability for training diverse models. These meta-attributes provide additional information about the data, potentially including details like image source, location, or object labels.

Figure 15: The COYO-700M Dataset

DeepFashion-MultiModal, a goldmine for fashion image analysis and generation, holds over 44,000 high-resolution human images [36], as illustrated in Figure 16. Each image boasts a wealth of annotations: precise body part labels, keypoints, DensePose for intricate pose details, clothing attributes like sleeve length and fabric, and even textual descriptions. This multifaceted approach empowers researchers to explore exciting avenues like generating fashion images based on text descriptions, manipulating existing images with textual instructions, and delving into the multi-modal world of understanding fashion through both visual and textual data. Besides, these kinds of datasets can be used in recommendation systems such as [45].

Figure 16: DeepFashion-MultiModal Dataset

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 3, pp 67-77, March-2025

https://doi.org/10.47001/IRJIET/2025.903009

The ANNA (Abstractive News Captions dataset) is designed to advance news image generation using abstractive captions [37], as illustrated in Figure 17. It contains 30,000 image-caption pairs from the New York Times, focusing on contextual information while minimizing named entities. Preprocessing ensures captions are suitable for generative tasks. The dataset offers insights into image-text composition, including caption statistics and image analysis. Split into training, validation, and testing sets, ANNA benchmarks text-to-image models on quality, similarity, and contextual relevance. It's a valuable resource for studying abstractive image generation and the link between content and context in news media.

Figure 17: The ANNA Dataset

The DreamBooth dataset caters to researchers and enthusiasts exploring personalized text-to-image generation [38], as illustrated in Figure 18. Unlike traditional datasets, it focuses on specific subjects or styles with corresponding textual prompts. This allows users to train models to generate images based on their own descriptions, hobbies, or appearances, fostering creativity and personalization. Applications like personalized image generation, style transfer, and rapid concept creation become possible, making DreamBooth a valuable tool in the ever-evolving field of text-to-image generation.

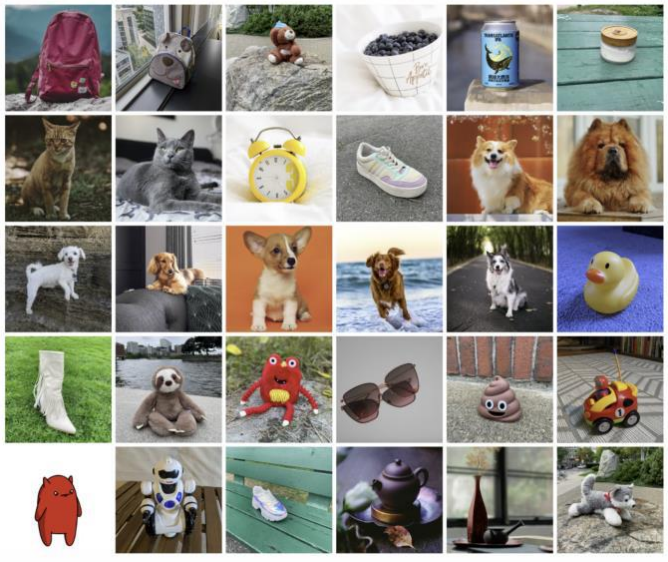

Figure 18: The DreamBooth dataset

III. CHALLENGES AND LIMITATIONS

Notwithstanding the considerable promise of T2I technology across several domains, it continues to encounter numerous hurdles.

3.1 Bias and Fairness

Although the considerable promise of text-to-image (T2I) technology in diverse domains, numerous hurdles remain in its advancement and implementation. A primary concern is the biases included in training data, especially with racial and gender representation [46]. These biases frequently arise from datasets that primarily represent specific demographics, leading to models that may inadequately represent minority populations or provide stereotypical pictures. Such biases can result in unjust performance, particularly when directed against varied demographic groups, yielding erroneous or inequitable outcomes [47]. For example, when required to provide images of professionals, models may predominantly depict men, highlighting the cultural biases embedded in the training data. Addressing this challenge requires the creation of more diverse and inclusive statistics that encompass a broader spectrum of cultural, linguistic, and demographic differences, hence ensuring equitable representation [48].

3.2 Computational and the Need for Efficient Training

A notable problem in dataset construction is the computational burden associated with training complex models on extensive collections of image-text pairs [49]. Most models trained on cloud that supports mutli GPUs with noval techniques such as [50]. Prominent systems such as DALL-E 2 [51] and GLIDE [15] rely on extensive datasets, necessitating considerable computational resources and infrastructure for processing. To resolve this issue, there is an increasing necessity to concentrate on assembling smaller, high-quality datasets that can facilitate quick training while preserving model correctness [52]. This method would enhance training accessibility for smaller institutions and academics lacking substantial computational resources. Furthermore, refining dataset architecture and augmenting algorithmic efficiency might diminish dependence on extensive data repositories, hence improving the accessibility and efficacy of T2I models [53]. Besides, using reliable software while training will reduce the computation resources [54].

3.3 Ethical Concerns in Dataset

Ethical considerations are paramount in the creation and application of datasets for text-to-image (T2I) models [55]. The datasets utilized to train these models can influence the ethical ramifications of the produced material, especially

ISSN (online): 2581-3048
Volume 9, Issue 3, pp 67-77, March-2025
https://doi.org/10.47001/IRJIET/2025.903009

concerning matters such as representation, privacy, and permission [56]. Datasets must be meticulously curated to prevent the perpetuation of damaging stereotypes, misrepresentations, or the exploitation of sensitive information [57]. Moreover, the issue of informed consent is crucial when utilizing photos obtained from public platforms, guaranteeing that the individuals depicted in the datasets have authorized their usage [58]. The construction of ethical datasets necessitates the assurance that created content does not endorse harmful views or perpetuate discriminatory actions. This necessitates a balance among data diversity, representation, and ethical data utilization to guarantee that T2I models are implemented in accordance with society norms and the respect for human rights.

IV. FUTURE WORKS

In the future, a primary emphasis should be on the creation of domain-specific datasets tailored for specialized applications in text-to-image generation. These datasets would serve businesses including healthcare, architecture, fashion, and entertainment, where accuracy and specialized knowledge are essential. In the realm of education, specifically for young children, these datasets may encompass age-appropriate pictures, cartoons, and visual aids pertinent to early childhood education subjects such as fundamental mathematics, literacy, colors, shapes, animals, and nature. Moreover, they may incorporate interactive components, such as visual narratives, to facilitate language acquisition and cognitive advancement in young learners.

V. CONCLUSION

The efficacy of text-to-image generative models depends fundamentally on the presence of high-quality, diversified datasets that capture the complex.

ACKNOWLEDGEMENT

The authors wish to express their appreciation and gratitude to the College of Computer Sciences and Mathematics, as well as the College of Education for Pure Science at the University of Mosul, for their invaluable support in the advancement of this study.

REFERENCES

[1] Y. X. Tan, C. P. Lee, M. Neo, and K. M. Lim, "Text-to-image synthesis with self-supervised learning," *Pattern Recognit Lett*, vol. 157, pp. 119–126, May 2022, doi: 10.1016/j.patrec.2022.04.010.

[2] S. K. Alhabeeb and A. A. Al-Shargabi, "Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction," *IEEE Access*, vol. 12, pp. 24412–24427, 2024, doi: 10.1109/ACCESS.2024.3365043.

[3] N. Zhang and H. Tang, "Text-to-Image Synthesis: A Decade Survey," Nov. 2024.

[4] S. O. Hasoon and M. M. Al-Hashimi, "Hybrid Deep Neural network and Long Short term Memory Network for Predicting of Sunspot Time Series," *Int J Math Comput Sci*, vol. 17, no. 3, pp. 955–967, 2022, [Online]. Available: http://ijmcs.future-in-tech.net

[5] R. Talal Ibrahim and F. Mahmood Ramo, "Hybrid Intelligent Technique with Deep Learning to Classify Personality Traits," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 2210–142, 2023, doi: 10.12785/ijcds/130119.

[6] "Disentangled Diffusion: T2I model to extract multiple concepts from a single image | AI-SCHOLAR | AI: (Artificial Intelligence) *Articles and technical information media."* Accessed: Mar. 01, 2025. [Online]. Available: https://ai-scholar.tech/en/articles/image-generation/T2I-DisenDiff

[7] Ian J. Goodfellow et al., "Generative Adversarial Networks," *Adv Neural Inf Process Syst*, vol. 27, Jun. 2014, doi: 10.48550/arXiv.1406.2661.

[8] Y. Namani, I. Reghioua, G. Bendiab, M. A. Labiod, and S. Shiaeles, "DeepGuard: Identification and Attribution of AI-Generated Synthetic Images," *Electronics (Basel)*, vol. 14, no. 4, p. 665, Feb. 2025, doi: 10.3390/electronics14040665.

[9] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," Nov. 2014.

[10] A.Ramesh et al., "Zero-Shot Text-to-Image Generation," in International Conference on Machine Learning, PMLR *(Proceedings of Machine Learning Research)*, Feb. 2021, pp. 8821–8831. doi: 10.48550/arXiv.2102.12092.

[11] A.Vaswani et al., "Attention Is All You Need," Jun. 2017.

[12] M. Dalal, A. C. Li, and R. Taori, "Autoregressive Models: What Are They Good For?," Oct. 2019.

[13] M. Khoshnoodi, V. Jain, M. Gao, M. Srikanth, and A. Chadha, "A Comprehensive Survey of Accelerated Generation Techniques in Large Language Models," May 2024.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society*, 2022, pp. 10674–10685. doi: 10.1109/CVPR52688.2022.01042.

ISSN (online): 2581-3048
Volume 9, Issue 3, pp 67-77, March-2025
https://doi.org/10.47001/IRJIET/2025.903009

[15] A.Nichol et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," *arXiv preprint arXiv*:2112.10741, Dec. 2021.

[16] L. Li et al., "T2ISafety: Benchmark for Assessing Fairness, Toxicity, and Privacy in Image Generation," Jan. 2025.

[17] M. Ahsan Habib et al., "Exploring Progress in Text-to-Image Synthesis: An In-Depth Survey on the Evolution of Generative Adversarial Networks," *IEEE Access*, vol. 12, pp. 178401–178440, 2024, doi: 10.1109/ACCESS.2024.3435541.

[18] R. Shelby, S. Rismani, and N. Rostamzadeh, "Generative AI in Creative Practice: ML-Artist Folk Theories of T2I Use, Harm, and Harm-Reduction," *in Proceedings of the CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM*, May 2024, pp. 1–17. doi: 10.1145/3613904.3642461.

[19] M.-E. Nilsback and A. Zisserman, "Automated Flower Classification over a Large Number of Classes," *in 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE*, Dec. 2008, pp. 722–729. doi: 10.1109/ICVGIP.2008.47.

[20] Wah, Catherine, Steve Branson, Peter Welinder, and Serge Belongie., "The Caltech-UCSD Birds-200-2011 Dataset," *California Institute of Technology*, 2011.

[21] Li Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]," *IEEE Signal Process Mag*, vol. 29, no. 6, pp. 141–142, Nov. 2012, doi: 10.1109/MSP.2012.2211477.

[22] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[23] R. Krishna et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *Int J Comput Vis*, vol. 123, no. 1, pp. 32–73, May 2017, doi: 10.1007/s11263-016-0981-7.

[24] X. Chen, L. Qing, X. He, X. Luo, and Y. Xu, "FTGAN: A Fully-trained Generative Adversarial Networks for Text to Face Generation," Apr. 2019.

[25] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "TediGAN: Text-Guided Diverse Face Image Generation and Manipulation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE*, Jun. 2021, pp. 2256–2265. doi: 10.1109/CVPR46437.2021.00229.

[26] Y. Zhou, "Generative Adversarial Network for Text-to-Face Synthesis and Manipulation," *in Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA: ACM*, Oct. 2021, pp. 2940–2944. doi: 10.1145/3474085.3481026.

[27] Z. Zhang et al., "UFC-BERT: Unifying Multi-Modal Controls for Conditional Image Synthesis," *Adv Neural Inf Process Syst*, vol. 34, 2021.

[28] Y. Jiang, Z. Huang, X. Pan, C. C. Loy, and Z. Liu, "Talk-to-Edit: Fine-Grained Facial Editing via Dialog," Sep. 2021.

[29] T. Wang, T. Zhang, and B. Lovell, "Faces à la Carte: Text-to-Face Generation via Attribute Disentanglement*," in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE*, Jan. 2021, pp. 3379–3387. doi: 10.1109/WACV48630.2021.00342.

[30] C. Schuhmann et al., "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," Nov. 2021.

[31] A.Gatt et al., "Face2Text: Collecting an Annotated Image Description Corpus for the Generation of Rich Face Descriptions," Mar. 2018.

[32] Y. Zhou and N. Shimada, "ABLE: Aesthetic Box Lunch Editing," *in Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and related Applications, New York, NY, USA: ACM*, Oct. 2022, pp. 53–56. doi: 10.1145/3552485.3554935.

[33] C. Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," Oct. 2022.

[34] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, "DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models," Oct. 2022.

[35] "GitHub - kakaobrain/coyo-dataset: COYO-700M: Large-scale Image-Text Pair Dataset." Accessed: Feb. 27, 2024. [Online]. Available: https://github.com/kakaobrain/coyo-dataset

[36] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu, "Text2Human: Text-Driven Controllable Human Image Generation," May 2022.

[37] A.A. Ramakrishnan, S. X. Huang, and D. Lee, "ANNA: Abstractive Text-to-Image Synthesis with Filtered News Captions," Jan. 2023.

[38] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," Aug. 2022.

[39] M. Chen et al., "Evaluating Text-to-Image Generative Models: An Empirical Study on Human Image Synthesis," Mar. 2024.

[40] X. Wu, S. Huang, and F. Wei, "Multimodal Large Language Model is a Human-Aligned Annotator for Text-to-Image Generation," Apr. 2024.

[41] M. M. Ismael, I. A. Saleh, and M. S. Student, "Software Maintenance Potential Prediction Based on

ISSN (online): 2581-3048
Volume 9, Issue 3, pp 67-77, March-2025
https://doi.org/10.47001/IRJIET/2025.903009

Machine Learning," *International Research Journal of Innovations in Engineering and Technology (IRJIET)*, vol. 6, no. 10, pp. 56–62, 2022, doi: 10.47001/IRJIET/2022.610009.

[42] H. Altememi and Y. Al-Irhaim, "A Comparative Study for Speech Summarization Based on Machine Learning: A Survey," *AL-Rafidain Journal of Computer Sciences and Mathematics*, vol. 16, no. 2, pp. 89–96, Dec. 2022, doi: 10.33899/csmj.2022.176595.

[43] R. S. Mahamed Najeeb and I. O. Abdul Majjed Dahl, "Brain Tumor Segmentation Utilizing Generative Adversarial, Resnet And Unet Deep Learning," *in 2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM), IEEE, Aug. 2022*, pp. 85–89. doi: 10.1109/ICCITM56309.2022.10031760.

[44] A.Ali and A. Taqa, "Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches," *JOURNAL OF EDUCATION AND SCIENCE*, vol. 31, no. 1, pp. 8–25, Mar. 2022, doi: 10.33899/edusj.2021.131895.1192.

[45] A.M.S. Saleh and A. Y. Taqa, "A Recent Trends in eBooks Recommender Systems: A Comparative Survey," *Iraqi Journal of Science*, pp. 487–511, Jan. 2024, doi: 10.24996/ijs.2024.65.1.39.

[46] V. Joynt et al., "A Comparative Analysis of Text-to-Image Generative AI Models in Scientific Contexts: A Case Study on Nuclear Power," Dec. 2023.

[47] A.F.de C. Vázquez and E. C. Garrido-Merchán, "A Taxonomy of the Biases of the Images created by Generative Artificial Intelligence," May 2024.

[48] T. Sandoval-Martin and E. Martínez-Sanzo, "Perpetuation of Gender Bias in Visual Representation of Professions in the Generative AI Tools DALL·E and Bing Image Creator," *Soc Sci*, vol. 13, no. 5, p. 250, May 2024, doi: 10.3390/socsci13050250.

[49] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and Applications of Large Language Models," Jul. 2023.

[50] Z. N. Al-Kateeb and D. B. Abdullah, "AdaBoost-powered cloud of things framework for low-latency, energy-efficient chronic kidney disease prediction," *Transactions on Emerging Telecommunications Technologies*, vol. 35, no. 6, Jun. 2024, doi: 10.1002/ett.5007.

[51] A.Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv*:2204.06125, vol. 1, no. 2, p. 3, Apr. 2022.

[52] I.Hussain Rather, · Sushil Kumar, · Amir, and H. Gandomi, "Breaking the data barrier: a review of deep learning techniques for democratizing AI with small datasets," *Artif Intell Rev*, vol. 57, p. 226, 123AD, doi: 10.1007/s10462-024-10859-3.

[53] K. Pilz, L. Heim, and N. Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities," Nov. 2023.

[54] S. I. Khaleel and L. F. Salih, "A survey of predicting software reliability using machine learning methods," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 1, p. 35, Mar. 2024, doi: 10.11591/ijai.v13.i1.pp35-44.

[55] X. Wang, X. Yi, X. Xie, and J. Jia, "Embedding an Ethical Mind: Aligning Text-to-Image Synthesis via Lightweight Value Optimization," Oct. 2024, doi: 10.1145/3664647.3681652.

[56] M. Al-kfairy, D. Mustafa, N. Kshetri, M. Insiew, and O. Alfandi, "Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective," *Informatics*, vol. 11, no. 3, p. 58, Aug. 2024, doi: 10.3390/informatics11030058.

[57] A.Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," Oct. 2021.

[58] G. A. Tahir, "Ethical Challenges in Computer Vision: Ensuring Privacy and Mitigating Bias in Publicly Available Datasets," Aug. 2024.

AUTHORS BIOGRAPHY

Haitham ALHAJI is a Ph.D. candidate in computer science with an M.Sc. from Bahcesehir University (2021) and a B.Sc. from the University of Mosul (2017). His research focuses on computer vision, machine learning, and deep neural networks. Contact: haithamtalhaji@yahoo.com

Prof. Asst. Dr. Alaa, at the University of Mosul, holds a Ph.D. in Computer Science focused on anti-spam filtering and an M.Sc. in Applied Object-Oriented Software Engineering. Her research interests include pattern recognition, AI applications, and machine learning. Contact: alaa.taqa@uomosul.edu.iq

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 3, pp 67-77, March-2025

https://doi.org/10.47001/IRJIET/2025.903009

Citation of this Article:

Haitham ALHAJI, & Alaa Yaseen Taqa. (2025). Text-to-Image Datasets: Characteristics, Challenges, and Opportunities. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(3), 67-77. Article DOI https://doi.org/10.47001/IRJIET/2025.903009

****

