Comparative Analysis of State of The Art Methods for Text-To-Image Generation

Anand Khanna*
Arizona State University
Tempe, Arizona, USA
akhann40@asu.edu

Fatemeh Elhambakhsh*
Arizona State University
Tempe, Arizona, USA
felhamba@asu.edu

Pranav Hegde*
Arizona State University
Tempe, Arizona, USA
phegde7@asu.edu

Rahul Shah*
Arizona State University
Tempe, Arizona, USA
rshah107@asu.edu

ABSTRACT

Leveraging large language models for interpreting and converting textual descriptions into visual content represents a significant advancement in bridging the gap between linguistic and visual domains. Within this context, the task of text-to-image synthesis emerges as a crucial area of study, aiming to transform narrative text into corresponding images with high fidelity. This research focuses on comparing prominent methodologies employed for this purpose: Generative Adversarial Networks (GANs), diffusion-based models, Variational Auto Encoders (VAEs). Our comparative analysis hinges on the Fréchet Inception Distance (FID) metric. By employing FID, we aim to quantitatively evaluate the performance of GANs, diffusion-based models, and VAEs in producing images that are not only visually compelling but also accurately reflect the textual descriptions.

KEYWORDS

Large Language Models, Text-to-Image Synthesis, Generative Adversarial Networks, Diffusion Models, AutoEncoders

ACM Reference Format:
Anand Khanna, Fatemeh Elhambakhsh, Pranav Hegde, and Rahul Shah. 2024. Comparative Analysis of State of The Art Methods for Text-To-Image Generation. In *Proceedings of Data Mining (CSE 572).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Text-to-image generation involves models that can interpret text inputs and produce high-quality images according to the provided descriptions. The origin of text-to-image generation with neural networks dates back to the development of Generative Adversarial Networks (GAN). Diffusion models, a leading category of generative models, create images by methodically introducing noise in a series of steps. The remarkable achievements of diffusion models in image synthesis have established them as the primary image decoders for text-to-image models [1].

Large language models (LLMs) are becoming more popular across academic and industrial sectors due to their exceptional performance across a wide range of applications. As LLMs become more integral to research and everyday tasks, assessing their impact is becoming more essential. This evaluation is important not just in terms of task-specific performance but also for a broader societal understanding of their possible risks [2].

Diffusion models, Variational Autoencoders (VAEs), and Generative Adversarial Networks (GANs) each possess unique advantages and drawbacks when it comes to image-to-image, text-to-text, and text-to-image applications. No single approach consistently surpasses the others in performance. In this research, we aim to evaluate and contrast the efficacy of these three methodologies specifically for text-to-image tasks.

The remainder of the paper is as follows: Section 2 reviews the literature. Section 3 introduces the methodologies. Section 4 presents a dataset. Section 5 provides the results. Section 6 discusses the findings, and Section 7 concludes the study and outlines future work.

2 LITERATURE REVIEW

The rapid advancements in text-to-image generation models have highlighted the dynamic interplay between machine learning technologies and their applications in creating visual content from textual descriptions. In the seminal work by Aditya Ramesh et al. [14] the paper introduces a hierarchical approach utilizing the Contrastive Language-Image Pre-Training (CLIP) model to generate images from text. This model enhances the diversity and photorealism of generated images through a two-stage diffusion process, addressing the intricate balance between style and semantic accuracy. Despite its strengths, the model reveals limitations in attribute-object binding and the rendering of complex details, underscoring the challenges in achieving comprehensive visual coherence.

Expanding the accessibility and efficiency of diffusion models, Li et al. [8] present SnapFusion, a method that dramatically accelerates the deployment of text-to-image models on mobile platforms. By

*All authors contributed equally to this research.

CSE 572, Spring 2024, ASU, Tempe, AZ

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

optimizing the model architecture for reduced redundancy, SnapFusion achieves remarkable inference speeds without compromising image quality, demonstrating the potential for real-world applications on handheld devices. However, its performance variability across different hardware and the large model size highlights the ongoing need for optimization and broad compatibility in model deployment.

In parallel, the integration of large language models (LLMs) with diffusion techniques has opened new avenues for enhancing text comprehension and image generation fidelity. Lian et al. [9] introduce a training-free, two-stage generation process that significantly improves the generation of accurate images from complex prompts, emphasizing the synergy between textual understanding and visual representation. Meanwhile, innovative approaches to text-image alignment and multimodal learning are explored by Liu et al. [11] through Language Quantized AutoEncoders (LQAE), and the combined strengths of Conditional Variational Autoencoders and Conditional Generative Adversarial Networks are harnessed by Tibebu et al. [18] to synthesize high-resolution images from text. Each of these contributions, along with the straightforward yet effective DF-GAN model proposed by Tao et al. [17] and the dual-attention DTGAN model by Zhang and Schomaker [19], collectively push the boundaries of text-to-image synthesis, addressing both the technical and creative challenges in the field

3 METHODOLOGY

In this section, we will provide a background of GANs, DDPMs and VAEs. Then, we will explain four different generative text-to-image methods based on GANs, DDPMs, and VAEs that we will be using in our comparative analysis.

3.1 Background

3.1.1 ***Generative Adversarial Networks (GANs).*** GANs [4, 6] are a type of generative model. Within a GAN framework, two neural networks function as differentiable functions engaging in a sort of competition. These networks, known as the generator and the discriminator, each have distinct roles. The generator's objective is to create data that mimics a specific probability distribution. On the other side, the discriminator acts as a judge, determining whether its input is sourced from the generator or the actual training dataset. As a result: The generator aims to increase the likelihood that the discriminator will mistakenly classify its outputs as authentic. The discriminator steers the generator towards crafting more lifelike images. Figure 1 shows the architecture of a GAN model.

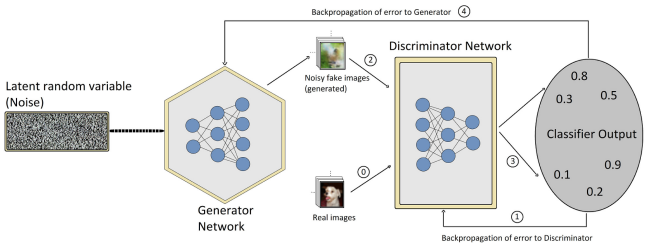

Figure 1: A Generative Adversarial Network (GAN) [6].

3.1.2 ***Denoising Diffusion Probabilistic Models (DDPMs).*** A diffusion probabilistic model [7] is a type of Markov chain that has been parameterized and trained through variational inference to generate samples that resemble the data after a certain period. This model learns to reverse a diffusion process, which is another Markov chain that incrementally introduces noise to the data in the reverse direction of the sampling process until the original signal is obliterated. When the diffusion involves only minor amounts of Gaussian noise, the transitions in the sampling chain can also be modeled as conditional Gaussians. This approach enables straightforward neural network parameterization. Figure 2 the architecture of a DDPM.

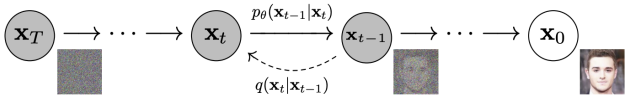

Figure 2: A Denoising diffusion Probabilistic Model (DDPM) [7].

3.1.3 ***Variational Auto Encoders (VAEs).*** Variational Autoencoders (VAEs) [3] are a type of advanced deep learning generative model commonly applied to capture and represent complex, high-dimensional data using a condensed, low-dimensional latent space. In the foundational VAE framework, each input data vector is treated as an independent entity during processing. Figure 3 indicates the architecture of a VAE.

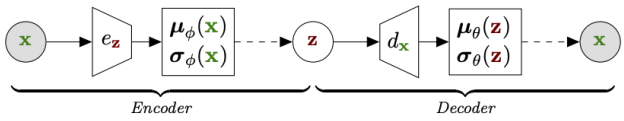

Figure 3: A Variational Auto Encoder (VAE) [3].

3.1.4 ***Contrastive Language-Image Pre-training (CLIP).*** Contrastive Language-Image Pre-training (CLIP) [13] is a cutting-edge model developed by OpenAI. It consists of a neural network that has been trained on roughly 400 million pairs of text and images. The training process employs a method known as contrastive learning, which is designed to bridge the gap between text and images. This integration facilitates operations such as image classification, which can be performed based on the similarity between text and images. Figure 4 illustrates the architecture of CLIP.

3.2 Models

3.2.1 ***Generative Adversarial CLIPs (GALIP).*** Generative Adversarial CLIPs [16], or GALIP, utilize the robustly pre-trained CLIP model within both the generator and discriminator components. The sophisticated scene comprehension capabilities of CLIP enhance the discriminator's ability to precisely evaluate the quality of images. This integration allows GALIP to effectively leverage CLIP's advanced understanding in the generative adversarial network setup. Figure 5 represents the architecture of GALIP.

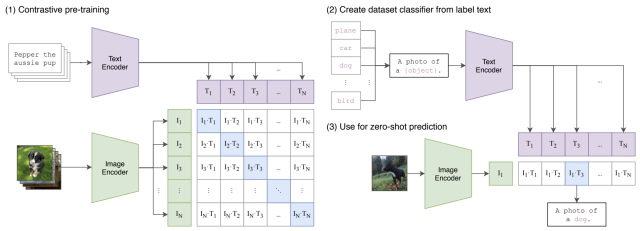

Figure 4: A Contrastive Language-Image Pre-training method (CLIP) [13].

Figure 5: A Generative Adversarial CLIP method (GALIP) [16].

3.2.2 ***Language-Free Training for Text-to-Image Generation (LAFITE).*** LAFITE [20] represents a flexible system adept across various text-to-image generation scenarios, encompassing language-free, zero-shot, and fully supervised learning approaches. Figure 6 illustrates the architecture of LAFITE.

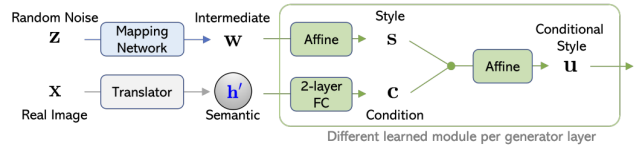

Figure 6: A Language-Free Training for Text-to-Image Generation METHOD (LAFITE) [20].

3.2.3 ***Vector Quantized Diffusion (VQ-Diffusion).*** Vector Quantized Diffusion (VQ-Diffusion) model [5] integrates a vector quantized variational autoencoder (VQ-VAE) with a conditional version of DDPM. This combination exploits the latent space for text-to-image generation tasks, offering significant improvements over existing methods. It effectively removes the unidirectional bias found in traditional approaches and introduces a masking and replacement diffusion strategy. This innovation helps prevent the buildup of errors—a notable issue in previous methods—enhancing the model's overall efficiency and accuracy in generating images from text. Figure 7 indicates a VQ-Diffusion architecture.

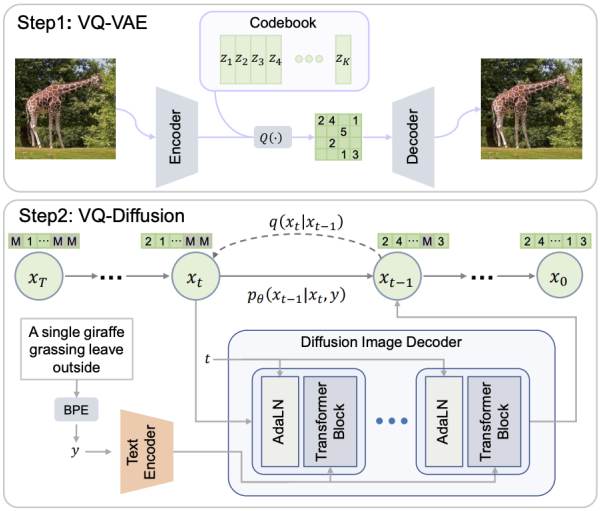

Figure 7: A Vector Quantized Diffusion (VQ-Diffusion) [5]

3.2.4 ***Stable Diffusion.*** Stable Diffusion is a form of Latent diffusion model [15] which represents a straightforward and effective approach to enhancing the efficiency of training and sampling in denoising diffusion models without compromising their performance. This technique employs an autoencoder model to discover a space that is perceptually similar to the image space but with greatly reduced computational demands. Figure 8 provides an in-depth look at the latent diffusion model's specifics.

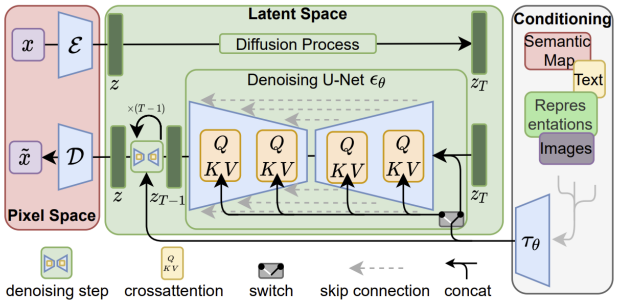

Figure 8: Structure of Latent Diffusion Models [15]

4 DATASET

In this section, we present information about the dataset used to evaluate all these models.

The Microsoft Common Objects in Context (MS COCO) dataset [10] is a very large-scale dataset for testing object detection, segmentation, and captioning. It contains 328,000 total images, and 121,000 for text-to-image generation for 80 object categories—ranging from everyday items like "person," "bicycle," and "car" to animals like "cat," "dog," and "elephant.

The dataset is also accompanied by image captioning annotations, such that each image is annotated with at least five descriptive

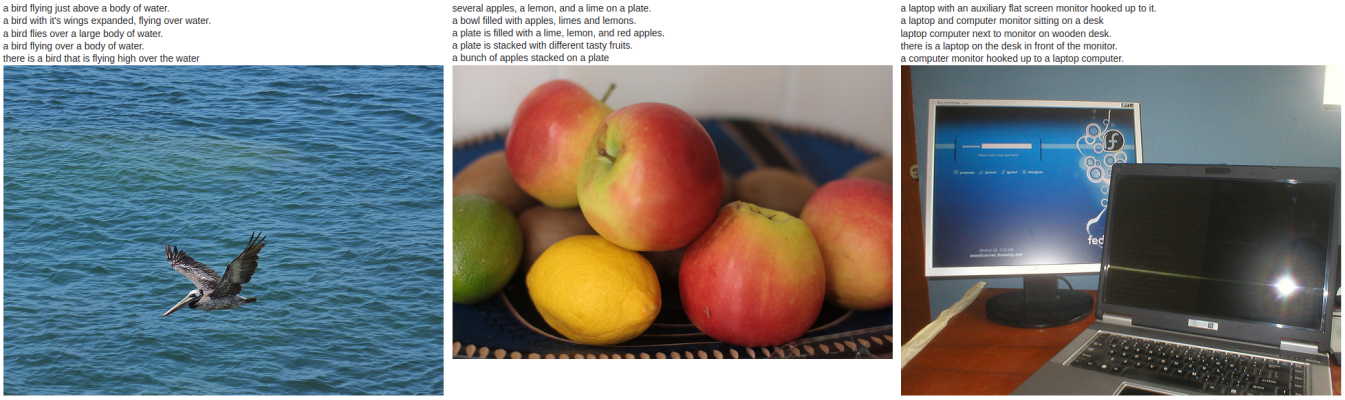

Figure 9: Sample Images with their annotations from the MS COCO Dataset

sentences. The dataset is typically split into training (82,000 images), validation (40,000 images), and test sets (40,000 images). The largest set i.e. training is used for modelling, model tuning with validation, and finally for evaluation of the models using the test set. During training, we generate the corresponding text embedding as the average of four randomly selected descriptions corresponding to the image. Figure 9 shows a select few samples from the dataset along with the captions accompanying them that can be used for text-to-image generation.

5 RESULTS

5.1 Generated images

Figure 10 illustrates the generated images of the different models for the prompt "a bunch of furniture in a home living room." The results indicate that stable diffusion is capable of generating high-quality images as compared to the other methods. In general, the diffusion-based models generate better-quality images compared to the GAN-based models.

5.2 Quantitative results

For the quantitative analysis of our text-to-image generation models, we employ the widely recognized Fréchet Inception Distance (FID) [12] metric.

The Frechet Inception Distance assesses the similarity between two groups of images, one being the output from an image creation process and the other a reference set of real images. It quantifies the distance between feature vectors extracted from InceptionV3 network layers. FID calculates the difference in the mean and covariance of these feature vectors, expressed mathematically as

$$FID = \left\|\mu_r - \mu_g\right\| + Tr\left(\sum r + \sum g - 2\left(\sum r \sum g\right)^{1/2}\right) \quad (1)$$

Here, μ_r and $\sum r$ denote the mean and covariance of the real images' feature vectors, while μ_g and $\sum g$ correspond to those of the generated images.A reduced FID score implies a greater similarity between the created images and actual images, indicative of higher image quality.

We also use the average generation time for one image and the model size in our comparisons. All the experiments were performed on an NVIDIA A30 GPU.

We plot our findings in the below graphs. Figure 11 contains the graph of FID against Model Size in MB.

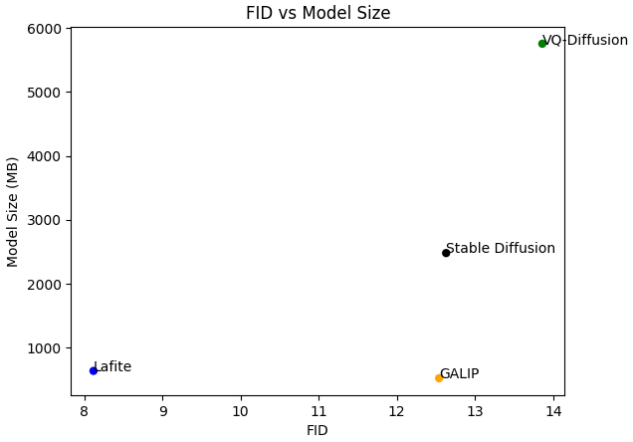

Figure 11: FID vs Model Size (MB)

Figure 12 contains the graph of FID against the average time in seconds taken by the models to generate one imaage.

Figure 10: Generated samples for the prompt "a bunch of furniture in a home living room".

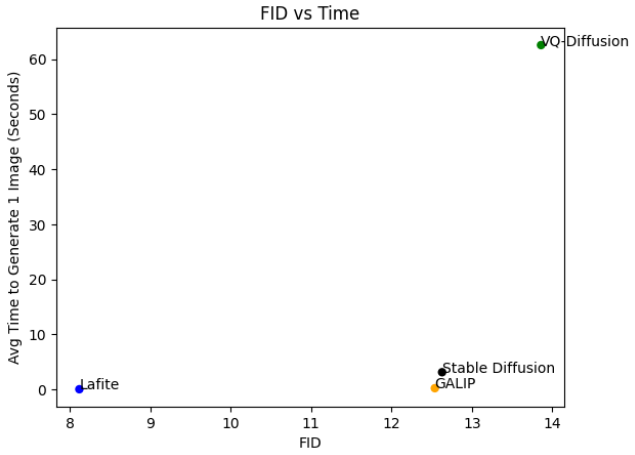

Figure 12: FID vs Model Size (MB)

Figure 13 contains the graph of Model Size in MB against the average image generation time in seconds.

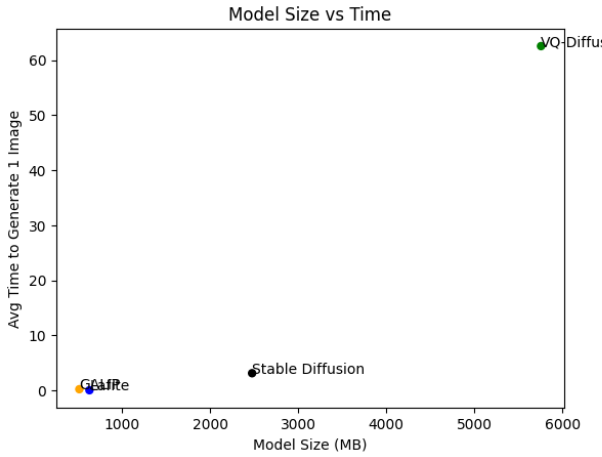

Figure 13: FID vs Model Size (MB)

In all these graphs, the bottom left quadrant indicates a better model.

6 DISCUSSION

Analysing Figure 11, we see that LAFITE is the best-performing model, while also being the smallest. GALIP is of similar size to LAFITE but falls a bit behind in the FID score. Both the Diffusion models' FID score is lower than the GAN models, and they also have bigger model sizes

From Figure 12, we can see that LAFITE is the best-performing model followed by GALIP. Even though Stable Diffusion is bigger than GALIP, its image generation time is just a bit higher.

Comparing the Model Size and Image Generation Time in Figure 13 we can see that both the GAN-based models of LAFITE and GALIP perform the best, with the diffusion models lagging. VQ-Diffusion in particular is the biggest and slowest, while also having the worst FID score.

In all these three graphs we observe the trend of GAN-based models outperforming the Diffusion based models, being smaller, faster and having a better FID score. This might indicate that GAN-based models are better, but based on Figure 10, we can see that Diffusion Models produced a better-looking image. This indicates that the FID metric alone is not enough to quantify these models and we might need more metrics.

7 CONCLUDING REMARKS AND FUTURE WORK

In conclusion, we tested four models of different architectures for text-to-image generation. In general, Diffusion models were larger and slower but also produced better results. The FID scores for all the models were pretty similar, slightly favouring GANs, but the images produced by Diffusion models were better on visual inspection, indicating FID might not capture everything.

In future works, we could evaluate more models on more datasets as we were limited to pre-trained models and the MS COCO dataset due to limited resources. We also plan to use more metrics such as Inception Score and human evaluation since FID alone is not enough to accurately evaluate these models.

REFERENCES

[1] Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, and Shuaiwen Leon Song. 2023. RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model. arXiv:2309.00810 [cs.CV]

[2] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A Survey on Evaluation of Large Language Models. arXiv:2307.03109 [cs.CL]

[3] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. 2020. Dynamical Variational Autoencoders: A Comprehensive Review. *arXiv preprint arXiv:2008.12595* (2020).

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative Adversarial Networks. *Commun. ACM* 63, 11 (2020), 139–144.

[5] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10696–10706.

[6] G. M. Harshvardhan, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. 2020. A Comprehensive Survey and Analysis of Generative Models in Machine Learning. *Computer Science Review* 38 (2020), 100285.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[8] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. 2023. SnapFusion: Text-to-Image Diffusion Model on Mobile Devices within Two Seconds. arXiv:2306.00980 [cs.CV]

[9] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2024. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. arXiv:2305.13655 [cs.CV]

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV]

[11] Hao Liu, Wilson Yan, and Pieter Abbeel. 2023. Language Quantized AutoEncoders: Towards Unsupervised Text-Image Alignment. arXiv:2302.00902 [cs.LG]

[12] Artem Obukhov and Mikhail Krasnyanskiy. 2020. *Quality Assessment Method for GAN Based on Modified Metrics Inception Score and Fréchet Inception Distance*. 102–114. https://doi.org/10.1007/978-3-030-63322-6_8

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[14] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[16] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. 2023. Galip: Generative Adversarial Clips for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14214–14223.

[17] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. arXiv:2008.05865 [cs.CV]

[18] Haileleol Tibebu, Aadin Malik, and Varuna De Silva. 2022. *Text to Image Synthesis Using Stacked Conditional Variational Autoencoders and Conditional Generative Adversarial Networks*. Springer International Publishing, 560–580. https://doi.org/10.1007/978-3-031-10461-9_38

[19] Zhenxing Zhang and Lambert Schomaker. 2021. DTGAN: Dual Attention Generative Adversarial Networks for Text-to-Image Generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. https://doi.org/10.1109/ijcnn52387.2021.9533527

[20] Y Zhou, R Zhang, C Chen, C Li, C Tensmeyer, T Yu, J Gu, J Xu, and T Sun. 2021. Lafite: Towards Language-Free Training for Text-to-Image Generation. *arXiv preprint arXiv:2111.13792* 2 (2021).