DOI: 10.24996/ijs.2025.67.1.41

ISSN: 0067-2904

Leveraging GANs and Vision Transformers for Text-to-Image Synthesis

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

[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*

Received: 18/11/2024 Accepted: 28/1/2025 Published: xx

Abstract

Text-to-image (T2I) in recent advances has proven an important headway, but high-quality and efficient image generation is still a challenge. We proposed GANViT, a new model for generative adversarial vision transformer. It is designed for fast and efficient high-quality T2I synthesis. GANViT addresses limitations in existing models, such as extensive data requirements for training, multi-phase processes that slow down synthesis speed, and many parameters needed to achieve adequate performance. GANViT consists mainly of a generator and a discriminator, but these parts are separately based on a vision transformer (ViT). The generator ViT utilizes a feature bridge for fine-tuning, thus enhancing image generation capabilities. Conversely, the discriminator ViT interprets complex scenes through a feature extraction module and assessment phase. GANViT demonstrates substantial improvements in synthesizing images. It achieves a 5.01 Frechet Inception Distance score on the Common Objects in Context dataset and over 10 on the Caltech-UCSD Birds dataset, with 33 times data reduction, 4.5 times fewer parameters, and 23 times faster processing.

Keywords: Text-to-Image Synthesis, Generative Adversarial Networks (GANs), Vision Transformers, Generative Models, GAN-based Image Generation, Transformer Models in Image Synthesis.

توليد الصور من النص باستخدام الشبكات التوليدية الخصومية والمحولات البصرية

هيثم الحاجي[1]*, الاء ياسين طاقة[2]

[1]قسم علوم الحاسوب, كلية علوم الحاسوب والرياضات, جامعة الموصل, نينوى, العراق

[2]قسم علوم الحاسوب, كلية التربية للعلوم الصرفة, جامعة الموصل, نينوى, العراق

الخلاصة

لقد حققت أساليب توليد الصور من النص (T2I) تقدمًا ملحوظًا في السنوات الأخيرة؛ ومع ذلك، لا تزال هناك تحديات مستمرة في هذا المجال البحثي. من اجل توليد صور ذات دقة عالية ومبنية على النص المدخل بكمية بيانات اقل مع وقت معالجة اقل قد اقترحنا نموذج GANViT، وهو نموذج جديد يمثل المحول التوليدي الخصومي البصري. تم تصميم هذا النموذج ليكون سريعًا وفعالًا في توليد صور عالية الجودة من النصوص. يعالج GANViT القيود الموجودة في النماذج الحالية، مثل الحاجة الكبيرة للبيانات لتدريبها، والعمليات متعددة

* Email: haitham.22csp52@student.uomosul.edu.iq

المراحل التي تبطئ سرعة التوليد، وعدد المعلمات الكبير اللازم لتحقيق أداء جيد. يتألف GANViT بشكل رئيسي من مولد ومميز، لكن هذه الأجزاء تعتمد بشكل منفصل على المحول البصري (ViT). يستخدم في المولد مع ViT جزءا يدعى جسر الخصائص لضبط الأداء، مما يعزز من قدرات توليد الصور. في المقابل، يقوم المميز مع ViT بفهم المشاهد المعقدة من خلال جزء يدعى وحدة استخراج الميزات ووحدة التقييم. يُظهر GANViT تحسينات كبيرة في توليد الصور، حيث يحقق نتيجة ملحوظة حسب مقياس Frechet Inception Distance قدرها 5.01 على مجموعة بيانات Common Objects in Context، وأكثر من 10 على مجموعة بيانات Caltech–UCSD Birds، مع تقليل حجم البيانات بنسبة 33 مرة، وتقليل عدد المعلمات بمقدار 4.5 مرة، وزيادة سرعة المعالجة بمقدار 23 مرة.

1. Introduction

Conventional image synthesis methods, which rely on image labels or attributes, offer limited adaptability compared with text-based image generation. Text-based synthesis has expanded applications, such as auxiliary structure design, photo-editing, and scene restoration [1], [2]. The use of generative adversarial networks (GANs) [3] has produced promising outcomes in text-to-image (T2I) synthesis. However, generating realistic images based on textual descriptions remains challenging, as demonstrated in previous studies [4].
Numerous pretraining methods have been recently developed, using diverse frameworks such as latent diffusion model (LDM) [5], DALL-E [5], and GLIDE [6] and employing techniques from diffusion and autoregressive modeling. These models have demonstrated remarkable efficacy in transforming textual descriptions into visual content, outperforming earlier GANs. Three limitations suggest the need for alternative approaches to the previous models.

First, they require substantial data and many parameters, thus increasing risks and costs. Second, the aforementioned models rely on several sequential phases and smooth latent space, unlike GANs, which partition the generation process into a structured latent space. Third, their training and sampling processes can be inefficient because of the extensive steps required.

To overcome these limitations, we proposed using GANs with a smooth latent space for rapid image generation. Utilizing vision transformer (ViT) technology [7], such as OpenAI's large-scale pretraining models, which supports robust textual content transformation into visual representations.

The primary reason for using ViT is to extract information from images through an image encoder. This process aligns the captured image with the textual representation. Understanding complex images is further supported by utilizing these pretraining datasets. This pretraining model is constructed from a broad range of publicly available resources, encompassing various visual forms such as images, drawings, and similar materials across diverse domains. Including this feature contributes to the overall generalization of the model and serves as a secondary rationale for employing ViT.
Our model incorporates unique features by integrating the ViT model within the generator and discriminator components, hence enhancing text-to-image synthesis through advanced scene comprehension and domain generalization capabilities. Components such as the Discriminator ViT, Generator ViT with Feature Bridger, and Prompt Tuning collaborate to enhance image synthesis by extracting salient visual features and augmenting generating capacity.

Our model has superior capability in generating high-quality, intricate images characterized by meticulous details, precise object contours, and remarkable realism in human facial

synthesis. It enhances efficiency and speed, attaining remarkable outcomes with reduced computational resources while preserving high quality.

This research is structured into several sections. The second section reviews relevant related works on contemporary models and methodologies for generating images from text. The following section presents our model's structural framework. The final section provides the outcomes of the models across various datasets.

2. Literature review

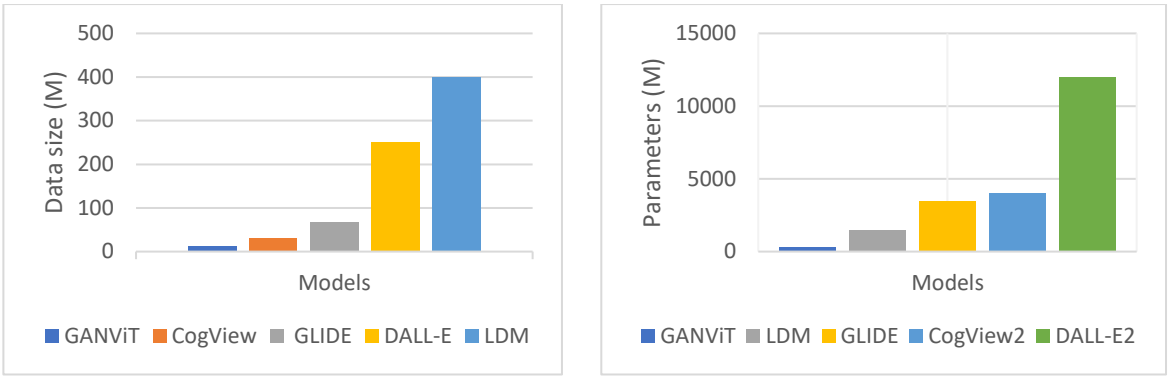
Generative adversarial network with interpolated conditioning and character-level sentence [8] is a method that uses character-level convolutional neural networks and recurrent neural networks to transform textual descriptions into visual features. It is considered the first conditional GAN, along with attentional GAN [9], another classical model. Using attention and alignment mechanisms, various T2I models have been developed, including channel-wise feature attention hierarchical attention GAN [10], deeply convolutional GAN [11], dual attention GAN [12], object-aware spatial attention GAN [13], and multi-discriminator GAN [14].

In addition, several models employ multi-stage networks, including multi-resolution progressive GAN [15], multi-task conditional GAN [16], knowledge-aware hierarchical GAN [17], divergence GAN [18], dual-generator GAN [19], balanced attention GAN [20], deep semantic generation GAN [21], semi-supervised T2I GAN [22], and differential evolution GAN [23], or cGAN as used in [24]

Conversely, the vector quantized variational autoencoder [25] approach produces high-quality images across various datasets by adjusting specific hyperparameters. DALL-E [5] utilizes a transformer architecture trained on an extensive dataset to generate superior images without labelled data. GLIDE [6] introduces a text-guided diffusion method for creating and editing photorealistic images. Human assessors prefer this model over DALL-E [5] for its ability to capture photorealism and retain caption alignment with contrastive language–image pretraining (CLIP) guidance.

The LDM [26] achieves effective image synthesis while notably reducing computational costs compared with pixel-based models. Figure 1 (a) shows the data size of each model. Moreover, LDM offers a versatile conditioning mechanism suitable for multi-modal training. Imagen integrates transformer language models with diffusion techniques to generate photorealistic images and enhance language comprehension. This integrated approach leverages models like bidirectional encoder representations from transformers, text-to-text transfer transformer, and CLIP, each tailored to specific tasks.

CogView2 [27] improves the efficiency and accuracy of image generation through hierarchical transformers and local parallel autoregressive generation, enabling interactive text-guided editing. Figure 1 (b) shows the number of parameters for each model.



(a) (b)

Figure 1: This figure shows an overview of the data size (a) and parameters used (b) for diverse text-to-image generation models, including DALL-E [5], GLIDE [6], LDM [26] CogView2 [27] and our proposed model GANViT.

3. Proposed methods

This research proposes a model that combines the GAN framework with ViT [28], referred to as GANViT. GANViT incorporates the pre trained CLIP [28] model within both the generator and discriminator. GANViT consists of two main parts, a discriminator and a generator, which are integrated with ViT, as shown in Figure 2. First, the model's discriminator, known as discriminator ViT, comprises an image encoder with frozen weights and a learnable discriminator to derive significant visual features from images. The ViT gathers information from complex scenes and extracts features from several layers through weight freezing.

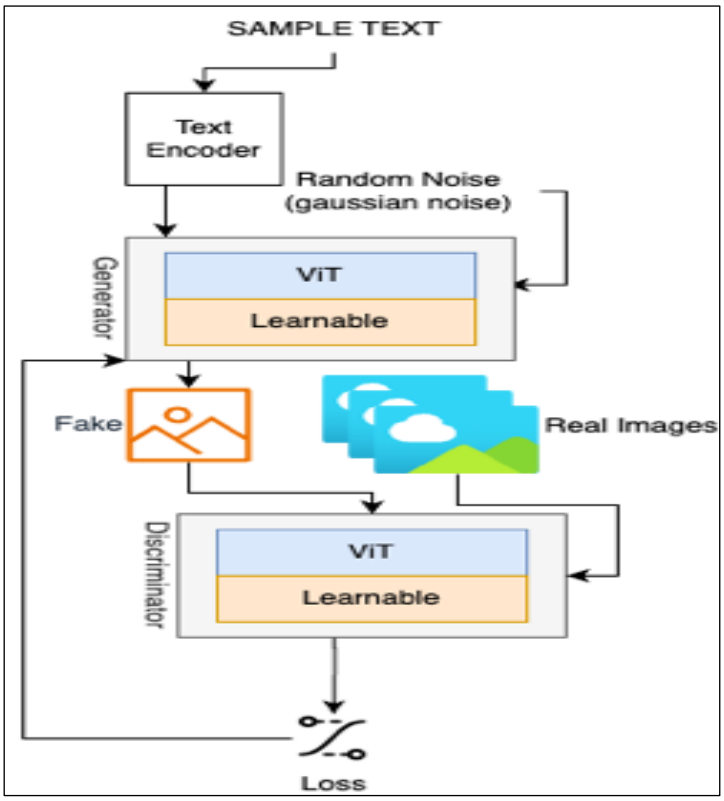


Figure 2: Overview of the proposed model GANViT

Moreover, the discriminator's evaluation can be precisely adjusted to enhance the quality of generated images. The second component, known as generator ViT, enables model generalization without relying on sketching techniques [29] used by other models. Like the discriminator, it includes a ViT with frozen weights and a learnable generator. The CLIP [28] model functions as the text encoder within the GANViT architecture, producing a global sentence vector from textual descriptions, which is used in the generator and discriminator for T2I synthesis. The discriminator ViT aggregates CLIP features from specific layers and applies prompts for task adaptability in image generation. The text encoder is crucial for converting text descriptions into global sentence vectors, supporting efficient, high-quality T2I synthesis in GANViT.

In text encoding, we begin by tokenizing the caption as input for our encoder using CLIP. The resulting tokens are fed into a transformer-based encoder, producing two types of embeddings: word embeddings and sentence embeddings for the whole caption. Our text encoder converts these tokens into dense vectors, incorporating positional embeddings to maintain word order. The transformer layers apply self-attention to identify contextual relationships among tokens. Finally, the model extracts the complete phrase embedding and projects it into latent space. These embeddings are then passed to the GAN generator, conditioning the generation process to ensure alignment between the text and the generated image, as shown in Algorithm 1.

Algorithm 1: Text Encoding for T2I Synthesis

Step 1: Tokenization
Utilize CLIP's tokenizer to tokenize the input text T.
***Output**: Sequence of tokens tokens = tokenize(T).*
Step 2: Token Embedding
Transform tokens into their vectors (dense) representation.
***Output**: token embeddings, denoted as te = e (tokens).*
Step 3: Positional Embedding
Incorporate positional embeddings into the token embeddings to represent the sequence of tokens.
***Output**: Positional token embeddings are derived by adding position to token embeddings, as pe = add_p(te).*
Step 4: Transformer Processing
Process positional embeddings via Transformer network.
***Output**: embeddings context derived from the transformer to positional embeddings*
Step 5: Extract Sentence and Word Embeddings
Obtain the s embedding from output of the Transformer.
Obtain word embeddings for each token.
***Output**: se=extract_s(context), word_e=extr_w (context).*
Step 6: Projection to Latent Space
Map sentence e onto latent space with a projection layer
***Output**: latent_sentence_e = project(sentence_e).*
Step 7: Return the Sentence, Word Embeddings
Return the latent_sentence_e and word_e to influence the image generation process in the GAN.
end

The text is initially divided into separate tokens via CLIP's tokenizer, resulting in a sequence of tokens. Subsequently, these tokens are transformed into dense vector representations known

as token embeddings. Positional embeddings are incorporated into the token embeddings to represent the sequence order of the tokens. The Transformer network processes this combined representation, yielding a contextual embedding. Both sentence-level and word-level embeddings are derived from this. The sentence embedding is subsequently mapped into a latent space via a projection layer, resulting in a final latent sentence embedding. The sentence and word embeddings are utilized to impact the image production process, guaranteeing coherence between the input text and the produced image.

3.1. Generator ViT

Figure 3 shows a frozen ViT linked with another component called a generator, which is learnable. This module takes two inputs: the encoded text, processed through a text encoder, and a noise vector modeled using the Gaussian distribution to increase generation diversity. The learnable generator comprises a feature bridge, prompt tuning functionality, and an image generator.

These inputs undergo three processing phases. First, the feature bridge, as demonstrated in Figure 3, adjusts each sampled text and noise input before passing it to the ViT. Next, the prompt tuning uses multiple transformer blocks to improve adaptability. Finally, the image generator produces the output based on the sampled phrases, Gaussian noise, and specific features from the bridge. The feature bridge contains fully connected layers and several fusion blocks to refine the visual representation from ViT. This module receives vectors representing both sentence and noise. The noise is reshaped to dimensions (7, 7, 64) for height, width, and depth. Sequential fusion blocks [30] then reduce noise from the previous layer. Each fusion module includes convolution (Conv) layers, and two additional fusion blocks merge text features with bridged features.

The prompt tuning module [31], next in generator ViT, bridges the data gap between text and image, using input vectors of text and noise and following the predictions made by pre-trained ViT [28]. As the last layers are specific to visual representation, prompts are generally not used here. In practice, the generator ViT effectively maps CLIP features to the visual space of the pre-trained CLIP model, aligning image features with CLIP's representations to produce high-quality, meaningful images. We then apply several GBlocks modules that further refine and enhance input features, adding visual attributes through combined convolution operations, specific feature conditioning, and residual connections. The final enhanced features are transformed into RGB images.

3.2. Discriminator ViT

The second module consists of two key components: the frozen ViT [28] and a learnable discriminator, as shown in Figure 3. The frozen ViT component uses convolution layers and a set number of transformer blocks to transform the input image and extract its features. Additionally, the learnable discriminator includes an extraction features module, which gathers image features from several ViT layers. These visual data are then compiled to assess the accuracy of the generated images. The adversarial loss, indicating quality, is calculated based on extracted features and sampled sentences. The extraction features module enables ViT to interpret complex scenes by taking features from sequential image layers. This module includes multiple Conv layers and two activation functions using the Rectified Linear Unit (ReLU) [32], as Eq. (1).

$$f(x) = max(0, x) \tag{1}$$

High-quality feature extraction in this module enables the identification of fake images. For stabilizing the training process, a loss called hinge [33] has been used. Besides, a one-way

discriminator is utilized for the same purpose. The model's formulation is shown in Eq. (2) to Eq. (9).

$$\alpha = \mathbb{E}_{x\sim\mathbb{P}_r}\left[min\left(0, -1 + D(C(x), e)\right)\right] \tag{2}$$

$$\beta = (1/2)\mathbb{E}_{G(z,e)\sim\mathbb{P}_g}\left[min\left(0, -1 + D\left(C\left(G(z,e)\right), e\right)\right)\right] \tag{3}$$

$$\gamma = (1/2)\mathbb{E}_{x\sim\mathbb{P}_{mis}}\left[min\left(0, -1 + D(C(x), e)\right)\right] \tag{4}$$

$$\delta = k\mathbb{E}_{x\sim\mathbb{P}_r}\left[\left(\left\|\nabla_{C(x)}D(C(x), e)\right\| + \|\nabla_e D(C(x), e)\|\right)^p\right] \tag{5}$$

$$L_D = -\alpha - \beta - \gamma + \delta \tag{6}$$

$$\mathbb{A} = \mathbb{E}_{G(z,e)\sim\mathbb{P}_g}\left[D\left(C\left(G(z,e)\right), e\right)\right] \tag{7}$$

$$\mathbb{B} = \lambda\mathbb{E}_{G(z,e)\sim\mathbb{P}_g}\left[S\left(C\left(G(z,e)\right), e\right)\right] \tag{8}$$

$$L_G = -\mathbb{A} - \mathbb{B} \tag{9}$$

Here, the discriminator lossL_D, integrates multiple components to assist the model in differentiating between authentic and counterfeit image-text combinations, as well as between congruent and incongruent pairs. The Gaussian distribution is used as a vector, abbreviated with z, to input to the model as input, whereas the sentence vector is marked by e. The generator ViT and discriminator ViT are denoted as G and D, respectively, with C indicating the frozen ViT in the Discriminator. α ensures that the discriminator accurately classifies real data by minimizing hinge loss, which is crucial for distinguishing real images. β improves training by lowering the hinge loss for generated data samples, helping the generator produce more realistic images, and ensuring the discriminator accurately classifies them as fake. Additionally, γ is applied in hinge loss to enhance the discriminator's ability to differentiate mismatched data samples, thus distinguishing between real and synthetic images.

The cosine similarity, denoted as S, measures the similarity between encoded visual and textual representations. Two parameters, k, and p, are used for the gradient penalty. Furthermore, the text-image similarity is evaluated using the coefficients of λ. The distributions for generated data, real data, and mismatched data are represented by $\mathbb{P}_g$, $\mathbb{P}_r$, and $\mathbb{P}_{mis}$, respectively.

GANViT, as shown in Figure 3, is trained and assessed on Caltech-UCSD Birds (CUB), which is publicly available, containing 11,788 images across 200 bird species, each with 10 detailed annotations. Common Objects in Context (COCO), a large-scale, publicly available dataset containing over 330,000 images featuring a wide variety of common, iconic objects, with detailed 5 annotations. The CC3M and CC12M datasets are large-scale, publicly available image-text datasets with 3 million and 12 million pairs, respectively, providing diverse, high-quality captions from web-sourced images.

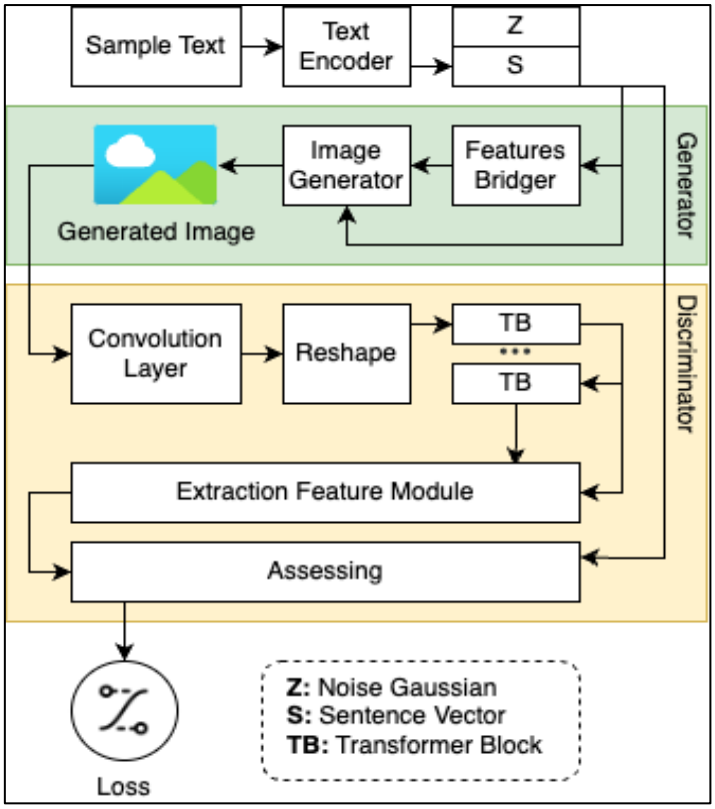


Figure 3: Details Structure of GANViT

It utilizes a CLIP-based architecture with defined hyperparameters and evaluates image fidelity using Frechet inception distance (FID) metrics. The model maintains a smooth latent space, allowing for gradual modifications in generated images based on textual inputs, and outperforms alternative models in complex image generation tasks. The key components, such as the discriminator ViT, generator ViT with feature bridge projection, and prompt tuning, as well as the choice of CLIP-ViT layers, affect the quality and complexity of the generated images.

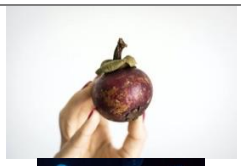
4. Datasets

There are several datasets available for image generation. The MS-COCO 2014 dataset [34] contains over 83,000 image–text pairs for training, 41,000 pairs for testing, and an additional 41,000 for validation, providing detailed information about objects, locations, and attributes at various resolutions.

The CUB-200-2011 dataset is divided into 200 bird subcategories, each image annotated with detailed information, including subcategory labels, 15-part locations (e.g., beak, wing, and tail), 312 binary attributes (e.g., migratory or ground-nesting), and bounding boxes.

The CC12M [35] dataset is designed for vision-and-language pretraining, with 12.4 million image-text pairs that feature diverse visual concepts and longer captions than CC3M. Created by relaxing filters to enhance recall, CC12M offers an expanded collection of image–text pairs, balancing precision with recall relative to CC3M. Images were sourced from the internet, with accompanying alt-text captions that are generally concise descriptions of visual content. Samples of the CC12M dataset are shown in **Error! Reference source not found.**.

Table 1: CC12M dataset examples

Sentence	Images
Hand holding a fresh mangosteen	
#jellyfish #blue #ocean #pretty Sea Turtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn and <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life.	
<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo).	

This dataset encompasses a diverse range of visual concepts with minimal biases related to gender, age, color, or ethnicity. It addresses limitations found in other datasets, such as COCO, by offering a larger volume of diverse images and text, thus enhancing training. The dataset is publicly available for non-commercial use, with paired image–text data downloadable via specified URLs.

5. Experiment and results

The proposed model, GANViT, was evaluated and trained on several datasets such as CC12M and MS-COCO.

GANViT models are trained in a cloud-based environment [36] using eight GPUs, each with 24 GB memory and 282 TFLOPs, which stands for Tera Floating Point Operations Per Second, of computational capacity. The system's CPU is an AMD EPYC 7302 16-core Processor.

The model employs ViT-B/32 in the generator and discriminator components, as shown in Figure 3. The extraction feature module comprises a single set of stacked extraction blocks, whereas the feature bridge produces 224 × 224 images using four fusion blocks and generates six additional blocks. Prompt tuning is implemented with TransBlocks in the range of two to fewer than 10. The model optimizer, Adam [37], is configured with β_1 set to 0.0 and β_2 to 0.9. Furthermore, the discriminator uses hyperparameters k and p set to values of 2 and 6, respectively.

Conversely, λ is set to 4 for the datasets used. Figure 4 displays the required time for inference in seconds. **Error! Reference source not found.** compares the performance of GANViT with other models based on FID, where our proposed model is benchmarked against several recent models [38], [39], [40], [41], [42].

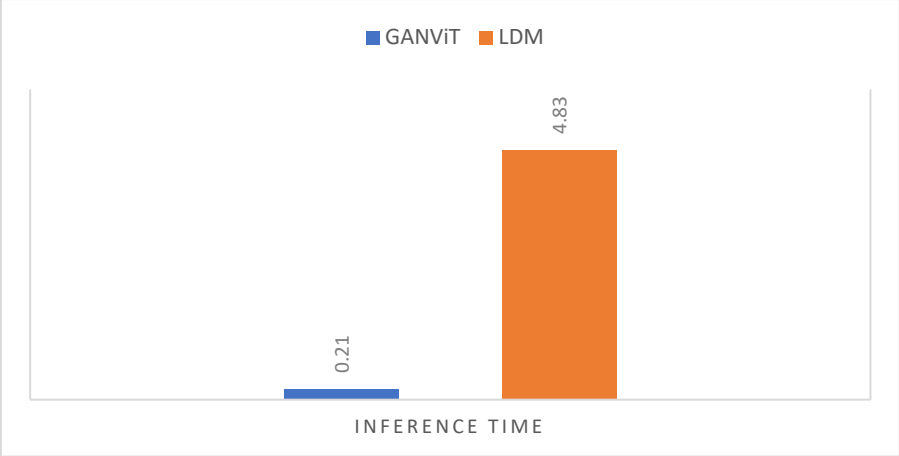
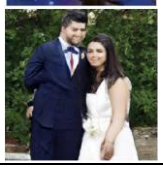


Figure 4: Required time of inferencing in seconds

Table 2: Inferencing via GANViT dataset, and LDM

Sentence	GANViT	Images
A smiling statue to be found in the Pavilion		
Watercolor of Blue Lighthouse. Lovely blue sky and dark blue lighthouse on a hillside.		
Cute cartoon little owl with big shining eyes sitting on a tree branch at starry night		
newly - weds after the wedding ceremony		

In Eq. (10), real and generated image features are represented by v and $\bar{v}$, respectively, with their mean and covariance denoted as μ and Σ. The trace of the matrix is denoted as T_r.

$$FID(v,\bar{v}) = \|\mu_v - \mu_{\bar{v}}\|^2 + T_r(\sum_v + \sum_{\bar{v}} - 2(\sum_v \sum_{\bar{v}}))^{1/2} \quad (10)$$

A lower FID score indicates greater similarity between the distributions of real and generated images in the feature space, reflecting higher image quality.

The results evaluated on the CUB-200-2011 [45] and Microsoft Common Objects in Context 2014 [36] datasets indicate a new level of improvement. Applying contrastive loss to the CUB dataset, as demonstrated in [41], the proposed model, GANViT, improved the FID from 14.58 to 10.09. Similarly, GANViT reduced the FID on the COCO dataset from 8.21 to 5.01 and from 13.86 to 10.08 on CUB, as shown in Table 3.

Table 3: FID results on COCO and CUB datasets

Model	MSCOCO2014	CUB-200-2011
[38]	32.64	16.09
[39]	28.12	15.19
[40]	19.32	14.81
[42]	8.21	14.58
[41]	10.32	13.86
Our	5.01	10.08

In contrast, [41] applies a diffusion model to both datasets with larger data, showing potential for performance gains. GANViT has limitations when synthesizing complex imaginary scenes, as shown in **Error! Reference source not found.**. Increasing dataset size and model capacity could address these limitations.

Table 4: Instances of failure in text-to-image

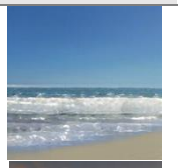
Sentence	GANViT
A lion educator dressed in a suit stands before a blackboard.	
A cute cat lives in a house made out of food.	
A robot is riding a horse under the cloudy sky.	

Conversely, short phrases or single words can introduce ambiguity. For example, using a single word, like “beach,” may result in an ambiguous daytime scene, whereas specifying “sunset at the beach” clarifies the desired outcome

Firstly, the contextual elements could be presented with sentences that enable the model to comprehend descriptions and relationships between items. Consequently, this contributes to improving the precision of image generation.

Error! Reference source not found. demonstrates the performance of GANViT with various text inputs, showing clear advantages over other methods. Sentences provide contextual detail, allowing the model to understand descriptions and relationships between objects, thus enhancing image precision.

Table 5: Text-to-Image Generation Analysis Table

Sentence		GANViT
Single Word	Beach	
Phrase	Sunset at the beach	
Sentence	A group of guys relaxing by the ocean on a sunset beach surrounded by golden sand and blue sky	

6. Conclusions

This work introduces GANViT, a new model designed for accurately generating aligned, high-quality images. Integrating ViT components in the generator and discriminator improves understanding of complex scenes, enhancing model generalization. Our model also addresses specific dataset challenges, demonstrating significant improvement in T2I synthesis. GANViT attained FID scores of 5.01 and 10.08 on the COCO and CUB-200-2011 datasets, respectively, and decreasing parameters by a factor of 4.5 and processing time by a factor of 23. As Future work, incorporating a language model in the text encoder could improve T2I generation. Additionally, training on larger datasets would increase output diversity, and supporting longer, richer descriptions could allow for more precise scene depiction.

References

[1] Y. Zhou and N. Shimada, "Vision + Language Applications: A Survey," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, May 2023, pp. 826–842.

[2] Y. X. Tan, C. P. Lee, M. Neo, K. M. Lim, J. Y. Lim, and A. Alqahtani, "Recent Advances in Text-to-Image Synthesis: Approaches, Datasets and Future Research Prospects," *IEEE Access*, vol. 11, pp. 88099–88115, 2023, doi: 10.1109/ACCESS.2023.3306422.

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

[4] H. serag Ibrahim and N. M. Shati, "A Survey on Image Caption Generation in Various Languages," *Iraqi Journal of Science*, pp. 4030–4046, Jul. 2024, doi: 10.24996/ijs.2024.65.7.38.

[5] 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.

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

[7] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, Oct. 2020.

[8] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," in *International conference on machine learning*, PMLR, May 2016, pp. 1060–1069.

[9] T. Xu *et al.*, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Nov. 2018, pp. 1316–1324. [Online]. Available: http://arxiv.org/abs/1711.10485

[10] Q. Cheng and X. Gu, "Cross-modal Feature Alignment based Hybrid Attentional Generative Adversarial Networks for text-to-image synthesis," *Digital Signal Processing: A Review Journal*, vol. 107, Dec. 2020, doi: 10.1016/j.dsp.2020.102866.

[11] M. Tao, S. Wu, X. Zhang, and C. Wang, "DCFGAN: Dynamic Convolutional Fusion Generative Adversarial Network for Text-to-Image Synthesis," in *2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, IEEE, Nov. 2020, pp. 1250–1254. doi: 10.1109/ICIBA50161.2020.9277299.

[12] H. Zhang, H. Zhu, S. Yang, and W. Li, "DGattGAN: Cooperative Up-Sampling Based Dual Generator Attentional GAN on Text-to-Image Synthesis," *IEEE Access*, vol. 9, pp. 29584–29598, 2021, doi: 10.1109/ACCESS.2021.3058674.

[13] R. Li, W. Li, Y. Yang, and Q. Bai, "Obj-SA-GAN: Object-Driven Text-to-Image Synthesis with Self-Attention Based Full Semantic Information Mining," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 339–350. doi: 10.1007/978-3-031-20862-1_25.

[14] B. Jiang, Y. Huang, W. Huang, C. Yang, and F. Xu, "Multi-scale dual-modal generative adversarial networks for text-to-image synthesis," *Multimed Tools Appl*, vol. 82, no. 10, pp. 15061–15077, Apr. 2023, doi: 10.1007/s11042-022-14080-8.

[15] Z. Qi, C. Fan, L. Xu, X. Li, and S. Zhan, "MRP-GAN: Multi-resolution parallel generative adversarial networks for text-to-image synthesis," *Pattern Recognit Lett*, vol. 147, pp. 1–7, Jul. 2021, doi: 10.1016/j.patrec.2021.02.020.

[16] M. Zhang, C. Li, and Z. Zhou, "Text to image synthesis using multi-generator text conditioned generative adversarial networks," *Multimed Tools Appl*, vol. 80, no. 5, pp. 7789–7803, Feb. 2021, doi: 10.1007/s11042-020-09965-5.

[17] N. Ge, Y. Zhu, X. Xiong, B. Zheng, and J. Huang, "KnHiGAN: Knowledge-enhanced Hierarchical Generative Adversarial Network for Fine-grained Text-to-Image Synthesis," in *Proceedings - 2021 14th International Symposium on Computational Intelligence and Design, ISCID 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 357–360. doi: 10.1109/ISCID52796.2021.00088.

[18] Z. Zhang and L. Schomaker, "DiverGAN: An Efficient and Effective Single-Stage Framework for Diverse Text-to-Image Generation," *Neurocomputing*, vol. 473, pp. 182–198, Feb. 2022, doi: 10.1016/j.neucom.2021.12.005.

[19] X. Luo, X. He, X. Chen, L. Qing, and J. Zhang, "DualG-GAN, a Dual-channel Generator based Generative Adversarial Network for text-to-face synthesis," *Neural Networks*, vol. 155, pp. 155–167, Nov. 2022, doi: 10.1016/j.neunet.2022.08.016.

[20] T. Yang, X. Tian, N. Jia, Y. Gao, and L. Jiao, "BA-GAN: Bidirectional Attention Generation Adversarial Network for Text-to-Image Synthesis," in *IFIP Advances in Information and Communication Technology*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 149–157. doi: 10.1007/978-3-031-14903-0_16.

[21] H. Sun and Q. Guo, "DSG-GAN: Multi-turn text-to-image synthesis via dual semantic-stream guidance with global and local linguistics," *Intelligent Systems with Applications*, vol. 20, Nov. 2023, doi: 10.1016/j.iswa.2023.200271.

[22] Y. X. Tan, C. P. Lee, M. Neo, K. M. Lim, and J. Y. Lim, "Enhanced Text-to-Image Synthesis with Self-Supervision," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3268869.

[23] B. Jiang, W. Zeng, C. Yang, R. Wang, and B. Zhang, "DE-GAN: Text-to-image synthesis with dual and efficient fusion model," *Multimed Tools Appl*, Mar. 2023, doi: 10.1007/s11042-023-16377-8.

[24] H. Majid and K. H. Ali, "Automatic Diagnosis of Coronavirus Using Conditional Generative Adversarial Network (CGAN)," *Iraqi Journal of Science*, pp. 4542–4556, Jul. 2023, doi: 10.24996/ijs.2023.64.7.40.

[25] A. Razavi, A. van den Oord, and O. Vinyals, "Generating Diverse High-Fidelity Images with VQ-VAE-2," *Adv Neural Inf Process Syst*, vol. 32, Jun. 2019.

[26] 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.

[27] M. Ding, W. Zheng, W. Hong, and J. Tang, "CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers," *Adv Neural Inf Process Syst*, vol. 35, pp. 16890–16902, Apr. 2022.

[28] A. Radford *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *International conference on machine learning*, PMLR, Feb. 2021, pp. 8748–8763.

[29] H. Tibebu, A. Malik, and V. De Silva, "Text to Image Synthesis using Stacked Conditional Variational Autoencoders and Conditional Generative Adversarial Networks," Jul. 2022, doi: 10.1007/978-3-031-10461-9_38.

[30] M. Tao, H. Tang, F. Wu, X. Jing, B.-K. Bao, and C. Xu, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 16494–16504. doi: 10.1109/CVPR52688.2022.01602.

[31] M. Jia *et al.*, "Visual Prompt Tuning," in *European Conference on Computer Vision*, Springer, Mar. 2022, pp. 709–727.

[32] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," *arXiv preprint arXiv:1803.08375*, Mar. 2018.

[33] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-Attention Generative Adversarial Networks," May 2018.

[34] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision--ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, May 2014, pp. 740–755.

[35] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training to Recognize Long-Tail Visual Concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Feb. 2021, pp. 3558–3568.

[36] "Rent GPUs | Vast.ai." Accessed: Aug. 31, 2024. [Online]. Available: https://vast.ai/

[37] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014.

[38] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang, "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39] S. Ruan *et al.*, "DAE-GAN: Dynamic Aspect-aware GAN for Text-to-Image Synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, Aug. 2021, pp. 13960–13969.

[40] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "DF-GAN: A simple and effective baseline for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.[41] S. Gu *et al.*, "Vector Quantized Diffusion Model for Text-to-Image Synthesis," *Computer Vision and Pattern Recognition*, 2022.

[41] Y. Zhou *et al.*, "Towards Language-Free Training for Text-to-Image Generation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2022, pp. 17886–17896. doi: 10.1109/CVPR52688.2022.01738.