Journal of Artificial Intelligence Research 82 (2025) 2453-2469 Submitted 02/2025; published 04/2025

MAT2I: Enhancing Perceptual Authenticity in Text-to-Image Synthesis Using Multi-Attribute Generative Adversarial Networks

Varsha Singh VARSHAGAUR@GMAIL.COM
IT Dept., Indian Institute of Information Technology Allahabad, Prayagraj, U.P.-211015, India.

Vijai Singh VIJAI.CS@GMAIL.COM
CSE Department, Inderprastha Engineering College, Ghaziabad, U.P.-201010, India.

Uma Shanker Tiwary UST@IIITA.AC.IN
IT Dept., Indian Institute of Information Technology Allahabad, Prayagraj, U.P.-211015, India.

Abstract

Generating visuals from text involves deriving visual representations from textual descriptions and transforming them into corresponding visuals. This technique finds vast application in various fields, such as graphic design and image editing. Generative adversarial networks (GANs) are the widely used and better performers for this task. A primary hurdle in this process is producing perceptually authentic visuals. This study introduces a Multi-Attribute Text to Image Synthesis Generative Adversarial Network (MAT2I) to address these challenges. The enhancements encompass attribute-control-net, feature alignment, and perceptual loss. The attribute-control-net is used for the fast and attribute-specific generation to maintain authenticity in perceptuality with adaptability. Feature alignment and perceptual loss motivate the generator to create visuals that closely resemble real visuals based on the accompanying text and to reduce randomness. The effectiveness of the proposed model is gauged on the CUB and COCO datasets. Empirical findings illustrate that this approach generates visuals with greater content diversity, enhanced realism, and improved semantic alignment with provided text descriptions. Furthermore, the proposed method surpasses comparative techniques in terms of inception score, further establishing its competitive performance.

1. Introduction

Visual generation from text has gained significant attention within the contemporary research landscape, aiming to create visual content from textual descriptions automatically. The primary hurdles in text-to-image synthesis involve generating visually authentic and semantically coherent visuals. Given the high-dimensional nature of image data distribution, which is difficult to directly approximate, several frameworks (Reed et al., 2016; Zhang et al., 2017; Xu et al., 2018) have been proposed to manage local and global visual information. These frameworks enhance the generation of visuals that are more visually plausible. Simultaneously, semantic consistency, which concerns aligning textual descriptions with image content is crucial.

Generative Adversarial Networks (GANs) are widely employed for text-to-image synthesis, but they come with inherent challenges, such as training instability and non-essential confidence. Taking inspiration from the accomplishments of working on attributes (Li et al., 2019) in various domains, this paper introduces a novel approach: Multi-attribute GAN (MAT2I) for visual generation from text. Integrating attribute control net (ACN) into visual generation from text contributes to the production of visually authentic visuals with adaptability. By employing an ACN, the model leverages change in attributes (Gidaris et al., 2018) to acquire diverse visual representations, thus yielding more visually plausible visuals that align better with the intended semantics.

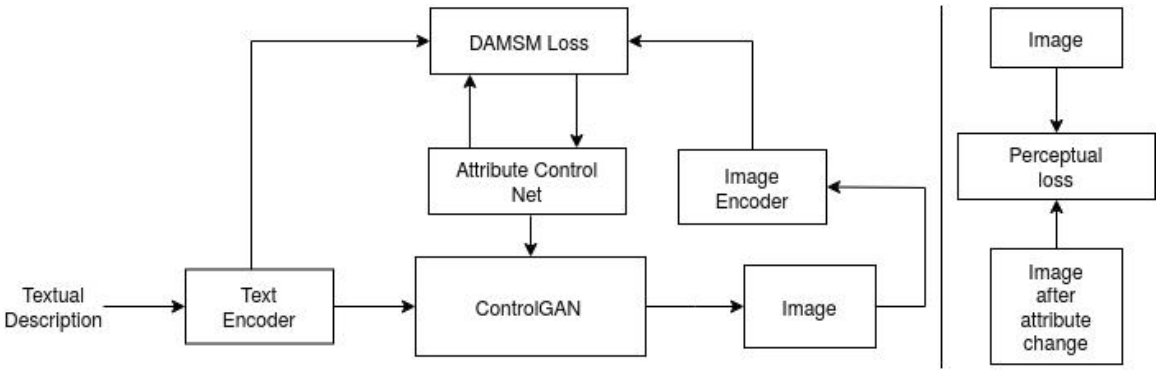

Figure 1: Architecture of MAT2I

The proposed MAT2I incorporates principal component analysis (PCA) derived direction to apply in both attribute control net and feature alignment(Härkönen et al., 2020) within the generator, enhancing its capacity to learn and adapt to a wide range of visual information facilitated by attribute control net.

Further enhancements are suggested to promote the synthesis of visually convincing and semantically congruent visuals. These enhancements encompass feature alignment (Salimans et al., 2016) and perceptual loss (Pihlgren et al., 2020). The feature alignment mechanism redirects the generator's focus from solely tricking the discriminator to creating visuals that closely resemble authentic ones. This shift improves the overall training stability of GANs. The perceptual loss function (Johnson et al., 2016) compels the generator to minimize disparities between the generated and real visuals. The contributions of this work are as follows:

- Introduction of attribute control net to provide modifications in generated visual content based on modified attribute-based textual descriptions, keeping other parts of the visual unchanged. The expanded variety of image representations subsequently enhances the discriminator's capacity for accurate classification.

- Integration of feature alignment and perceptual loss functions to redirect the generator's objective from merely outplaying the discriminator to producing visuals that closely resemble real-world counterparts. The generator's goal shifts to generating visuals that are less dissimilar from authentic ones.

- We perform extensive experiments on the CUB (Wah et al., 2011) and COCO (Lin et al., 2014) datasets. As a result, MAT2I-GAN demonstrates competitive performance.

2. Prior Related Work

This section outlines the evolution of text-to-image generation, highlighting key advancements in GAN and diffusion-based models. It also reviews existing techniques in compositional and attribute-specific image synthesis, forming the basis for the proposed MAT2I framework.

2.1 Visuals From Text

Generating visuals from text has witnessed substantial advancements, with numerous models emerging (Lee et al., 2022; Wu et al., 2022a,b; Zhou et al., 2022b). Recently, diffusion models (Nichol et al., 2021; Saharia et al., 2022; Gu et al., 2022) trained on extensive datasets have showcased remarkable potential. DALLE2 (Ramesh et al., 2022) introduces a diffusion decoder that generates images based on CLIP embeddings (Radford et al., 2021). Stable Diffusion (Rombach et al., 2022) enhances model architecture with cross-attention modules for potent diffusion-based image generation. GAN-based techniques (Qiao et al., 2019; Yin et al., 2019; Ma et al., 2020; Zhang and Schomaker, 2022; Singh and Tiwary, 2023) have spurred progress in text-to-image generation. AttnGAN (Xu et al., 2018) introduces an attention-driven multimodal similarity model for fine-grained image-text alignment, widely adopted in various GAN models (Li et al., 2019; Zhu et al., 2019; Liao et al., 2022; Wu et al., 2022c; Singh and Tiwary, 2023; Singh et al., 2023). XMC-GAN (Zhang et al., 2021) refines text-image matching through cross-modal contrastive learning. DAE-GAN (Ruan et al., 2021) incorporates sentence-level and attribute-derived information. Style-GAN (Karras et al., 2020), a successful image generation framework, extends to text-to-image synthesis. Our approach, MAT2IGAN, takes a different approach by incorporating principal component analysis (PCA) derived direction to apply in both attribute control net and feature alignment.

2.2 Visual Generation in Multidomain

Compositional image generation is studied in a benchmark (Park et al., 2021), evaluating attribute-based text-to-image models. StyleT2I (Li et al., 2022) incorporates a spatial constraint loss to disentangle attribute features by constraining spatial variance according to input attributes. However, these works often overlook the attribute-specific alteration without changing the image's other parts and keeping the background in consideration. Our framework, MAT2I, works on the PCA-derived direction for modifications by selectively applying them to specific layers and enhancing attribute compositional generalization by perceptual loss. Effective text representations are crucial for text-to-image synthesis. AttnGAN (Xu et al., 2018) employs an attention mechanism for fine-grained word-subregion connections, while ControlGAN(Li et al., 2019) uses perceptual loss to reduce randomness. GANSpace(Härkönen et al., 2020) contributes to text-to-image tasks, with prior studies (Li et al., 2022; Brock et al., 2018; Ramesh et al., 2022) showcasing its potent language-and-

vision feature space. Inspired by this, we use an attribute control net (ACN) to incorporate modifications done in attributes to visuals.

3. Visual Generation Using Multi-Attribute GAN

Drawn from the encouraging results of ControlGAN(Li et al., 2019), the proposed MAT2I method adopts the core concepts while incorporating several refinements. These refinements encompass attribute control net, feature alignment (Salimans et al., 2016) and perceptual loss (Pihlgren et al., 2020).

The attribute control net empowers the existing GAN to make modifications in generated visuals by changing only a few attributes using PCA-driven direction and keeping the others unaffected. PCA is utilized to compute the principal direction matrix p_x to guide attribute-specific modifications in the Attribute Control Net (ACN). It provides a mathematically grounded method to identify and manipulate the most significant directions of variation in the latent space. This, in turn, elevates the discriminator's learned representations, thereby spurring the generator to produce a more diverse set of images.

Feature alignment comes into play to align the characteristics of genuine and generated images. Doing so facilitates the synthesis of visuals that exhibit greater photorealism and a more extensive spectrum of content variations.

The integration of the perceptual loss functions to narrow the gap between authentic and synthesized images. This serves to lessen the unpredictability in the generation process and compels the generator to maintain the visual characteristics associated with the original text and semantic consistency.

These augmentations collectively ensure that the MAT2I method generates visually realistic images, semantically coherent and subsidised with a richer array of features. Furthermore, these enhancements contribute to alleviating the challenges of overconfidence and training instability encountered in visual from text generation.

3.1 Architecture

The proposed model aims to generate a realistic image, I_0, that semantically matches a given text description T. Furthermore, the model allows controlled modifications to the generated image based on changes in the text description. For example, if the original text T is modified to a new description T_{new}, the model generates a new image $\bar{I}_0$ that reflects the changes described in T_{new}. However, keeping other parts of the original image I_0 unchanged.

Our fundamental architectural framework is built upon the multistage ControlGAN (Li et al., 2019), which serves as the core structure (refer to Fig. 1). MAT2I's architecture consists of the following major components:

- **Text Encoder:** Extracts sentence-level and word-level features from the input text.
- **Attribute Control Net (ACN):** Leverages PCA-derived directions to enable precise attribute-specific modifications in the latent space.
- **Multi-Stage Generator:** A three-stage generator progressively refines the image from coarse to fine resolution.

- **Feature Alignment and Perceptual Loss:** Enhances visual realism by aligning generated images with real-world features and reducing noise.

Given an input sentence T, we employ a pre-trained bidirectional RNN (Xu et al., 2018) as the text encoder. This RNN encodes sentence T into a comprehensive sentence feature denoted as "t" in the D-dimensional space, providing an overall description of the sentence. Additionally, it generates the word feature "w" in a D×L-dimensional space, where L represents the word count, and D is the dimensionality. Building upon (Zhang et al., 2017), we also implement conditioning augmentation (CA) on "t". The augmented sentence feature, denoted as "t^", is combined with a random vector "z" to construct the input for the initial stage.

To ensure semantic alignment between text and generated visuals, we employ the Deep Attentional Multimodal Similarity Model (DAMSM) from Xu et al., 2018. DAMSM extracts text and image features and aligns them using a multimodal attention mechanism, which matches word-level features with corresponding image regions. By computing a similarity score using cosine similarity, DAMSM encourages the generated image to align semantically with the input text. This integration guides the generator to produce images that are not only visually realistic but also coherent with the input text's semantics.

In addition, the principal direction matrix p_x provides information to GAN through ACN regarding attributes or feature control. The latent direction for the principal component is computed using:

$$P = arg\,min \sum_{k} ||P_{x_k} - z_k||^2 \tag{1}$$

The computed coordinates are transformed into activations (y_k) through linear regression. These activations are then provided as input to the generator to guide the desired attribute alterations. By computing the coordinates, the ACN effectively translates abstract attribute changes into actionable modifications within the latent space, enabling controlled and semantically coherent visual generation.

The entire process of image generation follows a multi-stage approach, progressively generating images from coarse to fine levels. At each stage, the network generates a disguised visual feature "v_n", which acts as the input for the respective generator "G_n" to produce a synthetic image. To enhance the process, spatial attention (Xu et al., 2018) and a channel-wise attention module (Li et al., 2019), both of which utilize "w" and "v_n" to yield attentive word-context features. These attentive features are merged with the concealed feature "v_n" and fed into the subsequent stage with p_x as controlling feature input, which is activated when a new textual description is taken for generation.

Our novel ACN computed the coordinates of each feature tensor using random latent vectors, z_k and then converted to activations, y_k, transferred this to latent space by linear regression, which is given as input to the existing GAN structure to control the required feature alteration. Basically, to enhance control over attribute modifications during image generation, PCA is employed to compute a principal direction matrix. This matrix captures the primary directions of variation in the latent space, allowing for fine-grained adjustments to specific attributes while maintaining the overall realism of the generated images. Then, perceptual loss is used to compute the loss between the generated and previously generated

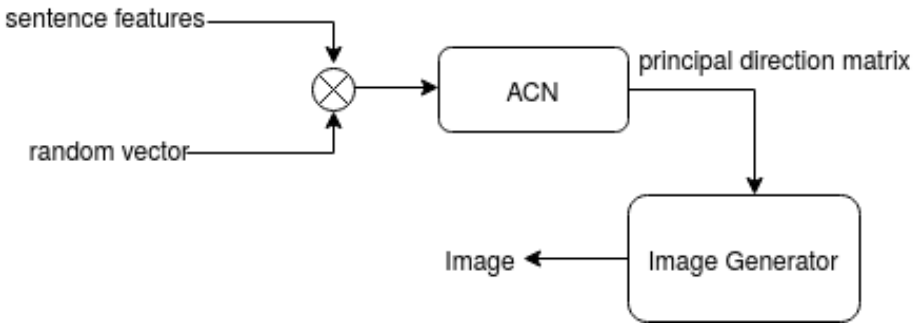

Figure 2: Generator Architecture. $\otimes$ of sentence feature and random vector denote concatenation

visual. The perceptual loss is used to make the image less noisy in less important areas like the background, unlike in controlGAN(Li et al., 2019), using the given equation:

$$\text{Perceptual Loss}, L_{\text{perceptual}} = \sum_{l} m_l \cdot \|\phi_l(G(\bar{I}_0)) - \phi_l(\bar{I}_0)\|^2 \tag{2}$$

where, $\sum$ denotes the sum over multiple layers in a pre-trained deep neural network, m_{l}represents the weight assigned to each layer l, $\phi_l(G(\bar{I}_0))$ is the feature representation of the generated visual $G(\bar{I}_0)$ extracted from layer "l" of the neural network. $\phi_l(\bar{I}_0)$ is the feature representation of the manipulated feature visual $(\bar{I}_0)$ extracted from the same layer "l" of the neural network.

3.2 Objective Functions

The model's objective function can be understood as a minimax game involving two participants who compete against each other, and it is formulated as follows:

$$\min_G \max_D O(D, G) = log\left[D(x, t)\right] + log\left[1 - D(G(z, t))\right] \tag{3}$$

3.2.1 Generator Objective

The generator loss, denoted as L_{Gen} in the given equation, encompasses several components. These include an adversarial loss referred to as L_{Gadv}, a loss promoting correlation between text and images known as L_{cor}, a perceptual loss termed L_{per}, and finally, a loss ensuring alignment between text and images called L_{DAMSM}(Xu et al., 2018).

$$L_{Gen} = \sum_{s=1}^{S}(L_{Gadv} + \lambda_2 L_{per} + \lambda_3 log(1 - L_{cor})) + \lambda_4 L_{DAMSM}, \tag{4}$$

Where 'S' denotes the number of stages, from the initial to the s^{th} stage of visuals. The hyperparameters λ_2, λ_3, and λ_4 regulate distinct losses. Specifically, L_{per} corresponds to the perceptual loss (Li et al., 2019), which is different from $L_{prceptual}$ in Eq 2. This loss

enforces constraints on the generative process to mitigate unpredictability. The L_{DAMSM} technique is adopted to gauge text-image correspondence, assessing matching through cosine similarity. Additionally, L_{cor} evaluates the correlation between the generated image and the provided textual description while accounting for spatial details.

3.2.2 Discriminator Objective

The loss function for the training of the discriminator includes the L_{cor}, which stands as the correlation loss, which assesses the presence of visually related attributes to words within the image. The text description that does not align with the visual is chosen randomly from the dataset and holds no relevance to computing this loss. λ_1 serves as a hyper-parameter that governs the significance of supplementary losses.

$$L_{Dis} = \sum_{s=1}^{S}(L_{Dadv} + \lambda_1(log(1 - L_{cor}) + log(L_{cor})), \tag{5}$$

The adversarial loss L_{Dadv} comprises dual elements: the unconditional adversarial segment gauges the authenticity of the image, while the conditional adversarial segment evaluates if the provided image aligns with the text description T.i.e.,

$$L_{Dadv} = L_{unconditional} + L_{conditional} \tag{6}$$

To determine the semantic consistency and visual authenticity of the input pair effectively, the discriminator must categorize three distinct pairs of samples:

1. A text description paired with the corresponding real image.
2. A text description paired with an unrelated real image.
3. A text description paired with a synthesized image.

Among these pairs, (1) represents the authentic sample, while (2) and (3) pertain to fabricated samples. By utilizing these sample pairs as input, the discriminator can grasp the correlation between the text description and the image. In particular, through pairs (1) and (3), the discriminator acquires the ability to distinguish the authenticity of the input image. Similarly, by examining pairs (1) and (2), the discriminator learns how to associate the image with the text description accurately.

4. Experiments

To assess the efficacy of the proposed MAT2I model, we conduct thorough experimentation using the CUB (Wah et al., 2011) and COCO (Lin et al., 2014) datasets as benchmarks. Our comparative analysis includes both GAN-based and diffusion-based text-to-image generation techniques to ensure a comprehensive evaluation.

For GAN-based models, we compare MAT2I with three widely recognized text-to-image synthesis methods: StackGAN (Zhang et al., 2017), StackGAN++ (Zhang et al., 2018), and AttnGAN (Xu et al., 2018). These models represent key advancements in GAN-based image generation, and we reproduce their results using the source code provided by their respective authors.

Given the recent success of diffusion-based models, we further acknowledge and discuss the advancements made by models such as DALL-E (Reddy et al., 2021), LAFITE (Zhou et al., 2022a), and Stable Diffusion (Rombach et al., 2022). While our primary focus remains on GAN-based approaches, we include a comparative discussion on diffusion-based methods to highlight key differences in architecture, computational complexity, and output quality.

This extended evaluation allows us to place MAT2I within the broader landscape of text-to-image synthesis, demonstrating its strengths in semantic alignment, attribute control, and computational efficiency compared to both GAN and diffusion-based methods.

4.1 Dataset

In this study, the proposed MAT2I model is evaluated on two benchmark datasets, CUB-200-2011 (CUB) and COCO (Common Objects in Context), to assess its performance in generating realistic and semantically aligned visuals. These datasets offer complementary challenges: CUB focuses on fine-grained object details, while COCO presents a diverse set of real-world scenes with multiple objects and complex textual descriptions.

Before diving into dataset-specific details, Table 1 provides an overview of the training and testing splits used in our experiments:

Table 1: Data Split for CUB and COCO datasets

Dataset	Training Split	Testing Split	Captions per Image
CUB	8,855	2,933	10
COCO	82,783	40,504	5

4.1.1 CUB-200-2011 (CUB)

The CUB dataset comprises 11,788 high-quality bird images, each paired with 10 detailed text descriptions. It is widely used in text-to-image synthesis due to its fine-grained object categories and detailed annotations. These annotations include specific body parts, colors, and patterns, allowing models to learn intricate semantic relationships between textual descriptions and visual features.

A distinct characteristic of the CUB dataset is its ability to capture wide variations in poses, backgrounds, and viewpoints across different bird species. This diversity, combined with the rich attribute annotations, facilitates nuanced analysis and attribute-driven learning approaches, making it an ideal benchmark for evaluating attribute-conditioned image generation.

4.1.2 COCO (Common Objects in Context)

The COCO dataset consists of over 80,000 images, each depicting diverse, real-world scenes with multiple objects. Unlike CUB, which focuses on single-object images with detailed attributes, COCO presents a greater challenge due to its cluttered backgrounds, highly diverse object categories, and complex textual descriptions. Each image is annotated with five different captions, making it a robust benchmark for evaluating MAT2I's ability to generalize across complex scene compositions and varied textual inputs.

By leveraging these complementary datasets, we assess MAT2I's effectiveness in fine-grained attribute synthesis (CUB) and complex scene generation (COCO), demonstrating its versatility and robustness across varying domains.

4.2 Implementation Details

The proposed MAT2I model consists of a three-stage generator (S = 3), where images are progressively refined at resolutions of 64×64, 128×128, and 256×256. Attribute Control Net (ACN), feature alignment, and spatial and channel-wise attention mechanisms are incorporated during stages 2 and 3 to enhance semantic consistency and fine-grained visual details. These mechanisms help guide the generator towards attribute-specific modifications, ensuring that the generated images align with the provided text descriptions.

A pre-trained bidirectional LSTM is employed to encode text descriptions. This text encoder translates the input text into a sentence feature with a dimension of 256, along with word features of length 18, each having a dimension of 256. These textual features are then used to guide the image generation process at multiple levels.

For the perceptual loss component, the content loss is calculated using the ReLU2_2 layer of an image encoder pre-trained on the ImageNet dataset (Russakovsky, 2015). This helps improve the perceptual realism of generated images by aligning their feature representations with real images.

To ensure stable training and optimized learning, we use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0002. The hyperparameters λ_1, λ_2, λ_3, and λ_4 are set to 0.5, 1, 1, and 5, respectively, to balance the different loss components.

4.3 Evaluation Metric

To quantitatively evaluate the generated images, we employ the Inception Score (IS), a widely used metric in GAN-based image synthesis. The Inception Score assesses two critical aspects of generated images:

1. Image Quality: The model should generate sharp, high-resolution images that closely resemble real samples.

2. Image Diversity: The generated images should be visually distinct and represent the full range of possible image classes.

The Inception Score is computed using an Inception-v3 network, which predicts class labels for the generated images. The metric is defined as:

$$\mathrm{IS} = \exp\left(E\left[\mathrm{KL}\left(P(y|x)\|P(y)\right)\right]\right) \tag{7}$$

Where,

- E denotes the expectation over generated images.
- $P(y|x)$ represents the conditional class distribution, which is the output of the Inception-v3 network when given a generated image x.
- P(y) is the marginal class distribution, which is the output of the Inception-v3 network when given real images from the dataset.

Table 2: Inception Score comparison on CUB and COCO datasets

Model	Inception Score (CUB)	Inception Score (COCO)
StackGAN	3.70	8.45
StackGAN++	3.82	8.92
AttnGAN	4.36	13.02
MAT2IGAN (Ours)	4.59	14.35

- KL divergence measures the difference between these two distributions, ensuring that generated images belong to well-defined, diverse categories.

A higher Inception Score indicates that the model produces high-quality and diverse images, while a lower score suggests mode collapse (i.e., the generator produces repetitive or low-diversity outputs).

The comparative Inception Scores on the CUB and COCO datasets are reported in Table 2, demonstrating that MAT2I achieves higher scores than previous GAN-based models:

4.4 Computational Cost and Complexity

The computational cost of MAT2I arises from its multi-stage generator, attention mechanisms, and additional loss functions, which contribute to improved image quality and semantic alignment. Below, we discuss the key factors affecting computational complexity and justify their inclusion.

4.4.1 Multi-Stage Generator

MAT2I employs a three-stage generator, progressively refining images from 64×64 to 256×256 resolution. Each stage enhances visual details and applies attribute modifications, ensuring high-quality synthesis.

Computational Complexity: The computational cost increases linearly with the number of stages. However, this approach significantly reduces mode collapse, stabilizing training and yielding more realistic outputs.

4.4.2 Attention Mechanisms

The spatial and channel-wise attention modules align text features with image regions, improving semantic consistency.

Computational Complexity: These mechanisms involve matrix multiplications, which add a computational overhead. However, their role in refining image details and improving text-image alignment outweighs this cost.

4.4.3 Attribute Control Net (ACN)

ACN utilizes PCA-derived latent directions to enable precise attribute-specific modifications.

Computational Complexity: The PCA transformation and feature regression introduce a lightweight computational cost, but since ACN works without additional retraining, it remains computationally efficient.

4.4.4 Additional Loss Functions

Perceptual loss and feature alignment improve training stability and realism.

Computational Complexity: Feature extraction from a pre-trained network (e.g., ImageNet-based encoder) slightly increases training cost but has no impact on inference speed.

4.4.5 Training and Inference Efficiency

Compared to single-stage GANs (e.g., StackGAN, AttnGAN), MAT2I requires 1.5× to 2× more training time due to multi-stage generation and additional loss terms. Inference remains efficient, as ACN transformations and attention-based conditioning operate within the existing GAN pipeline.

4.4.6 Trade-Off Justification

While MAT2I introduces additional computations, these enhancements directly contribute to improved training stability, semantic alignment, and visual quality. The model strikes a balance between computational efficiency and generation quality, making it a viable alternative to both traditional GANs and high-cost diffusion models.

4.5 Significance of ACN and Perceptual Loss

This section evaluates the contribution of the Attribute Control Net (ACN) and perceptual loss in enhancing the image generation process.

The perceptual loss ensures that less important areas, such as the background, are less noisy, while foreground elements are preserved with high fidelity. The ACN enables precise attribute modifications, improving semantic alignment between the generated image and the input text description.

To assess their impact, we conduct an ablation study by evaluating Inception Scores under three different conditions:

- With perceptual loss (L_per)
- With both ACN & perceptual loss
- Without ACN & perceptual loss

These results indicate that incorporating both ACN and perceptual loss leads to a notable increase in the Inception Score, improving image quality and diversity. The score rises from 4.52 ± 0.03 to 4.59 ± 0.05 (CUB) and from 13.89 ± 0.04 to 14.35 ± 0.07 (COCO), showcasing the effectiveness of ACN-driven attribute control and perceptual learning (see Table 3).

This improvement suggests that integrating attribute-based supervision in MAT2I enhances the semantic richness and variability of generated images, making them more realistic and aligned with textual descriptions.

Table 3: Significance of ACN and perceptual loss on CUB and COCO datasets

Score	**With L_{per}**	**With ACN & L_{per}**	**Without ACN & L_{per}**
Inception Score (CUB)	4.52 ± 0.03	4.59 ± 0.05	3.93 ± 0.03
Inception Score (COCO)	13.89 ± 0.04	14.35 ± 0.07	13.01 ± 0.02

5. Results

The proposed MAT2I model has been implemented in Python using the PyTorch framework. We evaluated its performance on both the CUB-200-2011 and COCO datasets to assess its ability to generate realistic and semantically aligned images.

5.1 Implementation Overview

The CUB dataset consists of 11,788 images, out of which 8,855 images were used for training and 2,933 images for testing. The COCO dataset, containing over 120,000 images, was split into 82,783 training images and 40,504 test images.

Our model optimizes the generator loss using cross-entropy loss, along with DAMSM loss and perceptual loss, which are crucial for improving text-image alignment and perceptual quality. These losses collectively contribute to the overall loss function, ensuring that generated images preserve semantic attributes while reducing artifacts and inconsistencies.

5.2 Word-Level Attention and Progressive Image Generation

The MAT2I framework incorporates word-level attention to enhance text-image alignment at different stages of generation (Fig. 3).

- Word-level attention mechanisms allow the model to focus on relevant parts of the text, ensuring that attributes described in the text are accurately reflected in the generated image.
- This ensures better semantic coherence, when generating fine-grained object details.

Figure 3: Word level attention

The MAT2I generator also operates in three stages, progressively refining image quality:

- In the first stage, the image is generated conditioned only on sentence features.

- In the second and third stages, the generator receives input from the fusion of word-level features and the feature map from the previous stage.
- This progressive refinement leads to better-structured, more detailed, and visually coherent images with each stage.

Figure 4: Generated image at each phase of generation

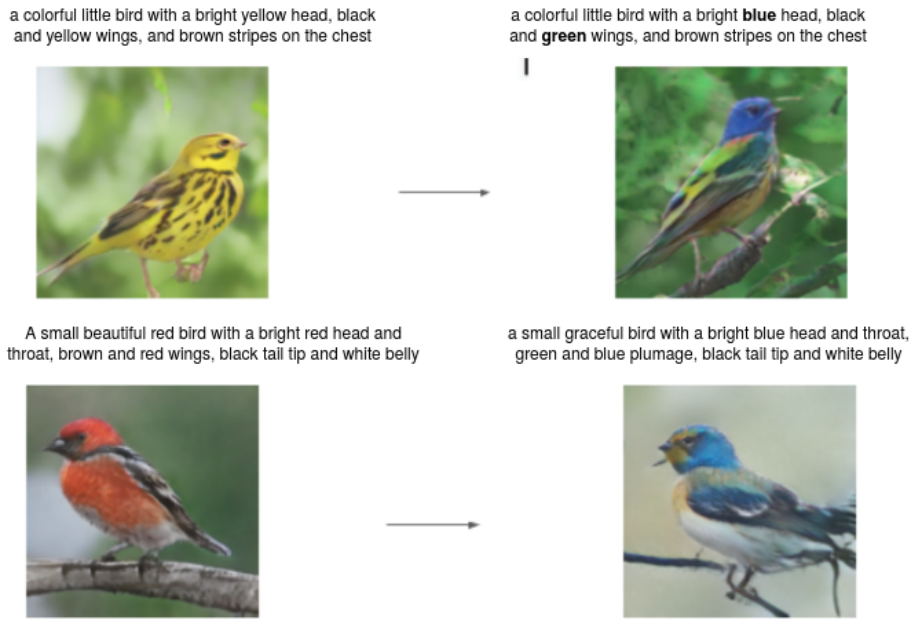

Figure 5: Visuals with change in attributes

Figure 4 illustrates how image quality improves across different stages, demonstrating enhanced resolution, texture details, and feature clarity. Figure 5 highlights the effectiveness of MAT2I in attribute modification, where changes in textual descriptions are accurately reflected in the generated visuals.

6. Conclusion

The proposed work introduces an enhanced method for visual generation from text descriptions. To achieve this, three improvements have been incorporated. Firstly, the approach integrates an attribute control net to control the feature that needs modification. This empowers the discriminator to grasp a wider range of variations in image content. Consequently, both the discriminator's classification ability and the generator's performance are enhanced, generating more diverse images.

Second, the feature-alignment technique with an objective function strives to increase the similarity between real and generated images. This tackles the issue of mode collapse, leading to a more diverse array of image content and contributes to the stability of the training process. Third, the perceptual loss motivates the generator to create visuals that closely resemble real visuals based on the accompanying text and to reduce randomness.

Experimental results obtained from the CUB and COCO datasets validate the effectiveness of the proposed method, MAT2I. In comparison to other techniques, MAT2I demonstrates superior performance, as indicated by higher inception scores.

Acknowledgments

The authors are thankful to SILP Lab, Center for Cognitive Computing, Indian Institute of Information Technology Allahabad, Prayagraj, India, for providing the necessary facilities and support.

References

Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Gidaris, S., Singh, P., and Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. (2022). Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. (2020). Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. (2022). Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532.

Li, B., Qi, X., Lukasiewicz, T., and Torr, P. (2019). Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32.

Li, Z., Min, M. R., Li, K., and Xu, C. (2022). Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18197–18207.

Liao, W., Hu, K., Yang, M. Y., and Rosenhahn, B. (2022). Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18187–18196.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Ma, J., Xu, H., Jiang, J., Mei, X., and Zhang, X.-P. (2020). Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995.

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

Park, D. H., Azadi, S., Liu, X., Darrell, T., and Rohrbach, A. (2021). Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Pihlgren, G. G., Sandin, F., and Liwicki, M. (2020). Improving image autoencoder embeddings with perceptual loss. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Qiao, T., Zhang, J., Xu, D., and Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural

language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Reddy, M. D. M., Basha, M. S. M., Hari, M. M. C., and Penchalaiah, M. N. (2021). Dall-e: Creating images from text. *UGC Care Group I Journal*, 8(14):71–75.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ruan, S., Zhang, Y., Zhang, K., Fan, Y., Tang, F., Liu, Q., and Chen, E. (2021). Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13960–13969.

Russakovsky, Olga, J. D. H. S. J. K. S. S. S. M. Z. H. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Singh, V., Gupta, S., and Tiwary, U. S. (2023). Hgan: Editable visual generation from hindi descriptions. In *International Conference on Intelligent Human Computer Interaction*, pages 3–14. Springer.

Singh, V. and Tiwary, U. S. (2023). Visual content generation from textual description using improved adversarial network. *Multimedia Tools and Applications*, 82(7):10943–10960.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., and Duan, N. (2022a). Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer.

Wu, F., Liu, L., Hao, F., He, F., and Cheng, J. (2022b). Text-to-image synthesis based on object-guided joint-decoding transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18113–18122.

Wu, X., Zhao, H., Zheng, L., Ding, S., and Li, X. (2022c). Adma-gan: Attribute-driven memory augmented gans for text-to-image generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1593–1602.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., and Shao, J. (2019). Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2327–2336.

Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., and Yang, Y. (2021). Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962.

Zhang, Z. and Schomaker, L. (2022). Divergan: An efficient and effective single-stage framework for diverse text-to-image generation. *Neurocomputing*, 473:182–198.

Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., and Sun, T. (2022a). Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17907–17917.

Zhou, Y., Zhang, R., Gu, J., Tensmeyer, C., Yu, T., Chen, C., Xu, J., and Sun, T. (2022b). Tigan: Text-based interactive image generation and manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3580–3588.

Zhu, M., Pan, P., Chen, W., and Yang, Y. (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810.