RESEARCH PAPER

Semantic-Aware Latent Diffusion With Text-Guided Data Augmentation For Text-To-Image Generation

Sonal Ajay Bankar [1]*, Satish Ket [2]

[1]*Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Lokmanya Tilak College of Engineering, Mumbai, India*
[2]*Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, India*

ABSTRACT
In this era, the Text-to-Image (TTI) generation has attained notable evolution. However, the TTI generation-based approaches face critical challenges in ensuring semantic consistency, high visual fidelity, efficient data utilization and fast generation speed. Specifically, semantic drift between textual descriptions and generated images has limited robustness under data imbalance and high computational cost during inference remain unresolved. To resolve these challenges, this research proposes a Semantic-Aware Latent Diffusion framework with Text-Guided Data Augmentation (SALD-TGA) for enhanced TTI generation. In particular, the proposed SALD-TGA framework presents a semantic-aware text encoding strategy that disentangles attribute, object and context-level information, which allows precise text-image alignment. Subsequently, to improve data efficiency and semantic diversity, a novel text-guided latent data augmentation mechanism is incorporated, which performs attribute-consistent perturbations in the latent space. Furthermore, an accelerated latent diffusion generator is designed to significantly decrease inference time while handling high-resolution and structurally coherent image synthesis. Therefore, results demonstrate the robustness of integrating semantic-aware representation learning, latent-space augmentation and fast diffusion sampling for next-generation TTI synthesis systems. The proposed SALD-TGA framework, in comparison with Context-Aware Generative Adversarial Network (CA-GAN) has acquired 9.36 Fréchet Inception Distance (FID) and 5.48 ± 0.06 Inception Score (IS) with respect to CUB-200-2011 dataset
Keywords: : Data augmentation, latent diffusion, semantic-aware learning, text-guided generation,Text-to-image synthesis.
How to cite this article: Bankar SA, Ket S, Semantic-Aware Latent Diffusion With Text-Guided Data Augmentation For Text-To-Image Generation .Int J Drug Deliv Technol. 2026;16(2s): 288-299; DOI: 10.25258/ijddt.16. 288-299
Source of support: Nil.
Conflict of interest: None

INTRODUCTION

In recent years, text-to-image generation has rapidly evolved with diffusion models, where images are created by iteratively refining random noise into meaningful visual representations guided by text prompts. Primarily, these diffusion models allowed users to input any text prompt for obtaining a desired output with enhanced quality [1]. Sequentially, modern diffusion models such as Stable Diffusion and Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers (eDiff-I) are also used, which optimise billions of parameters to produce high-quality images for the input text. [2]. Consequently, Artificial Intelligence (AI) has significantly advanced text-to-image generation by combining natural language processing with computer vision to synthesize images that align with user-provided textual descriptions. [3]. Subsequently, recent advancements in text-to-image generation have gained great attention in zero-shot text-to-3D generation, where natural language prompts are used to get a desired 3D image model [4]. However, text-to-image generation requires advanced levels of creativity and divergence and incorporation of natural language processing and computer vision [5]. Furthermore, user choices are diverse and complex, which makes it difficult to illustrate the desired image using limited keywords [6]. Subsequently, generating images from text with high quality requires high computational energy and is computationally costly. Consequently, converting text to image is challenging for memory modules that mimic the brain's ability to store, access and utilize information over time, which are becoming ineffective in capturing long-term dependencies [7]. Also, an attention mechanism focuses on the relevant part of the input data, thereby improving the text-to-image generation [8]. Additionally, images can be generated by matching the image with a specific word or phrase by achieving semantic alignment [9]. Sequentially, modern diffusion models incorporate fine-grained text–image alignment, where linguistic concepts are mapped to corresponding visual features, thereby improving the fidelity of images generated from input text. [10].

Further, prompt engineering technique, which is a skill of refining instruction to guide big Large Language Models (LLM's) to obtain the desired image from the input text [11]. State-of-the-art methods for text-to-image generation include Fast Composer, which is finetuning free, modified

**Author for Correspondence: Sonal Ajay Bankar*

multi subject text-to-image framework that employs a vision encoder along with a localized attention mechanism to generate high-quality composite images. Nevertheless, the method is unable to generate images that consist of more than three subjects [12]. Subsequently, RSDiff, a cascaded diffusion-based text-to-image model, has been introduced to generate highly realistic images from textual descriptions. Nonetheless, its inference process is computationally expensive, resulting in longer image generation time. [13]. Thereafter, SwinV2-Imagen, a text-to-image synthesis model integrated with a transformer, is utilized to produce effective and high-quality images. Although the model struggles with complex text description, it is unable to catch the semantic meaning [14]. Later, a Dynamic Local Affine Generative Adversarial Network (DLA-GAN) mechanism was incorporated to capture the key attributes for generating a high-quality image from the given input text. But the mechanism needs a wide range of tuning, and it also has the disadvantage of the vanishing gradient problem [15]. Recent advances in text-to-image synthesis have primarily focused on GAN based frameworks because of their capability in generating realistic images from the input text. Further, recent studies have demonstrated diverse frameworks with distinct methodological contributions for improving the quality of generated images; some are discussed in this section. To begin with, a Multi-level Text-driven Fine-grained Image Generation approach based on GAN, namely ML-GAN was introduced to improve text–image semantic consistency while enhancing the visual realism of the synthesized images. [16]. Here, the encoding of sentence-level textual features was carried out to generate a coarse image outline using a single-stage GAN framework. Then, attribute-level information extracted by a text parsing algorithm was combined with image features by utilization of a Dual-level Text Parallel Fusion Module (DPFM) for local details refinement.

Subsequently, the Triple-level Text Parallel Fusion Module (TPFM) integrates multi-level word features through parallel fusion, improving fine-grained image attributes and text–image consistency. In addition, a Cross-modal Attention Fusion Module (CAFM) was embedded in the discriminator for improving text-image matching. The ML-GAN framework enhanced the image quality on the CUB and COCO datasets. However, the model depended on accurate attribute extraction and irrelevant textual attributes that influence fine-detail refinement. Following this, a Context-Aware Generative Adversarial Network (CA-GAN) was determined for improving language-vision matching in text-to-image synthesis [17]. Initially, sentence embeddings extracted using a pre-trained text encoder and random noise vectors were given to a single-stage generator. Then, a Context-Aware Block (CA-Block) incorporating Context-Aware Conditional Batch Normalization (CACBN), residual connections and up-sampling was employed to effectively integrate textual semantics with image features. An Attention Convolution Module (ACM) was introduced to illustrate important visual attributes by suppressing irrelevant features. In the discriminator, mixed self-attention and convolution mechanisms are used for the enhancement of semantic consistency among generated images and text descriptions. The CA-GAN described enhanced visual detail and semantic alignment with decreased computational cost. However, the approach was evaluated on standard datasets and did not address complex multi-object scene generation. Similarly, a self-attention-based text encoder integrated with a modified Dynamic Memory GAN, namely DMGAN+MHA, was developed to mitigate information loss during text-to-image generation. [18]. Initially, textual descriptions were encoded by utilising a multi-head self-attention mechanism rather than a conventional Bi-LSTM for long-range dependencies and important keywords protection. The enhanced text embeddings were then fed to the DMGAN framework, which chose relevant words in different image refinement stages using a memory writing gate. However, the presented model focused on text encoding improvement and did not modify the core generator-discriminator architecture.

Another related work presented a single-stage GAN framework monitored by Masked Cross-Attention and Multi-Head Channel Attention for generating high-resolution images directly from text descriptions [19]. Initially, textual and visual features were combined in different scales using masked cross-attention for computational overhead reduction by preventing semantic relevance. Then, multi-head channel attention was implemented for highlighting text-related visual features and removing irrelevant information. A discriminator-based semantic consistency loss was developed for enhancing the consistency between text and generated images. The model prevented stacked generator-discriminator structures and obtained improved IS scores on both CUB and COCO datasets compared to ControlGAN. However, the masking strategy faced the problem of the model's ability to capture extremely fine-tuned global dependencies in highly complex scenes. In the same direction, a Sentence-Word Fusion Generative Adversarial Network (SWF-GAN) was recommended to address semantic inconsistency in text-to-image generation [20]. Initially, textual descriptions were encoded into sentence-level and word-level features by a pre-trained Bi-LSTM text encoder. Then, a Sentence Word Fusion Perceptual Module implemented affine transformations for guiding image synthesis from coarse sentence-level constraints to fine word-level adjustments. In addition, a weakly supervised Coordinate Mask Predictor depending on coordinate attention was established for capturing long-range dependencies and highlighting objects in complex backgrounds. However, the model relied on pre-trained text encoders and did not explore transformer-based textual representations.

Furthermore, a related work was presented in which a Cross-Domain Feature Fusion Generative Adversarial Network (CF-GAN) was introduced for improving semantic embedding and fine-tuned quality in text-to-image synthesis [21]. Initially, text descriptions were encoded by utilizing an LSTM-based text encoder to obtain sentence-level and word-level features and a low-resolution image

was produced in the first stage. Then, a Feature Fusion-Enhanced Response Module (FFERM) was implemented to deeply combine word-level vectors with image features after a Multi-Branch Residual Module (MBRM) to improve contour completeness and texture richness while fine-tuning image generation. The CF-GAN was evaluated on the CUB and MS-COCO datasets using the Inception Score (IS). The results demonstrated that CF-GAN achieves improved image realism and text–image semantic consistency when compared with AttnGAN and DM-GAN. However, CF-GAN adopted a multi-stage generator architecture that increased training complexity and computational cost. Thereafter, a Pretrained Model-based Generative Adversarial Network (PMGAN) was suggested for text-to-image generation with enhanced semantic consistency and image quality [22]. Initially, a CLIP text encoder was used for the extraction of initial image features directly from text prompts, whereas a DAMSM text encoder provided sentence-level and word-level embeddings. Then, different up-sampling fusion modules that included deep fusion and attention fusion mechanisms advanced image features in textual guidance. In the discriminator, a pretrained CLIP image encoder was used for evaluation of realism and text-image alignment which were supported by conditional and unconditional losses with a gradient penalty for training stability. PMGAN was evaluated on the CUB and COCO datasets by utilization of IS and FID metrics. The model obtained superior performance to the existing GAN-based approaches. However, the dependence on different pretrained models increased the parameter count and computational needs. Face generation GAN [23] presents effective mapping from linguistic features to facial visual features. The comparative analysis shows considerable improvement in semantic alignment in text description and generated image. While GAN-based approaches outperform existing models mentioned in the literature [24], recent diffusion-based TTI models signify latent efficiency and training stability but lack hierarchical semantic control, which motivates the proposed SALD-TGA framework.

The existing text-to-image GAN based approaches, including ML-GAN [16], ACM-based models [17], DMGAN [18], Control GAN [19], SWF-GAN [20], CF-GAN [21], PMGAN [22] and Face generation GAN[23] have enhanced text-image alignment. In contrast, these existing models struggle due to lack of a robust hierarchical guidance mechanism that jointly imposes global structure, attribute-level consistency and contextual coherence. In particular, the existing approaches lack attribute-level semantic constraints during intermediate generation stages [16], limits from semantic incoherence and structural distortions [17,18]. In addition, demonstrates convergence instability on complex datasets [18] or generates visually coarse outputs with low resolution and multi-object understanding [19,20]. Furthermore, cross-domain fusion models remain sensitive to dataset imbalance [21], and fine-grained attribute modelling often presents inconsistencies in capturing object parts and colour semantics. [22]. Consequently, there is no unified TTI framework that efficiently combines multi-level textual semantics ranging from attribute-level guidance to global scene composition that assists in handling structural stability, semantic fidelity and robustness across complex, multi-object scenarios. The key contributions of the research are as follows:

This research proposes a semantically grounded and computationally efficient framework for Text-To-Image (TTI) generation, namely a Semantic-Aware Latent Diffusion with Text-Guided Augmentation (SALD-TGA) framework.

A text-guided latent data augmentation strategy is incorporated to improve robustness against dataset imbalance and linguistic variability, which conducts controlled semantic perturbations directly in the latent space.

The proposed SALD-TGA combines a semantic-aware diffusion mechanism, where diverse semantic levels control distinct stages of the denoising process. This stage-wise semantic dominance facilitates stable layout formation at early stages and precise attribute refinement at later stages, which resolves the issue of semantic inconsistency and structural ambiguity.

The overall research is structured as follows: Section 2 describes block diagram and its system model, Section 3 demonstrates the proposed SALD-TGA framework. Section 4 illustrates results and discussion, finally the conclusion of this research paper is given in Section 5.

PROPOSED METHOD

The proposed SALD-TGA framework resolves the issues of semantic inconsistency and slow inference in TTI generation by combining semantic-aware representation learning with fast latent diffusion. Further, the text description is embedded as input and a transformer-based encoder first disentangles it into attribute, object and context-level semantic tokens, thereby preserving fine-grained semantic information. Furthermore, to improve data diversity and robustness, a text-guided latent data augmentation strategy is utilized, where controlled semantic-conditioned perturbations are performed directly in the latent space. Furthermore, the augmented semantic representations are then embedded into a semantic-aware latent diffusion generator, in which the semantic tokens guide each denoising step to ensure accurate TTI alignment and structural coherence. Thereby, an accelerated sampling mechanism is incorporated, which assists in significantly reducing generation time without compromising output quality. Finally, a lightweight semantic consistency enhancement module is used, which helps to refine the generated representation before decoding those results in high-resolution and semantically faithful images with fast inference performance. Figure 1 shows an overview of the proposed SALD-TGA framework.

2.1. Dataset Description

The combined use of Caltech-UCSD Birds (CUB-200-2011) and Common Objects in Context (COCO) demonstrates a comprehensive evaluation of the proposed SALD-TGA framework across different semantic

granularities. While CUB-200-2011 evaluates the model's ability to preserve fine-grained attribute-level semantics, COCO tests its efficiency in generating globally coherent, context-aware images involving multiple objects.

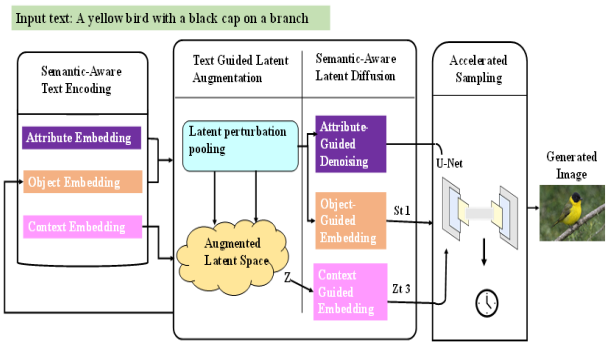

Figure 1. Overview of proposed SALD-TGA framework

2.1.1. CUB-200-2011 Dataset

The CUB-200-2011 is selected as it is a widely used benchmark for fine-grained TTI generation and semantic attribute learning. Specifically, it comprises 11,788 bird images spanning 200 species, where each image is annotated with 10 human-written textual descriptions that help in capturing detailed visual attributes such as colour patterns, wing shapes, beak structures and body textures. The dataset presents significant challenges due to high intra-class similarity and subtle inter-class differences, which makes this specifically suitable for evaluating attribute-level semantic consistency and fine-grained visual fidelity. Thus, following standard evaluation protocols, 150 species are used for training and 50 unseen species for testing, which allows an assessment of the model's generalization capability to unseen categories. The sample images of the CUB-200-2011 dataset are demonstrated in Figure 2.

2.1.2. COCO Dataset

The COCO dataset is a large-scale benchmark that was designed to support vision-language research that includes complex scenes and multiple interacting objects. Specifically, the dataset comprises over 120,000 images, each accompanied by five natural language captions defining the objects present in the scene, their attributes, actions and contextual relationships. When compared to fine-grained datasets, COCO signifies scene-level diversity and contextual richness, which features multiple objects with varying scales, occlusions and backgrounds within a single image. Thus, this dataset poses substantial challenges for TTI generation due to its semantic diversity and spatial complexity. The sample images of the COCO dataset are demonstrated in Figure 3.

Figure 2. Sample Images of CUB-200-2011 Dataset

Figure 3. Sample images of COCO dataset

2.2. System Model and Latent Image Representation

The TTI generation is formulated as a conditional generative learning problem, where a natural language description guides the synthesis of a visually coherent image. Let the training dataset be defined as mentioned in Equation (1), where a text description is denoted as T_i, $I_i \in R^{H\times W\times C}$ refers to the corresponding image. Further, the objective is to learn a conditional mapping that approaches the underlying distribution $p(I \mid T)$, which facilitates the generation of an image $\hat{I}$ that is both visually realistic and semantically aligned with the input text.

$$\mathsf{D} = \left\{\left(T_i, I_i\right)\right\}_{i=1}^{N} \tag{1}$$

Thereby, direct modelling of this distribution in pixel space is computationally intensive and often leads to slow convergence due to the high dimensionality and redundancy of image data. Therefore, to mitigate this issue, the proposed framework works in a compact latent space, which is obtained through a pre-trained variational encoder $E(\cdot)$ as demonstrated in Equation (2). Here, low-dimensional latent representation is defined as $z_i \in R^d$ and $d \ll HWC$.

$$z_i = \mathsf{E}\left(I_i\right) \tag{2}$$

Subsequently, this latent representation assists in capturing crucial semantic and structural characteristics of the image while discarding fine-grained pixel-level noise [25]. Consequently, operating in latent space demonstrates a stable and computationally efficient foundation for further diffusion modelling. When compared to pixel-space GANs and diffusion models, the latent representations determine smoother manifolds, improved numerical stability and significantly decreased inference time, which makes them specifically suitable for iterative generation processes.

2.3. Hierarchical Semantic Encoding of Text Descriptions

Further, accurate semantic interpretation of the input text is essential for realistic image synthesis, as any semantic ambiguity or loss at this stage propagates throughout the generation process. Let, the input text description be defined as T, a transformer-based text encoder, denoted as $F(\cdot)$ is incorporated to extract rich contextualised representations. In particular, the input text description is processed by incorporating a transformer-based text encoder and decomposed into attribute-level, object-level and context-level semantic tokens. Therefore, this operation of semantic separation facilitates accurate TTI alignment and prevents semantic drift during generation. Instead of compressing the entire text into a single global embedding, the proposed SALD-TGA framework decomposes the extracted semantics into multiple hierarchical components, as defined in Equation (3):

$$\left[s_a, s_o, s_c\right] = \mathsf{F}\ (T) \tag{3}$$

Here, attribute-level semantic tokens encoding visual properties such as colour, texture, material and shape is referred as s_a, s_o defining object-level semantics corresponding to entity identity and category and s_c captures context-level semantics demonstrating spatial relationships, background elements and overall scene configuration. Attribute-level semantic tokens are extracted from adjective–modifier spans describing visual properties such as color, texture, and shape. Object-level tokens are derived from syntactic noun heads corresponding to entity identities, while context-level tokens are obtained from relational phrases, including prepositional and spatial clauses. Token separation is performed within the transformer encoder using dependency-aware attention masking, where syntactic head relations are prioritised to resolve ambiguity in linguistic roles [26]. Further, this decomposition replicates the observation that different semantic elements play distinct roles during image formation. Context-level semantics primarily influence global layout and spatial composition, attribute-level semantics govern mid-level visual appearance, and object-level semantics refine fine-grained details.

By disentangling these components explicitly, the proposed method avoids semantic entanglement commonly observed in attention-only models, where dominant objects suppress attribute-level information. The transformer encoder assigns attention weights to tokens based on syntactic roles and semantic relevance. Attribute descriptors are separated from object nouns and contextual phrases, allowing the model to preserve fine-grained attribute information that is otherwise overshadowed in global embeddings. Furthermore, these semantic tokens are not treated as static descriptors; instead, they are designed to interact dynamically with the generative process, ensuring that semantic constraints remain active throughout image synthesis. While the existing TTI models depend on a single text embedding, which facilitates attribute-level semantics to decay during generation. Furthermore, in the proposed SALD-TGA framework, the hierarchical semantic tokens are particularly fed into all downstream modules, ensuring that each semantic level contributes to the generation process and facilitates image formation at the appropriate stages. Hence, this decomposition directly allows the controlled augmentation strategy described next.

SEMANTIC-AWARE LATENT DIFFUSION AND TEXT-GUIDED DATA AUGMENTATION

Although latent diffusion models are robust, their performance degrades under limited training data or imbalanced attribute distributions. Thereby, to resolve this issue, the proposed SALD-TGA framework presents a text-guided latent space augmentation strategy that improves data diversity while preserving semantic integrity. Thus, a latent representation is represented as z_i and its corresponding semantic tokens which are an augmented latent vector $\tilde{z}_i$ is evaluated using Equation (4) and the perturbation magnitude is expressed through Equation (5):

$$\breve{z}_i = z_i + \alpha_a \Delta_a\left(\boldsymbol{s}_a\right) + \alpha_o \Delta_o\left(\boldsymbol{s}_o\right) + \alpha_c \Delta_c\left(\boldsymbol{s}_c\right) \tag{4}$$

The semantic-conditioned perturbations incorporated in Eq. (4) are acquired through learnable projection functions, which assists to map hierarchical semantic tokens into the latent image space. Specifically, for each semantic level corresponding to attribute, object and context information, a linear projection function is defined as $\Delta x(s_x) = W_x s_x + b_x$, where $x \in \{a, o, c\}$ $s_x \in R^d$ denotes the semantic token extracted from the transformer-based text encoder, $W_x \in R^d$ determines a learnable projection matrix, and $b_x \in R^{dz}$ is a learnable bias term. Subsequently, these projections transform semantic embeddings into latent-space perturbation vectors which is defined as $\Delta x(s_x) \in R^{dz \times d_s}$ being dimensionally aligned with the latent image representation. Thereby, the projection parameters are learned jointly with the diffusion model through end-to-end optimization through the diffusion reconstruction loss and semantic alignment loss, without demonstrating additional supervision or auxiliary objectives. The final augmented latent representation is obtained by adding the weighted sum of semantic perturbations to the original latent vector, which is demonstrated as $z' = z + \lambda_a \Delta_a(s_a) + \lambda_o \Delta_o(s_o) + \lambda_c \Delta_c(s_c)$. Thus, to preserve semantic consistency and prevent excessive deviation from the original latent distribution, the magnitude of each perturbation is constrained within a bounded latent radius as defined in Equation (5).

$$\dot{o} \sim N\left(0, \sigma 2I\right), \mathbb{\Delta} z \mathbb{}_2 \leq \delta \tag{5}$$

The perturbation magnitude is constrained within a bounded latent radius δ to preserve semantic consistency, ensuring that augmented samples remain aligned with the original textual semantics. Subsequently, the semantic tokens guide latent-space augmentation, where controlled perturbations are applied to attributes, object structures and contextual features. Therefore, this improves data diversity, thereby mitigating the issue of dataset imbalance and preserves semantic consistency without introducing pixel-level artifacts. When compared to the conventional pixel-

level augmentation, which inadvertently alters semantic meaning, the employed latent-space augmentation demonstrates controlled, semantically informed variations. For example, attribute-level perturbations slightly modify colour intensity or texture patterns without changing object identity, while context-level perturbations assist in altering background composition while preserving foreground structure. Consequently, this augmentation approach is purposely incorporated before diffusion modelling, which enables the augmented semantics to facilitate the entire denoising trajectory instead of only post-generation refinement. As a result, the model learns an optimal conditional distribution that generalises better to unseen textual descriptions.

In particular, the augmented latent representations utilized as input to a diffusion-based generative model. In the subsequent diffusion process, Gaussian noise is added to the latent vector as demonstrated in Equation (6) and the diffusion process is mentioned in Equation (7). In specific, here the noisy latent at timestep t is defined as z_t, γ_t signifies the noise scheduling parameter, and $\epsilon \sim N(0, I)$ refers to Gaussian noise. A linear noise schedule $\{\beta_t\}_{t=1}^{T}$ is adopted, where $\beta_t \in [\beta_{min}, \beta_{max}]$, t indexes the diffusion timestep and the total number of diffusion steps is denoted by T.

$$z_t = \sqrt{\gamma_t}\tilde{z}_i + \sqrt{1-\gamma_t}\dot{o} \tag{6}$$

Unlike conventional latent diffusion models that inject text conditioning uniformly across all denoising timesteps, the proposed SALD-TGA introduces a stage-dependent semantic conditioning operator that explicitly modulates hierarchical semantic influence as a function of diffusion time [27]. This formulation establishes a mathematically distinct conditioning mechanism within the diffusion process. In the Equation (7), s_c, s_o and s_a represent context, object and attribute-level semantic tokens respectively, while α_t, β_t and γ_t state their timestep-dependent influence.

$$C_t = \alpha_t s_c + \beta_t s_o + \gamma_t s_a, \alpha_t + \beta_t + \gamma_t = 1 \tag{7}$$

Furthermore, the reverse diffusion process aims to reconstruct the original latent representation by iteratively removing noise. Hence, this process is formulated based on the hierarchical semantic tokens as expressed in Equation (8). The denoising network $G(\cdot)$ combines semantic information at each timestep, which allows different semantic levels that assist in guiding the denoising process at suitable stages. At early timesteps, context-level semantics are used to establish global layout. As denoising progresses, attribute-level semantics constrain shape and visual properties, while object-level semantics improve identity and fine details in Equation (9):

$$\dot{o}_t = \mathrm{G}\left(z_t, t, \mathrm{s}_a, \mathrm{s}_o, \mathrm{s}_c\right) \tag{8}$$

$$\alpha_t = 1 - T_t, \gamma_t = T_t, \beta_t = 1 - \alpha_t - \gamma_t \tag{9}$$

This schedule ensures that context-level semantics dominate early denoising stages to establish global layout, while attribute-level semantics gradually control later stages for fine-grained visual refinement. Thus, this progressive semantic guidance distinguishes the proposed SALD-TGA method from both GAN-based approaches, which generate images in a single step and conventional diffusion models, whereas many existing approaches apply textual conditioning only at the early or final stages of the synthesis pipeline. Hence, the continuous semantic conditioning significantly reduces structural artifacts and semantic drift.

3.1. Accelerated Sampling for Efficient Inference

Despite their generative quality, diffusion models are often criticised for slow inference. To resolve this limitation, the proposed SALD-TGA framework utilises an accelerated sampling strategy that decreases the number of denoising steps as illustrated in Equation (10). Where η_k signifies an adaptive step size and $K \ll T$ refers to the reduced number of sampling steps. The reverse diffusion process terminates when either $t = 0$ or semantic consistency stabilises across successive denoising steps.

$$z_{t_{k-1}} = z_{t_k} - \eta_k \dot{o}_{t_k} \tag{10}$$

Specifically, a lightweight semantic consistency module computes the correspondence between the generated image and the input text semantics, which demonstrates corrective feedback to improve attribute fidelity, object coherence and scene structure. While the semantic constraints are imposed at every denoising stage, the model requires fewer iterations to converge to a semantically coherent solution. Thereby, this significantly reduces inference time while maintaining image quality, which determines the proposed SALD-TGA framework suitable for practical deployment scenarios.

3.2. Semantic Consistency Enhancement and Image Reconstruction

Although semantic-aware diffusion significantly enhances alignment, minor discrepancies between generated content and textual descriptions persist. Subsequently, to resolve this, a lightweight semantic consistency enhancement module refines the final latent representation before decoding, as demonstrated in Equation (11). Where the refined latent vector is denoted as $z^*$and $D(\cdot)$ refers to the latent decoder determined in Equation (12):

$$\hat{I} = \mathrm{D}\left(z^*\right) \tag{11}$$

$$L_{ref} = 1 - cos\left(\phi\left(I\right), \phi\left(T\right)\right) \tag{12}$$

Here, $\phi(\cdot)$ denotes a frozen multimodal embedding network utilized to evaluate semantic similarity between the generated image I and the input text T. When compared to the adversarial refinement, this module operates deterministically in latent space, correcting residual mismatches without destabilising training. Consequently, the proposed framework ensures that the generated image effectively preserves the attribute-level details that were emphasized in the input text. Then, the refined latent representation is decoded into a final high-quality image, which resulted in improved visual realism, semantic accuracy and computational efficiency that determines the

proposed SALD-TGA suitable for real-time and large-scale applications. Thus, this refinement step operates deterministically in latent space and focuses on correcting residual attribute mismatches without demonstrating additional adversarial objectives. Finally, this operation improves semantic fidelity while maintaining training stability.

3.3. Training Objective and Optimization Strategy

The model is trained using a composite objective function which integrates diffusion reconstruction loss with semantic alignment loss as mentioned in the Equations (13)and (14). The semantic alignment loss applies consistency across attribute, object and context levels, while the balancing coefficient λ ensures smooth gradient interaction with the diffusion reconstruction loss. The diffusion loss assures accurate noise estimation and stable convergence, while the semantic alignment term penalises discrepancies across textual semantics and generated representations.

$$\mathsf{L} = \mathsf{E}_{t,\grave{o}}\left[\grave{o} - \hat{\grave{o}}_t^{\,2} \right] + \lambda \mathsf{L}_{sem} \quad (13)$$

$$\mathsf{L}_{sem} = \sum_{l \in \{a, o, c\}} \| \; (\;) - \; \|_2^2 \quad (14)$$

Further, the balancing parameter λ controls the relative influence of semantic supervision. Furthermore, the adversarial losses are employed in GANs, which are unstable and sensitive to hyperparameters. Thus, diffusion-based objectives assist in stable optimisation and smooth convergence, while semantic supervision mitigates semantic drift, thereby completing the end-to-end learning pipeline. Therefore, this objective avoids the instability associated with adversarial training and illustrates a principled optimization framework for text-conditioned image synthesis.

RESULTS AND DISCUSSION

All experiments follow standard evaluation protocols for TTI generation, ensuring reproducibility and enabling fair comparisons with existing methods. In particular, the proposed SALD-TGA framework is computed on the CUB-200-2011 and COCO datasets, which respectively demonstrate fine-grained object-centric and complex multi-object scene generation scenarios. Subsequently, text descriptions are tokenised and encoded through a transformer-based text encoder, and images are resized to a definite resolution of 256×256. Consequently, the model operates in a latent space learned through a pre-trained variational autoencoder, which assists in reducing computational complexity. Further, the training is conducted using the Adam optimizer with a fixed learning rate and batch size, which is chosen based on stability. Thereby, the diffusion process follows a predefined noise schedule with a reduced number of sampling steps for fast inference. Thus, the model performance is validated using Fréchet Inception Distance (FID) and Inception Score (IS), which are represented as Eq. (15) and Eq. (16), respectively. Here, all reported results are averaged over multiple runs to account for stochastic variability. Specifically, FID measures the distributional similarity between generated and real images, where lower values indicate better visual realism, while IS evaluates both image quality and diversity, with higher values reflecting superior generative performance. Hence, all baseline models are computed using identical text encoders, image resolutions, datasets and evaluation protocols, following the standard configurations to ensure fair and consistent comparison.

$$FID = \mu_r - \mu_g{}_2^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \, \Sigma_g \right)^{\frac{1}{2}} \right) \quad (15)$$

$$IS = exp\left(\mathsf{E}_{x \sim p_g}\left[D_{KL}\left(p(y|x) \, \| \, p(y) \right) \right] \right) \quad (16)$$

Here, $\| \cdot \|_2^2$ is Euclidean norm $Tr\ (\cdot)$ is the trace of a matrix, μ_r and μ_g are mean feature vectors of real and generated images, as well as $\sum_r$ and $\sum_g$ are covariance matrices. In addition, $p(x)$ is a conditional class probability distribution, $D_{KL}(\cdot \parallel \cdot)$ is Kullback–Leibler divergence and $E[\cdot]$ is the expectation.

4.1. Performance Analysis

To further assess the robustness and generalizability of the proposed SALD-TGA framework, performance evaluation is performed against widely adopted TTI generation models such as Latent Diffusion Model (LDM), StackGAN++ and more. Specifically, the considered benchmarks include attention-based GANs, transformer-guided generative models, and latent diffusion approaches, enabling a fair and comprehensive comparison across diverse modelling paradigms, as demonstrated in Table 1. All reported results are averaged over three independent runs, and standard deviations are illustrated where applicable to account for stochastic variability.

Table 1. Performance analysis of the proposed SALD-TGA with benchmark models

Model	FID		IS
	CUB-200-2011	COCO	CUB-200-2011
StackGAN++	13.02	18.52	4.89 ± 0.06
AttnGAN	11.62	15.01	5.09 ± 0.07
DF-GAN	10.78	14.36	5.27 ± 0.05
DALL·E-style Transformer TTI	10.21	13.98	5.31 ± 0.06
LDM	9.84	13.65	5.39 ± 0.05
Proposed SALD-TGA	9.36	13.21	5.48 ± 0.06

From Table 1, it is illustrated that the proposed SALD-TGA framework consistently attains superior performance across all evaluated metrics when compared with advanced benchmark TTI generation models. In particular, the traditional GAN-based architectures such as StackGAN++

and AttnGAN demonstrate higher FID values, specifically on the MS-COCO dataset, due to their limited ability to handle global semantic coherence and structural consistency in multi-object scenes. Although DF-GAN improves over earlier GAN-based methods by incorporating deep fusion mechanisms, its dependency on single-level semantic conditioning limits the preservation of fine-grained attribute details. As a result, DF-GAN produces lower FID and IS scores when compared with the proposed SALD-TGA framework. In addition, transformer-based TTI models achieve better performance due to their ability to capture global context; however, the lack of explicit semantic hierarchy often leads to suboptimal alignment between attribute-level descriptions and the generated visual features. Furthermore, the LDMs obtain strong baseline results due to efficient latent-space modelling, nevertheless, their uniform text-conditioning strategy does not explicitly differentiate semantic roles across diffusion stages. In contrast, the proposed SALD-TGA combines hierarchical semantic decomposition and stage-aware semantic injection, which allows more precise control over layout formation, attribute refinement and object detailing. Thus, the proposed SALD-TGA achieves the lowest FID scores on both datasets and the highest IS on CUB-200-2011 which determines the ability of the SALD-TGA framework in reflecting improved visual realism, semantic consistency and generative diversity.

4.2. Comparative analysis

Furthermore, to validate the robustness of the proposed SALD-TGA framework, a comparative analysis is conducted against several representative TTI generation models. In specific, ML-GAN [16], CA-GAN [17], DMGAN TTI [18], MCA-MHCA-SSGAN [19] and SWF-GAN [20] are the TTI generation models considered. The evaluation is performed on the existing models employed on CUB-200-2011 and MS-COCO datasets in terms of FID and IS as benchmark metrics as demonstrated in Table 2.

Table 2. Comparative analysis of the proposed SALD-TGA with existing models

Model	FID		IS
	CUB-200-2011	COCO	CUB-200-2011
ML-GAN [16]	11.98	16.81	5.05±0.08
CA-GAN [17]	11.78	14.92	5.21±0.05
DMGAN TTI [18]	15.11	NA	4.33±0.03
MCA-MHCA-SSGAN [19]	18.26	27.79	4.96±0.07
SWF-GAN [20]	11.41	19.08	5.15±0.07
Proposed SALD-TGA	9.36	13.21	5.48 ± 0.06

As shown in Table 2, the proposed SALD-TGA framework consistently outperforms all comparative methods across both datasets and evaluation metrics. On the fine-grained CUB-200-2011 dataset, the proposed SALD-TGA framework attains the lowest FID score of 9.36, which defines a substantial enhancement in visual fidelity and distributional alignment when compared to GAN-based approaches such as ML-GAN [16] and SWF-GAN [20]. Subsequently, on the more complex MS-COCO dataset, which contains diverse scenes and multiple objects, the proposed SALD-TGA framework obtains an FID of 13.21, outperforming CA-GAN and SWF-GAN by a significant margin. Hence, this demonstrates the robustness of the proposed SALD-TGA framework's semantic decomposition and context-aware diffusion process in handling complex textual descriptions and multi-object layouts. Specifically, in terms of image quality and diversity, the proposed SALD-TGA framework attains the highest IS of 5.48 ± 0.06 on CUB-200-2011. When compared to DMGAN TTI [18], which demonstrates degraded IS due to semantic fragmentation and convergence instability, the proposed SALD-TGA maintains strong semantic consistency by explicitly incorporating attribute, object and context-level semantics throughout the denoising process. Henceforth, these results validate that the proposed SALD-TGA framework effectively balances realism, diversity and semantic alignment, which determines proposed SALD-TGA is well-suited for high-quality and fast TTI generation.

4.3. Visualization of Generated Images

The qualitative image generation results obtained using the proposed SALD-TGA framework, which assists in validating the robustness in generating semantically consistent images is determined in Table 3. Specifically, both fine-grained and complex multi-object dataset which includes CUB-200-2011 and COCO, respectively are considered.

CUB-200-2011 Dataset			
A small gray bird with white and dark gray wing bars and white breast	A bird has a long bill that is red, as well as large wing bars	This is a bird with green back and head with prominent eye and striped feather with white belly	An all-white bird with darker colored legs and small eyes
			
COCO Dataset			
A cat with some items laid on top of him	A truck parked on a street next to a truck	A boat on the water with a sail boat in the water	
			

Table 3. Generated images using the proposed SALD-TGA framework

From Table 3, it is depicted that these images were generated using the proposed SALD-TGA framework on both the CUB-200-2011 and COCO datasets. Specifically, for the CUB dataset, the model generates fine-grained bird images that align with the corresponding class annotations, which demonstrates accurate capture of attribute-level details. Then, for the COCO dataset, the generated images replicate coherent multi-object scenes that are consistent with the provided labels and annotations. Hence, these results determine that the proposed SALD-TGA framework produces semantically aligned images across datasets of varying complexity without using adversarial training.

4.4. Ablation Study

To assess the individual contribution of each architectural component in the proposed SALD-TGA framework, an ablation study is conducted by progressively employing semantic modelling and optimization modules into a latent diffusion baseline. All experiments are conducted under identical training and evaluation settings on the CUB-200-2011 dataset, which helps to ensure fair comparison. Thus, the analysis focuses on understanding how each design choice affects semantic alignment, visual fidelity and generative diversity as illustrated in Table 4. Specifically, allowing different semantic levels to dominate specific denoising stages facilitates effective global structure formation, finer attribute refinement and sharper object details. This indicates that semantic guidance is robust when integrated directly into the diffusion dynamics, rather than being applied as static conditioning.

Table 4. Ablation study of model variants of proposed SALD-TGA

Model Variant	FID	IS
Latent Diffusion (Baseline)	9.84	5.39 ± 0.05
Baseline with Semantic Decomposition	9.62	5.41 ± 0.05
Baseline with Latent Augmentation	9.51	5.43 ± 0.06
Baseline with Semantic-Aware Diffusion	9.42	5.46 ± 0.05
Full SALD-TGA	9.36	5.48 ± 0.06

From Table 4, it is observed that the latent diffusion baseline is considered a strong baseline model, resembling the performance of modern diffusion-based TTI systems that depend on global text conditioning. Although it achieves reasonable visual quality, it struggles with semantic dilution, where fine-grained attribute details are overshadowed by dominant object-level semantics. Further, incorporating semantic decomposition facilitates improvement in both FID and IS, which determines that separating attribute, object and context-level semantics enhances TTI alignment. Thereby, this improvement defines that hierarchical semantic modelling reduces representational interference by allowing each semantic component to guide image synthesis more effectively. Additionally, the text-guided latent augmentation improves

performance by enhancing robustness to textual variability and dataset imbalance. Furthermore, by conducting augmentation in latent space under semantic constraints, the model explores a broader yet semantically consistent latent distribution. Subsequently, this leads to improved visual diversity, as indicated by lower FID and higher IS, without introducing unrealistic artifacts. Consequently, allowing semantic-aware diffusion demonstrates a significant performance gain among all components. Finally, employing semantic-preserving sampling results in additional improvement while significantly reducing inference time. Hence, the ablation results signify that each architectural component contributes incremental yet cumulative improvements. Thus, this determines that hierarchical semantic decomposition and semantic-aware diffusion jointly enhance alignment and visual fidelity rather than any single module demonstrating disproportionate gains.

4.5. Cross data Evaluation

To assess the robustness and generalization ability of the proposed SALD-TGA framework, a cross-dataset evaluation is conducted on CUB-200-2011 and COCO datasets Where each model is trained on one dataset and directly evaluated on another without additional fine-tuning, as demonstrated in Tables 5 and 6. Specifically, this cross-data evaluation is performed for adopting generalization as general k-fold cross-validation is not applicable in text-to-image generation. In addition, the main motive behind this evaluation is to assess domain-invariant semantic learning. Thereby, cross-dataset generalization is an essential indicator of semantic robustness in TTI generation, as it determines a model's ability to transfer learned textual-visual associations across domains with differing visual distributions, scene complexity and annotation styles. As shown in Tables 5 and 6, all models obtained a degradation in performance under cross-dataset evaluation due to a significant domain shift between CUB-200-2011 and COCO. Specifically, GAN-based approaches such as StackGAN++ and AttnGAN demonstrate the degradation, primarily due to their dependency on dataset-specific visual patterns and limited semantic abstraction capabilities. DF-GAN demonstrates improved generalization when compared to GAN-based methods; however, its performance remains constrained by single-level text conditioning, which limits its ability to transfer fine-grained attribute semantics across domains. Further, LDMs present stronger cross-dataset robustness due to latent-space modelling and stable training dynamics.

Table 5. Results of proposed SALD-TGA when trained on CUB-200-2011 dataset and evaluated on the COCO dataset

Model	FID	IS
StackGAN++	29.47	3.92 ± 0.06
AttnGAN	26.18	4.11 ± 0.05
DF-GAN	24.96	4.24 ± 0.07
LDM	22.41	4.36 ± 0.06
Proposed SALD-TGA	20.85	4.58 ± 0.05

Table 6. Results of proposed SALD-TGA when trained on COCO dataset and evaluated on COCO dataset

Model	FID	IS
StackGAN++	21.83	4.27 ± 0.07
AttnGAN	19.64	4.49 ± 0.06
DF-GAN	18.52	4.63 ± 0.05
LDM	17.21	4.71 ± 0.05
Proposed SALD-TGA	15.89	4.94 ± 0.06

However, they still struggle with semantic misalignment when encountering unseen or unfamiliar contextual distributions. In contrast, the proposed SALD-TGA framework consistently achieves the best performance in both transfer directions. These results determine that hierarchical semantic decomposition and text-guided latent augmentation significantly enhance semantic abstraction and domain-invariant representation learning. Thus, performance degradation under cross-dataset evaluation primarily arises from mismatched contextual distributions, while attribute-level semantics demonstrates stronger transferability across domains.

4.6. Discussion

The experimental results demonstrate that semantic inconsistency and slow inference remain challenges in existing TTI generation models. GAN-based approaches, although effective in capturing object-level features, often struggle with semantic coherence and multi-object reasoning, specifically under complex textual descriptions. When compared to attention-based models, which enhance alignment but remain sensitive to noisy or fine-grained attributes facilitates structural artifacts. Further, diffusion-based models provide improved stability and realism; however, dependency on uniform text conditioning limits semantic control and interpretability. Thereby, the proposed SALD-TGA framework resolves these challenges by embedding semantic awareness into the generative process and demonstrates consistent and measurable improvements. Furthermore, the hierarchical semantic decomposition plays an essential role in reducing representational conflict among textual components, which enables the model to preserve fine-grained attributes while handling global coherence. Thus, the ablation study determines that semantic-aware diffusion contributes to the significant performance improvement, signifying the importance of stage-dependent semantic guidance rather than static conditioning. Despite these strengths, the framework still demonstrates limitations in extremely complex scenes involving dense object interactions and long, ambiguous textual descriptions. Hence, these scenarios suggest that deeper reasoning mechanisms and stronger language–vision alignment may be required.

CONCLUSION

This research resolves a critical research gap in TTI generation, namely the lack of semantic guidance across the generative process and the high computational cost associated with diffusion-based models. When compared to

the existing GAN-based and attention-driven approaches, which depend on shallow or single-level text conditioning that facilitates semantic inconsistency, fragmented object structures and demonstrate limited attribute retention, specifically in complex scenes and fine-grained synthesis scenarios. While recent diffusion models improve visual realism, they often struggle with uniform text conditioning and slow inference, which limits their practical applicability. Specifically, to address these challenges, this research proposes SALD-TGA, a semantic-aware latent diffusion framework that integrates hierarchical text semantics and an efficiency-oriented design into the generation pipeline. Subsequently, the employed semantic decomposition mechanism disentangles attribute, object and context-level information, which ensures that each semantic component guides the image synthesis process at an appropriate stage. Consequently, this research addresses the semantic dilution and structural ambiguity observed in existing TTI models. Furthermore, the text-guided latent augmentation strategy improves robustness to data imbalance and linguistic variability, resolving the limited generalization capability of prior approaches. Consequently, by embedding semantic awareness into the diffusion dynamics and incorporating accelerated semantic-preserving sampling, the proposed SALD-TGA framework achieves improved semantic fidelity and reduced inference time, thereby resolving the long-standing trade-off between generation quality and efficiency, which demonstrates consistent and measurable improvements. In the future, the proposed SALD-TGA will explore handling dense multi-object interactions and long, compositional text descriptions by incorporating reasoning-aware language encoders and structured semantic graphs

REFERENCE

[1] H. Chen, Z. Zuo, L. Zhao, J. Li, and J. Yang, “ConceptCraft: One-shot personalized text-to-image generation via object-background disentanglement,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 36, no. 1, pp. 133-146, 2025.

[2] J. Zhu, H. Ma, J. Chen, and J. Yuan, “Isolated diffusion: Optimizing multi-concept text-to-image generation training-freely with isolated diffusion guidance,” IEEE Transactions on Visualization and Computer Graphics, vol. 31, no. 9, pp. 6280–6292, 2024.

[3] S. K. Alhabeeb and A. A. Al-Shargabi, “Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction,” IEEE Access, vol. 12, pp. 24412–24427, 2024.

[4] J. Zhang, X. Li, Z. Wan, C. Wang, and J. Liao, “Text2NeRF: Text-driven 3D scene generation with neural radiance fields,” IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 12, pp. 7749–7762, 2024.

[5] 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.

[6] H. Zhang, T. Wu, and Y. Wei, “Multi-view user preference modeling for personalized text-to-image generation,” IEEE Transactions on Multimedia, vol. 27, pp. 3082–3091, 2025.

[7] M. A. Habib, M. A. H. Wadud, M. M. Rahman, and M. F. Mridha, “MemAttn-CL: Unified memory, attention, and contrastive learning for enhanced text-to-image generation,” IET Image Processing, vol. 19, no. 1, p. e70185, 2025.

[8] M. A. Habib, M. A. H. Wadud, L. Y. Pinky, M. H. Talukder, M. M. Rahman, M. F. Mridha, Y. Okuyama, and J. Shin, “GACNet-text-to-image synthesis with generative models using attention mechanisms with contrastive learning,” IEEE Access, vol. 12, pp. 9572–9585, 2023.

[9] Z. Li, L. Liu, H. Zhang, D. Liu, Y. Song, and B. Li, “Locally controllable network based on visual–linguistic relation alignment for text-to-image generation,” Multimedia Systems, vol. 30, no. 1, p. 34, 2024.

[10] H. He, H. Yang, Z. Tuo, Y. Zhou, Q. Wang, Y. Zhang, Z. Liu, W. Huang, H. Chao, and J. Yin, “DreamStory: Open-domain story visualization by LLM-guided multi-subject consistent diffusion,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 12, pp. 11874–11891, 2025.

[11] H. Kim, J. H. Choi, and J. Y. Choi, “A novel scheme for generating context-aware images using generative artificial intelligence,” IEEE Access, vol. 12, pp. 31576–31588, 2024.

[12] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han, “FastComposer: Tuning-free multi-subject image generation with localized attention,” International Journal of Computer Vision, vol. 133, no. 3, pp. 1175–1194, 2024.

[13] A. Sebaq and M. ElHelw, “RSDiff: Remote sensing image generation from text using diffusion model,” Neural Computing and Applications, vol. 36, no. 36, pp. 23103–23111, 2024.

[14] R. Li, W. Li, Y. Yang, H. Wei, J. Jiang, and Q. Bai, “SwinV2-Imagen: Hierarchical vision transformer diffusion models for text-to-image generation,” Neural Computing and Applications, vol. 36, no. 28, pp. 17245–17260, 2023.

[15] Q. Lan and H. Wu, “Dynamic local affine transformation for enhanced text-to-image generation with GANs,” The Visual Computer, vol. 41, pp. 9691–9704, 2025.

[16] H. Zhao, H. Wang, and Y. Yang, “ML-GAN: Multi-level text-driven fine-grained image generation using generative adversarial network,” Neurocomputing, vol. 651, p. 130851, 2025.

[17] Y. Hou, W. Zhang, Z. Zhu, and H. Yu, “Language-vision matching for text-to-image synthesis with context-

aware GAN," Expert Systems with Applications, vol. 255, p. 124615, 2024.
[18] R. Gopalakrishnan, S. Naveen, S. Kalathil, and P. V. Sudeep, "Self-attention based text encoder for enhancing DMGAN text-to-image generation," IEEE Access, vol. 13, pp. 125442–125457, 2025.
[19] S. Hou, Z. Li, K. Wu, Y. Zhao, and H. Li, "Masked cross-attention and multi-head channel attention guiding single-stage generative adversarial networks for text-to-image generation," The Visual Computer, vol. 40, no. 12, pp. 8639–8651, 2024.
[20] C. Liu, J. Hu, and H. Lin, "SWF-GAN: A text-to-image model based on sentence–word fusion perception," Computer Graphics, vol. 115, pp. 500–510, 2023.
[21] Y. Zhang, S. Han, Z. Zhang, J. Wang, and H. Bi, "CF-GAN: Cross-domain feature fusion generative adversarial network for text-to-image synthesis," The Visual Computer, vol. 39, no. 4, pp. 1283–1293, 2022.
[22] Y. Yu, Y. Yang, and J. Xing, "PMGAN: Pretrained model-based generative adversarial network for text-to-image generation," The Visual Computer, vol. 41, no. 1, pp. 303–314, 2024.
[23] Bankar, S., & Ket, S. (2025). From words to faces: A GAN-based approach for text-driven face generation. In Artificial Intelligence and Sustainable Innovation (pp. 67–73). https://doi.org/10.1201/9781003731689-11
[24] Bankar, S. A., & Ket, S. (2021). An analysis of Text-to-Image synthesis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3852950
[25] Li, R., Li, W., Yang, Y., Wei, H., Jiang, J. and Bai, Q., 2024. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation. Neural Computing and Applications, 36(28), pp.17245-17260.
[26] Habib, M.A., Wadud, M.A.H., Rahman, M.M. and Mridha, M.F., 2025. MemAttn-CL: Unified Memory, Attention, and Contrastive Learning for Enhanced Text-to-Image Generation. IET Image Processing, 19(1), p.e70185.
[27] He, H., Yang, H., Tuo, Z., Zhou, Y., Wang, Q., Zhang, Y., Liu, Z., Huang, W., Chao, H. and Yin, J., 2025. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence.