International Journal of Innovative Computing, Information and Control
Volume **22**, Number **1**, February **2026**

 ISSN 1349-4198
pp. **261–281**

AMAC-T2I: A DIFFUSION FRAMEWORK FOR TEXT-TO-IMAGE GENERATION BASED ON ATTRIBUTE MATCHING AND ATTRIBUTE CORRECTION

Zhibo Dai[1,*], Ping Zong[2], Yuhua Cong[3], Mengyu Xu[1] and Wenhan Sun[1]

[1]School of Computer and Artificial Intelligence
Nanjing University of Science and Technology ZiJin College
No. 89, Wenlan Road, Qixia District, Nanjing 210046, P. R. China
mengyxu@hotmail.com; johnnysun991223@126.com
*Corresponding author: daizhibo@njust.edu.cn

[2]School of Computer Science
Nanjing University of Posts and Telecommunications
No. 9, Wenyuan Road, Yadong New District, Nanjing 210023, P. R. China
zong@njupt.edu.cn

[3]College of Automation Engineering
Nanjing University of Aeronautics and Astronautics
No. 29, Yudao Street, Qinhuai District, Nanjing 210016, P. R. China
congyuhua@nuaa.edu.cn

Received May 2025; revised August 2025

Abstract. *Diffusion models, as a paradigm shift in artificial intelligence, have demonstrated outstanding capabilities in image processing and modeling, marking a significant breakthrough in knowledge discovery and visual content creation. However, achieving precise alignment between textual prompts and the semantic and visual attributes of generated images remains a key challenge. To address this, we propose a diffusion-based text-to-image generation framework – AMAC-T2I – that aims to improve fine-grained attribute control and text-image consistency under textual guidance. The framework consists of two main components: the Attribute Correction Module (ACM) and the Attribute Repair Module (ARM). ACM incorporates prior image information via the multimodal encoder Q-Former, enabling a more targeted and guided generation process. It also uses adversarial loss as supervision to dynamically correct attribute discrepancies. ARM combines Q-Former with a cross-attention mechanism to effectively extract correspondence features between text and image and leverages the UNet network to generate category masks and calculate attention loss. This design addresses attribute omissions and mapping errors, thereby enhancing the generalization and consistency of generated results. Extensive experiments conducted on two authoritative attribute-level datasets, T2I-CompBench and DPG-Bench, demonstrate that AMAC-T2I significantly surpasses existing state-of-the-art methods in attribute alignment accuracy, generation controllability, and text-image alignment flexibility, highlighting its superior performance and broad application potential.*

Keywords: Diffusion model AMAC-T2I, ACM, ARM, Text-to-image generation, Visual attributes

1. **Introduction.** In recent years, diffusion models have made remarkable progress in the field of text-to-image generation [1]. Compared with earlier approaches such as Generative Adversarial Networks (GANs) and autoregressive models, diffusion models not only generate images with higher resolution and richer details but also exhibit significant

DOI: 10.24507/ijicic.22.01.261

advantages in authenticity and semantic alignment [2,3]. A high-quality generated image often encompasses a variety of visual attributes, including fine brushstroke textures, carefully designed composition, and realistic lighting effects [4,5]. The effective management and precise expression of these visual attributes are critical, as they not only determine the aesthetic standard of the image but also significantly enhance its emotional appeal and persuasive power in real-world applications – ranging from digital art creation and product design previews to advertising, self-media modeling, and virtual scene construction [6-9]. Consequently, enhancing the capacity of diffusion models for fine-grained attribute control has emerged as a key research challenge in the text-to-image generation domain [10,11].

At the same time, recent advances in generative models, especially GANs and diffusion-based frameworks, have achieved impressive performance in tasks such as image reconstruction, style synthesis, and image enhancement. For instance, Zhao et al. proposed a GAN architecture with improved batch normalization and feature extraction modules for image super-resolution, which significantly improves reconstruction quality under low-resolution settings [12]. Chen et al. developed a local expression diffusion model capable of synthesizing natural and controllable facial expressions through fine-grained modeling of local facial regions [13]. Furthermore, Zhao et al. introduced a dehazing GAN for remote sensing imagery that integrates color feature restoration, achieving superior clarity and structural consistency in degraded environments [14]. These studies demonstrate the importance of combining architectural innovations with task-specific conditioning strategies to enhance image generation quality and semantic controllability.

Despite recent progress, existing diffusion models still face limitations in capturing and expressing fine-grained attributes, such as color, texture, pose, lighting effects, and artistic styles [5,6]. These attributes are crucial across various real-world applications. For instance, the realistic depiction of material texture and lighting directly impacts the reliability of design previews, while in art creation, the detailed rendering of brushstroke styles and color schemes shapes the aesthetic value and visual impact of the work [10,11,15]. Models such as BERT and LSTM, known for their strong sequential modeling and contextual reasoning capabilities, have been integrated into image generation pipelines [16], with BERT, in particular, widely adopted due to its superior semantic modeling capacity [17]. However, most current diffusion models still struggle with accurate attribute regulation, often resulting in mismatched or missing attributes in generated images, which limits their practical utility in areas like advertising, digital art, and film production [10,15].

To address this, some studies have proposed fine-grained attribute control techniques for specific objects. These methods attempt to systematically analyze how a given object and its attributes manifest across different contexts, aiming to improve generalization and control accuracy [18,19]. However, such techniques are often limited to object-attribute pairs seen during training and exhibit poor generalization to unseen combinations, making them difficult to deploy in open-world applications [7]. Other approaches introduce auxiliary control signals such as style labels, color templates, or spatial layout guidance to direct the diffusion process [20-22]. Yet, these methods typically work only in predefined datasets or controlled environments, and their performance degrades significantly when applied to new or complex scenarios [23,24].

Moreover, many existing methods fail to consider the consistency and divergence of attribute descriptions across different images. That is, they often neglect the challenge of accurate attribute matching under multi-scene and multi-attribute conditions, which leads to attribute mismatches or omissions during generation [25,26]. Some recent research attempts to mitigate this issue by introducing external knowledge, such as extracting structured priors from literary texts, knowledge graphs, or large-scale image-text corpora

[10,15,22,23]. These approaches often explicitly model object-attribute relationships to enhance attribute comprehension and alignment. However, their dependence on manually designed matching rules increases system complexity and limits scalability to novel object-attribute configurations [18,27,28].

With the rapid advancement of Large Language Models (LLMs), recent work has explored leveraging their semantic reasoning capabilities to decompose complex text prompts into granular visual attributes, which are sequentially grounded during image generation [6,24]. While this step-by-step conditioning strategy improves global text-image alignment, it still suffers from ambiguity and semantic overlap among attributes. Consequently, errors or omissions in individual attribute expression remain common, limiting the effectiveness of fine-grained attribute control [25].

To tackle these challenges, we propose a novel framework: AMAC-T2I (Attribute Matching and Attribute Correction for Text-to-Image), which aims to enhance the controllability of fine-grained attributes and improve text-image alignment consistency in diffusion-based generation. The framework integrates two key components: the Attribute Correction Module (ACM) and the Attribute Repair Module (ARM).

The ACM, inspired by LED [13], introduces a discriminator and adversarial loss on top of a Denoising Diffusion Probabilistic Model (DDPM) to supervise attribute consistency. When an attribute mismatch is detected, the discriminator guides the model to dynamically correct discrepancies under adversarial supervision. In addition, ACM leverages a multimodal encoder (Q-Former) to provide strong image priors, offering reliable guidance throughout the generation process for more precise attribute alignment.

The ARM module incorporates Q-Former into a cross-attention mechanism to extract relevant text-conditioned visual features. Based on these features, ARM utilizes a UNet network to generate fine-grained category masks and calculates an attention-based loss, enabling the repair of missing or inconsistent attributes. This structure allows flexible and compositional attribute restoration, enhancing the model's adaptability across diverse scenes and tasks.

To comprehensively validate the effectiveness of AMAC-T2I, we conduct extensive experiments on multiple opendomain datasets with diverse attribute types. Results show that our method outperforms state-of-the-art approaches in attribute accuracy, generation controllability, and alignment flexibility, demonstrating its potential to drive the next wave of advancements in text-to-image generation.

Based on the above discussion, we summarize the contributions as follows.

1) A diffusion model for text-to-image generation based on Attribute Matching and Attribute Correction (AMAC-T2I) is proposed. This model is designed to boost the accuracy of attribute control and improve the consistency of text-image alignment within text-to-image generation tasks.

2) We bring in the ACM. By leveraging the multimodal encoder Q-Former as prior image information, we render the generation process more focused and directed. This, in turn, enhances the precision of text-image attribute alignment. Additionally, adversarial loss serves as a supervisory cue, steering the model to dynamically rectify attribute discrepancies.

3) We incorporate the ARM, which combines Q-Former with a cross-attention mechanism for the efficient extraction of text-image correspondence features. By integrating with the UNet network, it produces category masks and calculates attention loss. This approach efficiently addresses attribute omissions or discrepancies in the image, thereby strengthening the model's generalization capabilities.

4) We conducted comprehensive experimental validation on multiple attribute type datasets. The results show that our method outperforms existing methods in terms of

attribute matching accuracy, controllability during generation, and flexibility in text-image alignment.

2. Related Work.

2.1. Image generation and diffusion model.

Diffusion models have made significant progress in the field of image generation, primarily due to their outstanding generation capabilities. These models can transform Gaussian noise into clear and detailed images through a step-by-step denoising process. DDPM (Denoising Diffusion Probabilistic Model) utilizes this denoising process and is widely applied in various image generation tasks [1]. Latent diffusion models, developed further based on DDPM, apply score matching techniques in the latent space of the images, combining cross-attention mechanism control to significantly enhance the quality and stability of generated images. This approach not only enhances the generative ability of conventional diffusion models but also showcases its superiority in more intricate application scenarios [5].

Additionally, flow models have contributed to further improvements in diffusion models. By introducing flow matching techniques, flow models enhance the stability of the training process, avoid common collapse phenomena in generation, and significantly improve the image synthesis effects. Flow models can also reduce training time while generating high-quality images, making the models more practical and operable in real-world applications [29].

At the same time, another line of research focuses on innovations and variations in the diffusion model architecture. These new architectures improve the performance of traditional models on large-scale datasets. For instance, diffusion transformers and their associated variants substitute the U-Net backbone with a transformer architecture. This substitution offers greater efficiency and enhanced scalability during the processing of large datasets. These methods not only overcome the computational and memory limitations of the original architecture but also improve the model's performance and scalability by capturing finer details during image generation through optimized attention mechanisms [30-32].

Furthermore, recent research efforts have brought forward novel image editing methods that facilitate generation under particular image conditions. Approaches such as ObjectStitch [33], Paint by Example [34], and AnyDoor [35] leverage the CLIP [36] model. They utilize images as conditional factors to direct the generation procedure. These methods combine text and image information and use cross-modal learning to generate images that meet specific requirements, making the generated images more diverse and accurate. Through this approach, the model can understand the complex semantics within the images and generate visual content that aligns with specific requirements, thereby expanding the application scenarios of diffusion models, especially in image editing and modification.

In animation generation, Animate Anyone [37] introduced an innovative method that combines reference networks with skeletal structures, allowing image generation to have a more dynamic feel. This method is particularly suitable for generation tasks that require skeletal or motion structures. By integrating skeletal structure details sourced from reference images, it guarantees that the produced images uphold consistent movements and postures throughout the animation process. This integration significantly boosts the authenticity and fluidity of the generated images.

However, most existing methods mainly focus on the visual attributes of reference images, such as identity features and spatial structure. These methods often struggle to precisely control the style attributes of images.

2.2. **Text to image alignment.** Attention-based techniques [27,28,38] primarily focus on enhancing the attention modules in the UNet architecture. By introducing new attention mechanisms, these models pay more attention to key regions during image processing, thereby improving the quality and accuracy of image generation. However, these methods often require the design of specific heuristic approaches for each misalignment issue, which adds complexity to their application. Especially when dealing with complex visual tasks, ensuring that the attention mechanism effectively guides the model to accurately align images remains a challenge.

Another approach is based on planning techniques, where the layout of the image is used to guide the generation process. These layouts can come from explicit user input [18,39,40] or be automatically generated by Large Language Models (LLMs) [39,40]. Through this method, the generated image can better align with predefined structural requirements.

To enhance the alignment of generated images further, several studies have suggested strengthening the generation procedure with feedback from image understanding models. For instance, the use of Visual Question Answering (VQA) models [41] can help select images that align with the generated content, thereby fine-tuning the diffusion model to make the generated image better match the expected content. Additionally, [42] introduced Reinforcement Learning (RL) fine-tuning techniques, which optimize the generation process through general reward signals, further improving image quality and alignment. Some other research, as cited in [43], has incorporated the backpropagation of the reward signal within the denoising procedure. By fine-tuning the influence of the reward signals, this approach aims to enhance both the precision and quality of the images that are generated.

In contrast to these methods, our approach focuses more on refining each detail in the generation process and enhancing the generation results through precise control matching mechanisms.

3. **Method.** The overall framework is shown in Figure 1. In Section 3.1, we first introduce the Attribute Correction Module (ACM), followed by the introduction of the Attribute Repair Module (ARM) in Section 3.2.

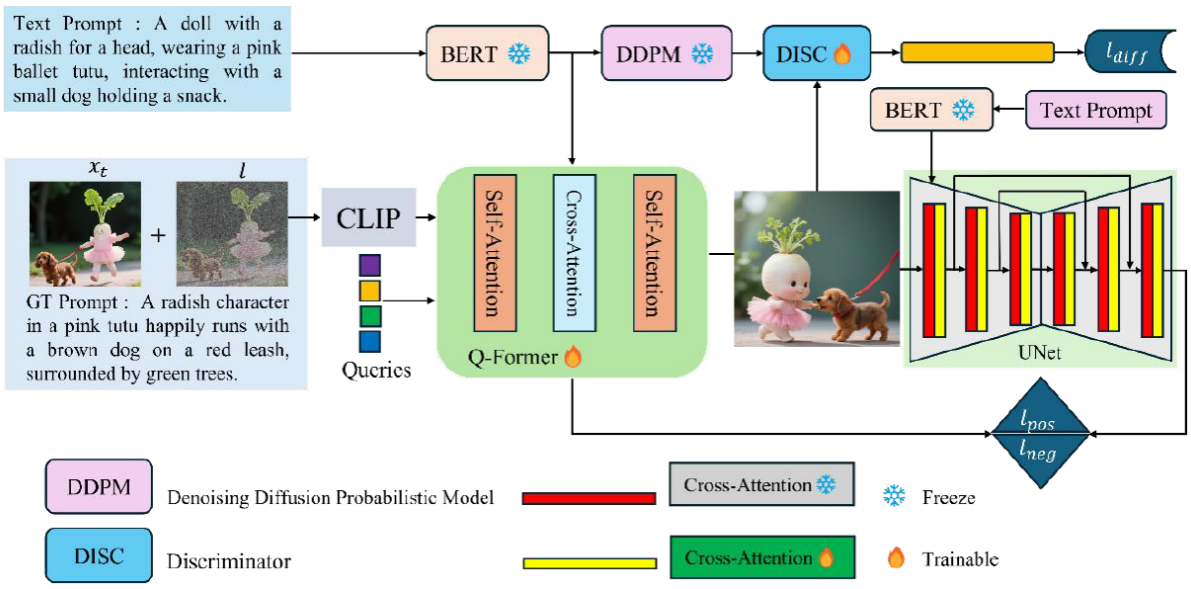

FIGURE 1. Overall framework diagram of AMAC-T2I

3.1. **Attribute Correction Module (ACM).** The ACM has the objective of enabling a pre-trained text-to-image diffusion model to produce images in response to image prompts. It is composed of a text encoder (BERT) utilized for the extraction of text features, a diffusion model (DDPM), a discriminator, as well as an adversarial loss. Once the text has passed through the encoder, the DDPM creates images according to the text. With the guidance of the discriminator and under the influence of the adversarial loss, the model rectifies errors in attribute matching.

We present a novel adversarial loss function. It employs the discriminator to differentiate the images which are generated by the diffusion models that have been pre-trained and fine-tuned. Traditionally, the discriminator typically uses real-world images as input for real data, but we take a different approach by using images generated by the original pre-trained diffusion model (DDPM). This choice is based on a key reason: there is a significant difference between images generated by the original diffusion model and real-world images. Directly aligning the image distribution produced by the fine-tuned diffusion model with the real-world image distribution can pose avoidable difficulties and challenges during the training phase. Such alignment has the potential to restrict the quality of the generated images.

In this research, we put forward multiple strategies for optimizing the design of the discriminator D_ϕ with the aim of boosting the model's performance and elevating the quality of the generated images. Initially, while the utilization of a pre-trained UNet can efficiently acclimatize to the domain of the input data, it might cause an excessive dependence on pre-trained knowledge. This, in turn, restricts the model's adaptability. To tackle this problem, we employ transfer learning to conduct fine-tuning on the pre-trained UNet. This allows the discriminator to modify its parameters in accordance with the present task, thus enhancing its performance for particular tasks and datasets. This fine-tuning mechanism not only avoids complete dependence on the pre-trained model's weights but also enhances the discriminator's adaptability to new tasks.

In the fine-tuned discriminator, the loss function is set as the adversarial loss. Its objective is to differentiate between the latent variables $\hat{z}_0$ of the images generated by the initially pre-trained diffusion model and the latent variables $\hat{z}_0'$ of the images produced by the fine-tuned model. This is presented in Equation (1):

$$L_{\text{adv}} = \log(D_\phi(\hat{z}_0)) + \log(1 - D_\phi(\hat{z}_0')) \tag{1}$$

Additionally, we introduce the concept of a conditional discriminator. Traditional discriminators only focus on the latent representations of the images $\hat{z}_0$ and $\hat{z}_0'$, but we add extra conditional inputs, such as text descriptions or image attributes, to this foundation. This allows the discriminator to evaluate the quality of image generation based on conditional information, not solely relying on the image itself. By combining conditional information with image latent variables, the discriminator can more accurately assess the match between the image and the conditions, enhancing the consistency of the generated image in terms of both details and content. Specifically, the improved loss function can be expressed as Equation (2):

$$L_{\text{diff}} = \log(D_\phi(\hat{z}_0, c)) + \log(1 - D_\phi(\hat{z}_0', c)) \tag{2}$$

To boost the quality and variety of the generated images, we integrate real-world image data into the learning procedure. In detail, apart from the latent variable z_{noise} derived from pure noise, we generate a noisy variant of the real latent variable $z_{\text{noisy}}(\tau)$. This is achieved by introducing noise to a real-world image $\mathbf{x}_{\text{real}}$ at an arbitrarily chosen time step τ, as presented in Equation (3):

$$z_{\text{noisy}}(\tau) = \mathbf{x}_{\text{real}} + \mathcal{N}\left(\mu, \sigma^2\right) \tag{3}$$

where $\mathcal{N}(\mu, \sigma^2)$ represents Gaussian noise, with μ and σ^2 being the mean and variance of the noise.

Subsequently, we denoise both types of latent variables z_{noise} and $z_{\text{noisy}}(\tau)$, and compute the loss $\mathcal{L}$ through the image-text model, as shown in Equation (4):

$$\mathcal{L} = \lambda_{\text{image}}\mathcal{L}_{\text{image}}(z_{\text{denoised}}) + \lambda_{\text{text}}\mathcal{L}_{\text{text}}(z_{\text{denoised}}, \mathbf{y}) \tag{4}$$

where $\mathcal{L}_{\text{image}}$ and $\mathcal{L}_{\text{text}}$ are the image loss and text loss, respectively, z_{denoised} is the denoised latent variable, $\mathbf{y}$ is the text prompt, and λ_{image} and λ_{text} are weight parameters.

This approach's central concept is to utilize the noisy variant of the real latent variable $z_{\text{noisy}}(\tau)$ to direct the generation process. By doing so, it enables the model to produce more lifelike images. In detail, the noisy version of the real latent variable $z_{\text{noisy}}(\tau)$ represents a perturbed form of the original image. It also exhibits a high degree of consistency with the relevant text prompt $\mathbf{y}$ (for example, descriptions or image attributes), as indicated by the relationship in Equation (5):

$$\mathbf{y} = f(z_{\text{noisy}}(\tau)) \tag{5}$$

Thus, with this method, the diffusion model learns directly from the noisy latent variable how to reconstruct the original image $\mathbf{x}_{\text{real}}$, providing an effective path for optimizing the generation process. This is achieved by minimizing the reconstruction loss, as shown in Equation (6):

$$\mathcal{L}_{\text{reconstruction}} = \|\mathbf{x}_{\text{real}} - \widehat{\mathbf{x}}\|_2^2 \tag{6}$$

where $\widehat{\mathbf{x}}$ is the image generated by the model.

This guiding strategy not only helps smooth the model's optimization process but also prevents gradients from overly relying on feedback from the image-text model, avoiding overfitting issues when generating images. The model combines the perturbed real-image version $z_{\text{noisy}}(\tau)$ with the generated latent variable. This combination encourages the model to generate images that are more in line with the text prompt. It also helps in maintaining a higher level of fidelity. Consequently, the produced images not only precisely correspond to the text prompt $\mathbf{y}$, but they also display more lifelike and natural visual appearances. This ultimately enhances the overall quality of the generated images.

By combining noise with the perturbed version of the real image $z_{\text{noisy}}(\tau)$, we enhance the diffusion model's learning ability, ensuring that it generates more stable and efficient images during the process, while better satisfying the set conditions and achieving high quality.

3.2. **Attribute Repair Module (ARM).** ARM is crafted with the aim of empowering pre-trained text-to-image diffusion models to create images in response to image prompts. The fundamental concept underlying this approach lies in augmenting the model's generation capacity via an image encoder and separated cross-attention modules. The function of the image encoder is to distill useful feature representations from the input image. The image features in question are not just basic descriptions of images. Instead, they are embedded representations that thoroughly mirror the image's content. To be more precise, the image encoder makes use of the pre-trained CLIP image encoder [21] for the extraction of image features. Subsequently, these features are processed via the Q-Former network. This process leads to the generation of a mapped image feature sequence $\mathbf{z}_{\text{image}}$. The Q-Former [44], initially presented as a trainable component, has the objective of reducing the representational disparity between the image encoder and Large Language Models (LLMs). In this work, we propose a multimodal conditional generation method that fuses text and images, with the goal of enhancing the coupled expression ability

between semantics and style. We aim to generate a target image based on the input text prompt and image attribute information. The modeling process is defined as follows:

$$\mathbf{I}_{\text{out}} = \mathcal{G}(f_v(\mathcal{X}), f_a(\mathcal{A}), f_t(\mathbf{y})) \tag{7}$$

Here, $\mathbf{I}_{\text{out}}$ denotes the generated image, and $\mathcal{G}(\cdot)$ is a multimodal generation function, which can be instantiated as a diffusion model, a Generative Adversarial Network (GAN), or other deep generative architectures. The term $f_v(\mathcal{X})$ represents the processed visual features extracted from an image encoder (e.g., CNN, Vision Transformer, or CLIP image branch), where $\mathcal{X}$ captures structural and appearance-level cues. The component $f_a(\mathcal{A})$ encodes the attribute control vector $\mathcal{A}$, which encodes fine-grained constraints such as object presence, layout, or style, guiding both local and global visual characteristics. Lastly, $f_t(\mathbf{y})$ transforms the semantic input $\mathbf{y}$ – typically encoded via a language model such as Transformer or CLIP text encoder – into a semantic embedding that provides high-level guidance during generation. This formulation enables flexible and fine-grained image synthesis under multimodal supervision, supporting both semantic fidelity and visual realism.

The architecture is composed of two transformer subcomponents. One is responsible for the extraction of visual features, while the other functions as an encoder-decoder for text. It uses learnable query tokens as its input. These tokens are then related to text via cross-attention layers. Finally, it produces outputs of image features that are in correspondence with the given text. ARM consists of two crucial elements. Firstly, it extracts image condition features associated with text attributes from the image. Secondly, it embeds the text prompt as a condition and the image features into the UNet. Here, two distinct cross-attention layers are employed to generate multi-condition image features. The process by which cross-attention handles text and image features is presented in Equation (8):

$$Y_{\text{new}} = \text{Softmax}\left(\frac{QK^\top + \gamma_1 \cdot Q'K'^\top}{\sqrt{d}}\right)V + \text{Softmax}\left(\frac{QK'^\top + \gamma_2 \cdot Q''K''^\top}{\sqrt{d}}\right)V' \tag{8}$$

Here, $Q = ZW_q$ represents the query vector generated from the image features Z through a transformation by the weight matrix W_q. $K = F_tW_k$ represents the key vector generated from the text features F_t through a transformation by the weight matrix W_k. $V = F_tW_v$ represents the value vector generated from the text features F_t through a transformation by the weight matrix W_v. $K' = F_vW'_k$ represents the visual key vector generated from the image features F_v through a transformation by the weight matrix W'_k. $V' = F_vW'_v$ represents the visual value vector generated from the image features F_v through a transformation by the weight matrix W'_v. Q' and Q'' represent query vectors generated from different transformations of the image features Z, used to capture deeper associations between image and text. K' and K'' represent multiple key vectors of the image features, allowing for more diverse information interaction.

Learnable weighting coefficients γ_1 and γ_2 act as soft gating mechanisms that control the relative influence of different queries and keys. Such gating structures are inspired by improvements to recurrent neural network gating mechanisms, as explored in [47]. The core idea is to perform weighted fusion of multi-source information, enabling dynamic path selection during information flow. This approach has been widely applied in time-series modeling and cross-modal learning tasks, demonstrating effectiveness in modeling long-term dependencies and maintaining semantic consistency.

Additionally, this mechanism benefits from the development of multi-head attention, which introduces learnable weights across multiple attention channels, allowing flexible integration and control of information from different modalities. For instance, in the CLIP

model [37], attention is guided by text-image similarity to steer information fusion. Similarly, diffusion-based generative frameworks like DALL·E 2 [2] and Imagen [48] utilize comparable fusion control strategies to achieve fine-grained alignment between semantic content and visual style.

Q, K, and V are the standard query, key, and value in the cross-attention mechanism, and they interact with the input data through their respective transformation matrices. Q', Q'', and K', K'' introduce additional query and key transformations, enhancing the information interaction between image and text, capturing finer-grained associations. γ_1 and γ_2 are learned weighting coefficients that adjust the influence of different query and key transformations.

We add two loss functions between the segmentation mask and attention feature map to encourage the diffusion model to correctly map text features to the corresponding image attribute regions, as shown in Equation (9):

$$
\begin{aligned}
L_{\text{entity}} &= -\frac{1}{N}\sum_{i=1}^{N} M_{i,u,v}\left(\alpha_i \sum_{k\in n_i\cup a_i}\left(\frac{A^k_{u,v}}{\sum_{x,y} A^k_{x,y}}\right) + \beta_i \log\left(\frac{\sum_{k\in n_i\cup a_i} A^k_{u,v}}{\max(A)}\right)\right) \\
L_{\text{background}} &= -\frac{1}{N}\sum_{i=1}^{N} M_{i,u,v=0}\left(\alpha_i \sum_{k\in n_i\cup a_i}\left(\frac{-A^k_{u,v}}{\sum_{x,y} A^k_{x,y}}\right) + \beta_i \log\left(1-\sum_{k\in n_i\cup a_i}\frac{A^k_{u,v}}{\max(A)}\right)\right) \\
L_{\text{regularization}} &= \lambda_1 \sum_{i=1}^{N}\left(\|\alpha_i\|_2^2 + \|\beta_i\|_2^2\right) + \lambda_2 \sum_{i=1}^{N}\|M_{i,u,v} - M_{i,u,v=0}\|_1
\end{aligned}
\tag{9}
$$

Here, L_{entity} represents the positive loss, aimed at strengthening the model's attention allocation to the entity regions (target regions). $L_{\text{background}}$ represents the negative loss, aimed at reducing attention to the background regions. $M_{i,u,v}$ is the segmentation mask, indicating whether a pixel belongs to the target region, and $A^k_{u,v}$ represents the k-th attention weight at position (u, v). α_i and β_i are dynamic weighting coefficients used to adjust the attention strength and region loss. $\max(A)$ represents the maximum attention weight, used for normalization to avoid numerical instability.

λ_1 and λ_2 are regularization coefficients, controlling the regularization strength of the weighting coefficients and mask differences, respectively. $\|\alpha_i\|_2^2$ and $\|\beta_i\|_2^2$ are the L_2 norms of the weighting coefficients, used to prevent overfitting. $\|M_{i,u,v} - M_{i,u,v=0}\|_1$ is the L_1 norm of the mask difference, helping the model improve generalization capability.

The core process of the ARM module is as follows. First, the target image is generated based on the text prompt. Then, Q-Former is used to extract the attention feature map. Next, a segmentation model is employed to obtain the mask from the image. Finally, the text and visual features are fused, and losses are computed on both the attention map and the mask regions for the entity and background areas respectively – thereby promoting positive attribute alignment and suppressing negative attribute interference. The detailed implementation steps are illustrated in the algorithm, which presents the interaction among ACM, ARM, and DDPM within the AMAC-T2I framework.

4. **Experimental Results.** Our approach is mainly carried out using DDPM (Denoising Diffusion Probabilistic Model) [1], and the evaluation is also conducted on DDPM. DDPM, a generative model relying on a diffusion process, generates high-quality images via a reverse diffusion procedure. It can effectively capture the details and variety of images, thus serving as the fundamental model for our experiments. For the image captioning model, we select BLIP [49], which has been fine-tuned on the COCO [50] image-caption

Algorithm: Interaction between ACM, ARM, and DDPM in AMAC-T2I

Step 1: Semantic and Attribute Feature Extraction

1. Encode the text prompt using a pretrained language model:
$y_{\text{sem}} \leftarrow \text{BERT}(y)$

2. Extract visual features from the reference image using the CLIP encoder:
$l_{\text{clip}} \leftarrow \text{CLIP}(l)$

3. Generate the attribute control vector via ACM, which encodes fine-grained priors such as style or object layout:
$\mathcal{A} \leftarrow \text{ACM}(l_{\text{clip}}, y_{\text{sem}})$

4. Fuse semantic and attribute representations using ARM, which utilizes Q-Former and cross-attention:
$\mathcal{Q} \leftarrow \text{ARM}(l_{\text{clip}}, \mathcal{A}, y_{\text{sem}})$

Step 2: Diffusion-Based Image Generation

5. Initialize a noisy latent image variable by sampling from a Gaussian distribution:
$z_T \sim \mathcal{N}(0, I)$

6. For $t = T$ down to 1, iteratively perform reverse diffusion:
- Predict the noise residual conditioned on $\mathcal{Q}$ and y_{sem}:
$\epsilon_\theta \leftarrow \text{UNet}\,(z_t, \mathcal{Q}, y_{\text{sem}})$
- Update latent variable using DDPM backward step:
$z_{t-1} \leftarrow G.\text{backward}_{step}(z_t, \epsilon_\theta, t)$

7. After T steps, obtain the generated image:
$I_{\text{out}} \leftarrow z_0$

Step 3: Adversarial and Attribute Loss Optimization

8. Compute CLIP-based perceptual loss to encourage alignment with the reference image:
$L_{\text{diff}} \leftarrow \|\text{CLIP}\,(I_{\text{out}}) - \text{CLIP}(l)\|^2$

9. Use a discriminator to assess the semantic consistency between I_{out} and y, and obtain positive/negative contrastive losses:
$L_{\text{pos}}, L_{\text{neg}} \leftarrow \text{DISC}(I_{\text{out}}, y)$

10. Compute the final training objective:
$L_{\text{total}} \leftarrow L_{\text{diff}} + L_{\text{pos}} - L_{\text{neg}}$

11. Update the trainable modules (ACM, ARM, UNet, DISC) using backpropagation based on L_{total}

Return:
Final synthesized image I_{out}

dataset. This model generates accurate image descriptions and is integrated with the generative effects of DDPM.

In the ARM module, we use a pre-trained UNet based on cross-attention as the segmentation mask generator, aiming to optimize the generation process by preserving image details and realism. UNet, with its encoder-decoder architecture, can effectively handle local details in images, ensuring that the quality of the generated image aligns with the original image.

4.1. **Datasets.** The Text-to-Image Comprehensive Benchmark (T2I-CompBench) [51] serves as a pivotal benchmark in the domain of open-world compositional text-to-image generation. It comprises 6,000 carefully crafted synthetic text prompts that are systematically categorized into three major types: attribute binding, object relationships, and complex compositions. Attribute binding includes variations such as color, shape, and

texture associations, while object relationships cover both spatial and non-spatial configurations. The complex composition category focuses on intricate multi-object, multi-attribute scenes. This fine-grained taxonomy provides a structured and diverse foundation for evaluating a model's capability to generate semantically accurate and compositionally coherent images from text. T2I-CompBench thus enables detailed performance assessment across a wide range of compositional challenges.

To further assess generative models in real-world multilingual and structurally diverse settings, this work also incorporates the DPG-Bench (Diverse and Pluralistic Generation Benchmark) [52] as a primary evaluation framework. DPG-Bench includes over 10,000 high-quality prompts spanning six languages – English, Chinese, Spanish, French, German, and Arabic – offering a rigorous platform for testing cross-lingual generalization and compositional reasoning. The prompts vary in complexity and include multi-attribute bindings, object relationships, and culturally grounded scenes with long-form descriptions. Evaluation is conducted using multiple metrics, including CLIP Score for semantic alignment, attribute matching accuracy for fine-grained consistency, and a multilingual alignment score for cross-language understanding. Given its comprehensive design, high diversity, and real-world applicability, DPG-Bench has been widely adopted to benchmark leading generative models such as Stable Diffusion, Emu, Janus, and FLUX, establishing itself as a standard for measuring parameter efficiency and alignment fidelity in text-to-image generation.

4.2. **Evaluation indicators.** The formula adopts FID (Fréchet Inception Distance) to evaluate image quality and CLIP Score to measure semantic consistency. The FID calculation formula is shown in Equation (10).

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{1/2}\right) \tag{10}$$

Here, μ_r, Σ_r represent the mean and covariance matrix of the real image features, μ_g, Σ_g represent the mean and covariance matrix of the generated image features, and Tr denotes the trace of the matrix (i.e., the sum of the diagonal elements).

CLIP Score calculation formula (11) shows

$$\text{CLIP Score} = \frac{\mathrm{Sim}(I, T)}{\|I\| \ \|T\|} \tag{11}$$

Here, I represents the feature embedding vector of the generated image. T represents the feature embedding vector of the input text description. $\mathrm{Sim}(I, T)$ denotes the cosine similarity between the image features and text features, and $\|I\|$ and $\|T\|$ represent the norms of the image features and text features, respectively.

4.3. **Implementation details.** To reduce computational overhead effectively, we optimized only the bottleneck regions of the model by introducing self-attention and cross-attention mechanisms specifically in these components. This design allows the model to capture critical contextual dependencies without imposing excessive computational burden.

For optimization, we employed the Adam optimizer with an initial learning rate set to 1×10^{-4}. A cosine warm-up schedule was applied during the first 5000 steps to gradually increasing the learning rate in the early stages of training, promoting more stable and smoother convergence. The batch size was set to 40, which we found to strike a good balance between computational efficiency and model performance.

To enhance the model's representational capacity, we incorporated a dedicated Q-Former module, whose output is utilized by both the Attribute Correction Module

(ACM) and the Attribute Repair Module (ARM). This design helps reduce computational complexity while maintaining strong feature expressiveness. The Q-Former consists of two parallel submodules: one for semantic encoding guided by text prompts and the other for style encoding based on image attributes. Each submodule is composed of 6 Transformer encoder layers, each with 8 attention heads, a total hidden dimension of 512, and a feed-forward network of dimension 2048 with GELU activation. Layer normalization follows a PreNorm configuration. The model uses 16 learnable query tokens, which interact with image or text features via cross-attention, producing fixed-length embeddings for the downstream generator. The two Q-Formers share the same structure but are independently trained, enabling decoupled modeling of content and style information.

For multimodal fusion, we adopt a dual-branch cross-attention mechanism as described in Equation (8), incorporating two key hyperparameters, γ_1 and γ_2, to control the relative influence of text and image attributes in the attention maps. We performed a grid search over the range 0.1, 0.5, 0.7, 1.2 and empirically selected $\gamma_1 = 0.5$ and $\gamma_2 = 0.7$. This configuration prioritizes semantic alignment while preserving sufficient stylistic flexibility for controllable image synthesis.

All modules, including the cross-attention mechanism, Q-Former, and DISC, are jointly trained from scratch in an end-to-end manner. This enables mutual adaptation across modules and promotes global consistency in the generation process.

All input images are resized to a fixed resolution of 256×256 to ensure consistent input dimensions. Pixel values are normalized to the range $[-1, 1]$, which enhances training stability and accelerates convergence.

The final model is trained for 180 epochs on 8 NVIDIA A100 GPUs. This extensive training schedule ensures that the network captures rich and complex multimodal representations, leading to high-quality image synthesis while maintaining efficient use of computational resources.

4.4. **Ablation experiments.** As shown in Table 1, we conducted a systematic ablation study on the two key components of the AMAC-T2I framework – namely, the Attribute Repair Module (ARM) and the Attribute Correction Module (ACM) – to evaluate their individual and combined contributions to the text-to-image generation task. The results demonstrate that both modules significantly enhance model performance and deliver even greater benefits when used together. When only the ARM module is applied, there is a clear improvement in key attributes such as color, texture, and complex scenes, indicating its strong capability in capturing fine-grained image details and maintaining semantic consistency. Likewise, the ACM module alone yields steady performance gains, particularly in color and shape attributes, confirming its effectiveness in correcting attribute biases and enhancing text-based guidance. Most importantly, the combination of both modules leads to the highest scores across all six subcategories, proving their complementary strengths in attribute modeling. Overall, this modular design greatly improves the model's ability to align fine-grained attributes and text-image semantics, resulting in more accurate and diverse image generation while enhancing adaptability in open-domain scenarios.

This ablation study systematically compares the model's performance under different γ_1 and γ_2 configurations, fully demonstrating the effectiveness and novelty of the proposed gating mechanism in Table 2. The configuration with learnable γ_1 and γ_2 (Experiment 2) significantly outperforms the fixed-weight setup (Group 1) and the no-gating mechanism (Group 5) in terms of semantic consistency measured by the CLIP Score, achieving improvements of approximately 8% and 18%, respectively. This indicates that dynamic weighted fusion enables more precise alignment between text and image semantics, effectively alleviating the limitations of traditional fixed fusion. The image quality metric FID

TABLE 1. The impact of ACM and ARM modules on the overall module. 'ACM' and 'ARM' represent Attribute Correction Module and Attribute Repair Module (ARM), respectively.

Model	ARM	ACM	Attribute binding			Object relationship		Complex
			Color	Shape	Texture	Spatial	Non-spatial	
VDMPDM			0.6990	0.5798	0.6410	0.3242	0.4230	0.4348
VDMPDM	✓		0.8566	0.6154	0.7363	0.3432	0.4282	0.4771
VDMPDM		✓	0.8779	0.6356	0.7423	0.3533	0.4275	0.4764
VDMPDM	✓	✓	0.8938	0.6440	0.7579	0.3540	0.4300	0.4791

TABLE 2. Performance comparison of different γ_1 and γ_2 configurations in the AMAC-T2I framework on T2I-CompBench dataset

Experiment group	γ_1 status	γ_2 status	CLIP score ↑	FID ↓	Attribute accuracy ↑	Attention loss ↓	Convergence speed (epochs) ↓
Fixed equal weights	Fixed at 0.5	Fixed at 0.5	0.72	35.4	78.5%	0.045	50
Learnable (Original design)	Learnable (random init)	Learnable (random init)	0.78	28.7	85.2%	0.031	38
Fixed different weights (0.3/0.7)	Fixed at 0.3	Fixed at 0.7	0.69	37.8	75.3%	0.050	55
Only γ_1 learnable	Learnable	Fixed at 0.5	0.75	31.2	81.0%	0.038	42
Only γ_2 learnable	Fixed at 0.5	Learnable	0.74	32.0	80.5%	0.039	43
No gating mechanism (Ablation)	Disabled	Disabled	0.66	40.1	70.8%	0.060	60

shows a similar trend, with Experiment 2 achieving the lowest score and producing more realistic images. Compared to the no-gating group, FID is reduced by nearly 12 points, highlighting the important role of the gating mechanism in preserving fine-grained attributes and enhancing image realism. The original design group also attains the highest attribute accuracy of 85.2%, outperforming the fixed equal-weight group by 7 percentage points, which fully confirms that the multi-source weighted fusion strategy significantly improves the model's ability to capture and express specific textual attributes. During training, the dynamic weighting group exhibits lower attention loss, further validating the stability and effectiveness of the mechanism in the attribute correction and repair modules, along with faster convergence and improved training efficiency.

To further illustrate the efficacy of the modules, as presented in Figure 2, we visualized the enhancements across all six sub-categories within T2I-CompBench. In the first row, first column, the DXL image depicts a relatively ordinary office scene with plain human interaction and environment, with average detail clarity. The ACM module enhances the focus on attributes, making the interaction between people more vivid and the office environment better focused, resulting in a clearer scene. When both the ARM and ACM modules are applied, the realism and clarity of the image are further improved, with more

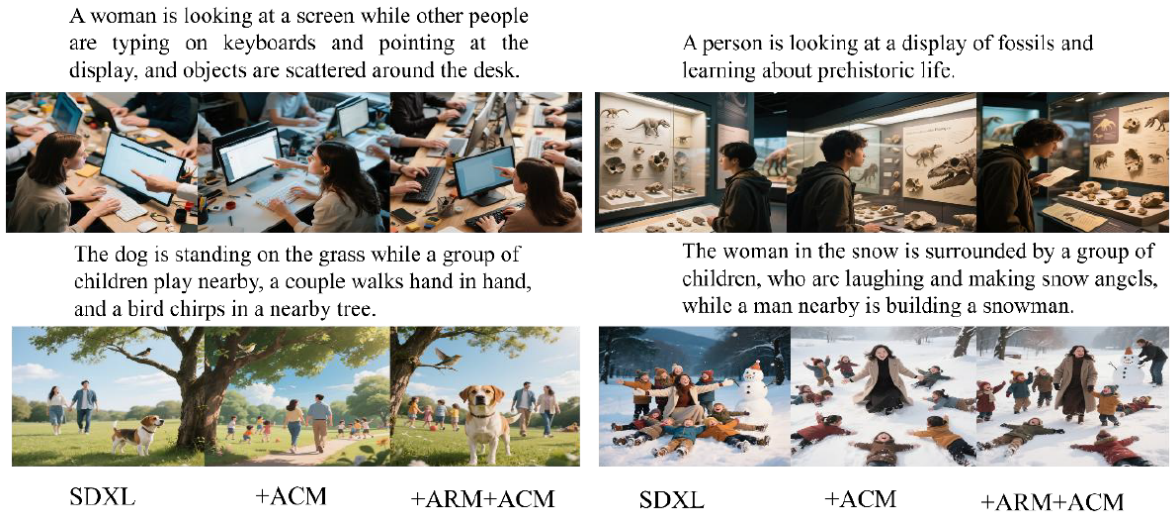

FIGURE 2. Visualization of the effectiveness of the proposed module. ACM helps to show objects mentioned in the prompt. ARM further guides the attribute's attention to the corresponding object on it.

refined interactions between people and more accurate positioning of the office environment and people, making the overall effect more natural and livelier. In the first row, the second column, the SDXL image shows a museum exhibition scene where the fossil display is clear, but the interaction between people and the exhibit is not prominent. The ACM module enhances the focus on the interaction between people and the exhibit, making the scene more vivid and establishing a stronger connection between the people and the exhibit. After applying both ARM and ACM, the focus of the image becomes clearer, highlighting the interaction between people and fossils, while further optimizing the details of the entire scene, making the image appear clearer and more realistic. In the second row, the second column, the SDXL image depicts a snowy scene, but the expression and activity of the people are not fully captured, lacking some subtle emotional expression. The ACM module makes the people and their activities clearer, especially enhancing the interaction between people in activities like snowball fighting and building snowmen. After applying both ARM and ACM, the expressions and activities of the people in the scene are further enhanced, with richer details, more expressive interactions between people, and an overall more vivid and dynamic image.

Table 3 lists the evaluation scores across different languages. From the perspective of parameter size (Param), many models pursue higher performance by using larger parameter counts, such as SD3-Medium with 2.0B parameters, SDXL with 2.6B, Janus-Pro with 7.0B, Emu3 with 8.0B. In contrast, our model has only 0.63B parameters, placing it at a lower level compared to many other models. However, in the DPG-Bench test, our model achieved a score of 80.57, which is competitive compared to models with several times the number of parameters, such as Janus (1.3B parameters, score 79.68) and SDXL (2.6B parameters, score 74.65). This demonstrates good parameter efficiency, achieving a good balance between parameter size and performance.

Regarding multilingual capability (Multi-Ling), many models lack this feature (marked as $\times$), such as SD1.5, and SD2.1. Our model, along with a few others like Sana, supports multilingual capabilities (marked as $\surd$). This means our model has an inherent advantage in handling multilingual tasks, making it more versatile and capable of meeting the needs of users from different language backgrounds. It is ahead of most comparison models in applications like multilingual content generation. In terms of DPG-Bench test scores,

Table 3. Model comparison based on DPG-Bench results

Method	Param	Multi-Ling	DPG-Bench
SD1.5 [5]	0.86B	×	61.18
LlamaGen [53]	0.78B	×	65.16
HART [54]	0.73B	×	80.89
Sana [55]	0.60B	√	83.6
ELLA [52]	0.93B	×	80.79
LLM4GEN [56]	0.86B	×	67.34
Pixart-α [57]	0.61B	×	71.11
MSCA [25]	0.81B	×	68.35
SD3-Medium [58]	2.0B	×	84.08
SDXL [59]	2.6B	×	74.65
Janus [60]	1.3B	×	79.68
Janus-Pro [61]	7.0B	×	84.19
Emu3 [62]	8.0B	×	80.60
DALL·E 3 [63]	–	×	83.50
Ours	0.63B	√	80.57

while some models like SD3-Medium (84.08) and Janus-Pro (84.19) slightly outperform our model, these models either have enormous parameter sizes with high training and deployment costs or lack multilingual capabilities.

To further demonstrate the advantages of our model, we visualize images from different scenes, as shown in Figure 3. In the first row, in the first column, the woman is not hugging the bear; in the second column, there is no dog; and in the third column, the man is not reading a book. Our model, however, accurately matches the content description. This indicates that the other models fail to fully capture key information when understanding the text semantics and miss important concepts during the process of converting text into images, showing their limitations in grasping the overall semantics of the text and mapping it to images. In the fourth row, in the first column, the man and woman are sitting separately. Other columns also lack elements corresponding to the text, while only our model presents the elements from the description. This reveals that other models miss the crucial concept of "intimate interaction", reflecting their poor understanding and representation of the relationships between elements in the scene.

In the second row, other models generate images where the color and style of the wall tiles do not match the description, resulting in color errors or mismatched textures. This is a case of object attribute mismatching. It indicates that these models exhibit biases when processing object attribute information and fail to correctly interpret the text's description of object features. Our model, on the other hand, accurately maps the text content.

As shown in Table 4, we compare several representative diffusion models from recent years, evaluating their performance using key metrics including CLIP Score, Fréchet Inception Distance (FID), Attribute Accuracy (Attr. Acc.), and the number of inference steps on the T2I-CompBench benchmark. The results clearly demonstrate the overall superiority of the proposed AMAC-T2I framework in text-to-image generation tasks.

First, in terms of text-image semantic alignment, AMAC-T2I achieves the highest CLIP Score of 0.78, outperforming Matryoshka (0.75), SDXL (0.74), and STAY Diffusion (0.73), indicating a stronger semantic consistency between generated images and input prompts. Additionally, for fine-grained attribute control, AMAC-T2I attains an attribute accuracy of 85.2%, which significantly exceeds that of SDXL (80.5%) and Matryoshka (81.7%), and

FIGURE 3. Qualitative comparisons of our model with the Pixart-α [62], Anydoor [40], and Tend-and-excite [30] models

TABLE 4. Performance comparison of recent diffusion models (2023-2025) on T2I-CompBench. Muse generates discrete tokens, so CLIP Score is not directly comparable. Muse reports FID on CC3M subset, which is significantly lower due to domain constraints. DiffiT reports FID on ImageNet-256.

Model	CLIP Score ↑	FID ↓	Attr. Acc. ↑	Steps ↓
SDXL [59]	0.74	27.8	80.5%	50
Muse [64]	0.32	6.06	–	25
DiffiT [65]	–	1.73	–	–
Matryoshka [66]	0.75	26.7	81.7%	35
STAY Diffusion [67]	0.73	29.1	83.4%	40
Diffusion-4K [68]	–	24.5	–	50
AMAC-T2I (Ours)	**0.78**	**28.7**	**85.2%**	**38**

even surpasses the strongest competing method STAY Diffusion (83.4%). This validates the effectiveness of our model in modeling attribute correspondence and achieving precise control over visual semantics.

Moreover, AMAC-T2I demonstrates competitive efficiency while maintaining high generation quality. Although Muse reports a low FID of 6.06, its evaluation is limited to the CC3M subset, and its CLIP Score is only 0.32, making it less suitable for general-purpose comparison. DiffiT reports an FID of 1.73 on ImageNet-256, which is not directly comparable to the other models evaluated on T2I-CompBench. In contrast, AMAC-T2I

achieves a balanced performance with an FID of 28.7 and reduces the sampling steps to 38, which is more efficient than SDXL and Diffusion-4K (both requiring 50 steps), thus maintaining a good trade-off between image quality and inference speed.

5. **Conclusion.** This paper presents AMAC-T2I, a diffusion-based framework for text-to-image generation with a specific focus on attribute-level precision and semantic alignment. To address common issues such as attribute mismatch and omission, we introduce two key components: the Attribute Correction Module (ACM) and the Attribute Repair Module (ARM). The ACM leverages adversarial loss and multimodal priors from Q-Former to correct attribute inconsistency during generation, while the ARM integrates Q-Former with a cross-attention-enhanced UNet to repair missing or weakly expressed attributes by generating fine-grained category masks.

Extensive experiments across multiple benchmarks – including compositional datasets like T2I-CompBench and multilingual settings such as DPG-Bench – demonstrate that AMAC-T2I achieves state-of-the-art performance in attribute accuracy, controllability, and cross-lingual generalization. Our model consistently outperforms baselines in both CLIP alignment and attribute fidelity, while maintaining competitive inference efficiency. These results highlight the effectiveness of our design in addressing the long-standing challenge of controllable and semantically aligned image generation from textual prompts.

However, our approach also presents several limitations. First, although AMAC-T2I demonstrates superior performance in attribute-level alignment, its FID score remains slightly higher than some models optimized specifically for image realism, such as token-based architectures (e.g., Muse) or super-resolution-focused methods (e.g., Diffusion-4K). Second, while Q-Former enables modular multimodal encoding, it introduces additional memory and training complexity that may limit deployment in lightweight or real-time applications. Third, AMAC-T2I currently relies on predefined category masks in ARM, which may constrain its adaptability in fully open-domain scenarios where category labels are unknown or ambiguous.

In future work, we plan to explore dynamic category-aware repair strategies, lightweight Q-Former variants, and prompt-adaptive diffusion decoding to further enhance both efficiency and scalability. Despite its current limitations, AMAC-T2I offers a promising and generalizable architecture for precise and interpretable text-to-image generation, especially in scenarios that demand fine-grained attribute control and semantic robustness.

REFERENCES

[1] J. Ho, A. Jain and P. Abbeel, Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems*, vol.33, pp.6840-6851, 2020.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, Hierarchical text-conditional image generation with clip latents, *arXiv Preprint*, arXiv: 2204.06125, 2022.

[3] D. Meena, H. Katragadda, K. Narva, A. Rajesh and J. Sheela, Text-conditioned image synthesis using TAC-GAN: A unique approach to text-to-image synthesis, *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp.454-462, 2023.

[4] T. Wu, Y. Xu, R. Po, M. Zhang, G. Yang, J. Wang, Z. Liu, D. Lin and G. Wetzstein, FiVA: Fine-grained visual attribute dataset for text-to-image diffusion models, *Advances in Neural Information Processing Systems*, vol.37, pp.31990-32011, 2024.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, High-resolution image synthesis with latent diffusion models, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10684-10695, 2022.

[6] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein and K. Aberman, DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.22500-22510, 2023.

[7] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Y. Zhu and S. Ermon, SDEdit: Guided image synthesis and editing with stochastic differential equations, *arXiv Preprint*, arXiv: 2108.01073, 2021.
[8] A. Cuzzocrea and E. Fadda, Modeling and supporting adaptive complex data-intensive web systems via XML and the OO paradigm: The OO-XAHM model, *Array*, vol.23, 100363, 2024.
[9] B.-H. Tsai, Assessment of IC clustering evolution by using a novel diffusion model and a genetic algorithm, *International Journal of Innovative Computing, Information and Control*, vol.9, no.4, pp.1493-1510, 2013.
[10] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch and D. Cohen-Or, Prompt-to-prompt image editing with cross attention control, *arXiv Preprint*, arXiv: 2208.01626, 2022.
[11] W. Feng, X. He, T. J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X. E. Wang and W. Y. Wang, Training-free structured diffusion guidance for compositional text-to-image synthesis, *arXiv Preprint*, arXiv: 2212.05032, 2022.
[12] L. Zhao, J. Wu and Y. Jia, Generative adversarial network with new batch normalization and feature extraction block for image super-resolution reconstruction, *International Journal of Innovative Computing, Information and Control*, vol.19, no.2, pp.385-401, 2023.
[13] R.-C. Chen, C. Sub-r-pa, M.-Z. Fan and H. Yu, Local expression diffusion for facial expression synthesis, *International Journal of Innovative Computing, Information and Control*, vol.20, no.1, pp.283-295, 2024.
[14] L. Zhao, Y. Qin and Y. Jia, Generative adversarial networks for remote sensing image dehazing with color feature restoration, *International Journal of Innovative Computing, Information and Control*, vol.21, no.2, pp.323-338, 2025.
[15] N. Liu, S. Li, Y. Du, A. Torralba and J. B. Tenenbaum, Compositional visual generation with composable diffusion models, *European Conference on Computer Vision*, pp.423-439, 2022.
[16] Z. He, T. Sun, K. Wang, X. Huang and X. Qiu, DiffusionBERT: Improving generative masked language models with diffusion models, *arXiv Preprint*, arXiv: 2211.15029, 2022.
[17] M. Hamdan and M. Cheriet, ResneSt-Transformer: Joint attention segmentation-free for end-to-end handwriting paragraph recognition model, *Array*, vol.19, 100300, 2023.
[18] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever and M. Chen, GLIDE: Towards pho-torealistic image generation and editing with text-guided diffusion models, *arXiv Preprint*, arXiv: 2112.10741, 2021.
[19] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato and E. Raff, VQGAN-CLIP: Open domain image generation and editing with natural language guidance, *European Conference on Computer Vision*, pp.88-105, 2022.
[20] K. Sohn, N. Ruiz, K. Lee, D. C. Chin, I. Blok, H. Chang, J. Barber, L. Jiang, G. Entis, Y. Li, Y. Hao, I. Essa, M. Rubinstein and D. Krishnan, StyleDrop: Text-to-image generation in any style, *arXiv Preprint*, arXiv: 2306.00983, 2023.
[21] L. Zhang, A. Rao and M. Agrawala, Adding conditional control to text-to-image diffusion models, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.3836-3847, 2023.
[22] O. Avrahami, D. Lischinski and O. Fried, Blended diffusion for text-driven editing of natural images, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.18208-18218, 2022.
[23] H. Chefer, S. Benaim, R. Paiss and L. Wolf, Image-based clip-guided essence transfer, *European Conference on Computer Vision*, pp.695-711, 2022.
[24] T. Brooks, A. Holynski and A. A. Efros, InstructPix2Pix: Learning to follow image editing instructions, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.18392-18402, 2023.
[25] L. Ma, T. Gao, H. Shen and K. Huang, Multi-scale cross-domain alignment for person image generation, *CAAI Transactions on Intelligence Technology*, vol.9, no.2, pp.374-387, 2024.
[26] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan and B. Guo, Vector quantized diffusion model for text-to-image synthesis, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10696-10706, 2022.
[27] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg and G. Chechik, Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment, *Advances in Neural Information Processing Systems*, vol.36, pp.3536-3559, 2023.
[28] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf and D. Cohen-Or, Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, *ACM Transactions on Graphics (TOG)*, vol.42, no.4, pp.1-10, 2023.

[29] X. Liu, C. Gong and Q. Liu, Flow straight and fast: Learning to generate and transfer data with rectified flow, *arXiv Preprint*, arXiv: 2209.03003, 2022.

[30] W. Peebles and S. Xie, Scalable diffusion models with transformers, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.4195-4205, 2023.

[31] X. Jing, Y. Chang, Z. Yang, J. Xie, A. Triantafyllopoulos and B. W. Schuller, U-DiT TTS: U-diffusion vision transformer for text-to-speech, *The 15th ITG Conference on Speech Communication*, pp.56-60, 2023.

[32] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu and Z. Li, Pixart-σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation, *European Conference on Computer Vision*, pp.74-91, 2024.

[33] Y. Song, Z. Zhang, Z. Lin, S. Cohen, B. Price, J. Zhang, S. Y. Kim and D. Aliaga, ObjectStitch: Generative object compositing, *arXiv Preprint*, arXiv: 2212.00932, 2022.

[34] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen and F. Wen, Paint by Example: Exemplar-based image editing with diffusion models, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.18381-18391, 2023.

[35] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao and H. Zhao, AnyDoor: Zero-shot object-level image customization, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6593-6602, 2024.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, Learning transferable visual models from natural language supervision, *International Conference on Machine Learning*, pp.8748-8763, 2021.

[37] L. Hu, Animate Anyone: Consistent and controllable image-to-video synthesis for character animation, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8153-8163, 2024.

[38] N. Tumanyan, M. Geyer, S. Bagon and T. Dekel, Plug-and-play diffusion features for text-driven image-to-image translation, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.1921-1930, 2023.

[39] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li and Y. J. Lee, GLIGEN: Open-set grounded text-to-image generation, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.22511-22521, 2023.

[40] M. Chen, I. Laina and A. Vedaldi, Training-free layout control with cross-attention guidance, *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.5343-5353, 2024.

[41] J. Sun, D. Fu, Y. Hu, S. Wang, R. Rassin, D. C. Juan, D. Alon, C. Herrmann, S. van Steenkiste, R. Krishna and C. Rashtchian, DreamSync: Aligning text-to-image generation with image understanding feedback, *arXiv Preprint*, arXiv: 2311.17946, 2023.

[42] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee and K. Lee, Reinforcement learning for fine-tuning text-to-image diffusion models, *The 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[43] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang and Y. Dong, ImageReward: Learning and evaluating human preferences for text-to-image generation, *Advances in Neural Information Processing Systems*, vol.36, pp.15903-15935, 2023.

[44] J. Li, D. Li, S. Savarese and S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, *International Conference on Machine Learning*, pp.19730-19742, 2023.

[45] D. Li, J. Li and S. Hoi, BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, *Advances in Neural Information Processing Systems*, vol.36, pp.30146-30166, 2023.

[46] T. Qi, S. Fang, Y. Wu, H. Xie, J. Liu, L. Chen, Q. He and Y. Zhang, DEADiff: An efficient stylization diffusion model with disentangled representations, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8693-8702, 2024.

[47] A. Gu, C. Gulcehre, T. Paine, M. Hoffman and R. Pascanu, Improving the gating mechanism of recurrent neural networks, *International Conference on Machine Learning*, pp.3800-3809, 2020.

[48] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. Jfleet and M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding, *Advances in Neural Information Processing Systems*, vol.35, pp.36479-36494, 2022.

[49] J. Li, D. Li, C. Xiong and S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *International Conference on Machine Learning*, pp.12888-12900, 2022.

[50] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, Microsoft COCO: Common objects in context, *Computer Vision – ECCV 2014: The 13th European Conference*, Zurich, Switzerland, pp.740-755, 2014.

[51] K. Huang, K. Sun, E. Xie, Z. Li and X. Liu, T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation, *Advances in Neural Information Processing Systems*, vol.36, pp.78723-78747, 2023.

[52] X. Hu, R. Wang, Y. Fang, B. Fu, P. Cheng and G. Yu, ELLA: Equip diffusion models with LLM for enhanced semantic alignment, *arXiv Preprint*, arXiv: 2403.05135, 2024.

[53] P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo and Z. Yuan, Autoregressive model beats diffusion: Llama for scalable image generation, *arXiv Preprint*, arXiv: 2406.06525, 2024.

[54] H. Tang, Y. Wu, S. Yang, E. Xie, J. Chen, J. Chen, Z. Zhang, H. Cai, Y. Lu and S. Han, HART: Efficient visual generation with hybrid autoregressive transformer, *arXiv Preprint*, arXiv: 2410.10812, 2024.

[55] E. Xie, J. Chen, J. Chen, H. Cai, H. Tang, Y. Lin, Z. Zhang, M. Li, L. Zhu, Y. Lu and S. Han, SANA: Efficient high-resolution image synthesis with linear diffusion transformers, *arXiv Preprint*, arXiv: 2410.10629, 2024.

[56] M. Liu, Y. Ma, Z. Yang, J. Dan, Y. Yu, Z. Zhao, Z. Hu, B. Liu and C. Fan, LLM4GEN: Leveraging semantic representation of LLMs for text-to-image generation, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.39, no.5, pp.5523-5531, 2025.

[57] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu and Z. Li, Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, *arXiv Preprint*, arXiv: 2310.00426, 2023.

[58] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English and R. Rombach, Scaling rectified flow transformers for high-resolution image synthesis, *The 41st International Conference on Machine Learning*, 2024.

[59] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna and R. Rombach, SDXL: Improving latent diffusion models for high-resolution image synthesis, *arXiv Preprint*, arXiv: 2307.01952, 2023.

[60] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan and P. Luo, Janus: Decoupling visual encoding for unified multimodal understanding and generation, *arXiv Preprint*, arXiv: 2410.13848, 2024.

[61] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu and C. Ruan, Janus-Pro: Unified multimodal understanding and generation with data and model scaling, *arXiv Preprint*, arXiv: 2501.17811, 2025.

[62] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, Y. Zhao, Y. Ao, X. Min, T. Li, B. Wu, B. Zhao, B. Zhang, L. Wang, G. Liu, Z. He, X. Yang, J. Liu, Y. Lin, T. Huang and Z. Wang, *Emu3: Next-Token Prediction Is All You Need*, https://doi.org/10.48550/arXiv.2409.18869, 2024.

[63] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee and Y. Guo, Improving image generation with better captions, *Comput. Sci.*, vol.2, no.3, 8, https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

[64] H. Chang, H. Zhang, J. Barber, A. J. Maschinot, J. Lezama, L. Jiang, M. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li and D. Krishnan, Muse: Text-to-image generation via masked generative transformers, *arXiv Preprint*, arXiv: 2301.00704, 2023.

[65] A. Hatamizadeh, J. Song, G. Liu, J. Kautz and A. Vahdat, DiffiT: Diffusion vision transformers for image generation, *European Conference on Computer Vision*, pp.37-55, 2024.

[66] J. Gu, S. Zhai, Y. Zhang, J. Susskind and N. Jaitly, Matryoshka diffusion models, *arXiv Preprint*, arXiv: 2310.15111, 2023.

[67] R. Wang, X. Hou, S. Schmedding and M. F. Huber, STAY Diffusion: Styled layout diffusion model for diverse layout-to-image generation, *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp.3855-3865, 2025.

[68] J. Zhang, Q. Huang, J. Liu, X. Guo and D. Huang, Diffusion-4K: Ultra-high-resolution image synthesis with latent diffusion models, *Proc. of the Computer Vision and Pattern Recognition Conference*, pp.23464-23473, 2025.

Author Biography

Zhibo Dai received Bachelor's degree in Computer Application from Henan University of Science and Technology, China, in 2000, Master's degree in Computer Application Technology from Hohai University, China, in 2006. He is a senior engineer, and worked in Nanjing University of Science and Technology ZiJin College, China. He was awarded Associate Senior Title, skilled in full-stack tech, AI integration, PMP-certified, CMMI5/agile expertise, and also awarded for R&D excellence. His research interests include computer vision, and machine learning area.

Ping Zong is a professor. He received Ph.D. degree from Hohai University, China, in 2008. He worked in Nanjing University of Posts and Telecommunications, China. He is a visiting scholar in Germany, an IEEE/ACM member, and national review expert. He authored 150+ papers, 3 books (1 national textbook), and awarded teaching/research prizes. His research interests include software engineering and artificial intelligence.

Yuhua Cong is a university lecturer. She received Master's degree from Nanjing University of Aeronautics and Astronautics, China, in 2009. She worked in Nanjing University of Aeronautics and Astronautics, China. Her research interests include computer vision, and UAV cluster planning.

Mengyu Xu is a university lecturer. She received Master's degree from Nanjing University of Science and Technology, China, in 2019. She worked in Nanjing University of Science and Technology ZiJin College, China. Her research interests include picture processing.

Wenhan Sun is a teaching assistant. He received Master's degree from University of Western Ontario, Canada, in 2023. He worked in Nanjing University of Science and Technology ZiJin College, China. His research interests include image compression.