AnyScene: Customized Image Synthesis with Composited Foreground

Ruidong Chen[1†] Lanjun Wang[1†] Weizhi Nie[1] Yongdong Zhang[2] An-An Liu[1*]
[1]Tianjin University [2]University of Science and Technology of China

Abstract

*Recent advancements in text-to-image technology have significantly advanced the field of image customization. Among various applications, the task of customizing diverse scenes for user-specified composited elements holds great application value but has not been extensively explored. Addressing this gap, we propose **AnyScene**, a specialized framework designed to create varied scenes from composited foreground using textual prompts. AnyScene addresses the primary challenges inherent in existing methods, particularly scene disharmony due to a lack of foreground semantic understanding and distortion of foreground elements. Specifically, we develop a foreground injection module that guides a pre-trained diffusion model to generate cohesive scenes in visual harmony with the provided foreground. To enhance robust generation, we implement a layout control strategy that prevents distortions of foreground elements. Furthermore, an efficient image blending mechanism seamlessly reintegrates foreground details into the generated scenes, producing outputs with overall visual harmony and precise foreground details. In addition, we propose a new benchmark and a series of quantitative metrics to evaluate this proposed image customization task. Extensive experimental results demonstrate the effectiveness of AnyScene, which confirms its potential in various applications.*

1. Introduction

Recently, text-to-image (T2I) synthesis models [24, 28, 30, 34] have seen rapid advancements and gained significant attention due to their ability to generate high-fidelity images from textual prompts. Among the various applications that harness T2I technology to enhance design efforts, generating diverse scenes tailored to specific composited foregrounds is valuable. It is effective in image editing tasks demanding background alteration and creation, such as situating e-commerce products in customized scenes, depicting objects in various environmental settings, etc. Despite its clear utility, this task of *"customizing diverse scenes for composited foreground"* has not been extensively explored.

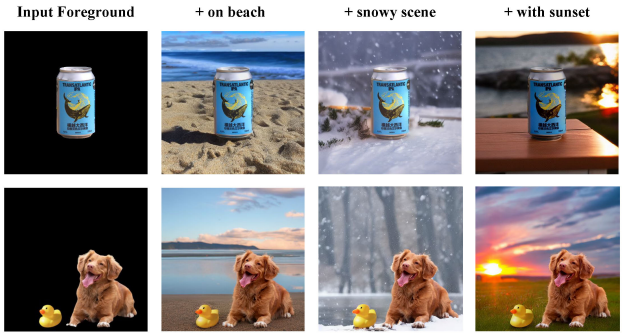

Figure 1. Our proposed AnyScene is capable of synthesizing high-quality scenes that align with textual prompts based on the given foregrounds. Compared to previous alternative methods [42, 45], AnyScene provides precise control over the introduced foreground and generates visually harmonious images.

Currently, several alternative methods [42, 45] can achieve this task. One such method is employing the canny edge as a control condition [23, 45], combined with overlaying the original foreground onto the generated image, often requiring manual adjustments to achieve visual harmony. Another strategy involves inpainting models [30, 42] that construct backgrounds by painting around the exterior of a masked foreground area. Furthermore, in cases where the foreground comprises only one or two specific objects, subject-driven techniques [17, 19, 32] have shown proficiency in generating customized images.

However, these methods have limitations in practical applications. The method of directly controlling the foreground edge [45] does not consider the visual context, leading to scenes that lack harmony (e.g. Fig. 2(a)). Moreover, inpainting-based approaches [30, 42], though capable of creating cohesive scenes, are commonly trained under a random area recovering paradigm, which may distort the foreground elements or cause unintended regeneration of elements from foreground descriptions (e.g. Fig. 2(b)). Finally, subject-driven methods [17, 19, 32] often compromise high-level details of objects (e.g. Fig. 2(c)), leading to significant distortions of details such as logos, text, etc, which is problematic for the applications like commodity poster design. Meanwhile, their limited capacity to com-

[1†] Equal contribution.
[2*] Corresponding author: An-An Liu (anan0422@gmail.com).

a) unharmonized scene b) redundant element c) low details preserving

Figure 2. The visualized examples of challenges faced by current alternatives [17, 42, 45] with the proposed task.

pose multiple objects into a complex scene also restricts their applicability.

To address these limitations, we present ***AnyScene***, a customized image synthesizing framework tailored explicitly for *"customizing diverse scenes for composited foreground"*. Firstly, facing the challenge of generating a visually harmonious scene based on the specific foreground and text prompt, we establish a training paradigm that reconstructs the overall scene considering the given foreground elements and propose the foreground injection module to guide pre-trained latent diffusion models in producing the images of overall scenes that cohesively incorporate the specific foreground. Secondly, to minimize severe distortion of the foreground and prevent redundant generation of foreground elements during synthesis, we implement the layout control strategy, which ensures that foreground information remains confined to its designated areas during model inference. Finally, acknowledging that generated images often fail to preserve intricate visual details of specific objects due to perceptual compression [9] inherent in latent diffusion models [30], it is necessary to recover the original details of the foreground onto the generated images. Unlike alternative methods [42, 45] that typically directly overlay the foreground on the generated image, we employ an image blending mechanism to achieve a visually harmonious integration, considering the lighting and color conditions of the generated images.

We conduct extensive experiments to evaluate the generative capability of AnyScene on the proposed task. We propose a benchmark and a series of evaluation metrics to assess the generative capacity of methods on this task. Additionally, we perform a detailed visualization analysis to compare the results produced by different approaches. These experiments conclusively demonstrate the effectiveness of AnyScene in synthesizing high-quality customized images.

2. Related Work

2.1. Text-to-Image Synthesis

Text-to-image (T2I) synthesis techniques aim to generate images from given textual descriptions. Early research [29, 40, 43, 44] employed Generative Adversarial Networks (GANs) [11] on text-image datasets. These initial methods suffered from training instability[1, 37] and a lack of high-quality datasets, which posed challenges in producing high-fidelity images. Subsequently, models based on autoregressive architectures [7, 8, 10, 27, 38, 41], trained on expansive datasets [35, 36], have made notable strides in generation quality, significantly advancing T2I synthesis research with their enhanced generative capabilities.

Recently, the study of denoising diffusion probabilistic models (DDPMs) [6, 14, 15] has significantly advanced the field of T2I technology. DDPMs utilize a process that progressively adds random noise to the original image and then iteratively removes this noise during inference to produce various high-quality images. Pioneering works such as [24, 30, 33, 34] have built powerful T2I models based on DDPMs, achieving remarkable improvements in the fidelity and clarity of the synthesized images. This progress has led to a range of innovations in image customization works, leveraging the capabilities of pre-trained diffusion models for practical applications, such as subject-driven generation [17, 19, 32, 39] and text-driven image editing [2, 12].

2.2. Controling Text-to-Image Diffusion Models

With the rapid advancement of Text-to-Image (T2I) diffusion models, recent studies have significantly advanced in precisely controlling the generative content of these models. In order to incorporate specific visual references into generative models, works such as [19, 21, 39] utilize CLIP-based visual adapters to introduce visual concepts from reference images into the image generation process. For layout control, some studies [23, 45] introduce additional conditions like sketches, depth maps, or canny edges to pre-trained diffusion models, constraining models to generate images conforming to these layouts. Additionally, attention control methods [5, 12] have achieved image editing or layout guiding capabilities without the need for any training by editing cross-attention maps between text tokens and intermediate visual features during the model's inference process.

3. Preliminary

We adopt the Stable Diffusion model [30] for its proven stability and efficacy in text-to-image synthesis tasks. Our method aims to utilize the powerful generative capabilities of the pre-trained diffusion model to customize diverse scenes based on the given foregrounds guided by textual prompts. Before introducing our method, it is essential first to discuss the mechanism of Stable Diffusion, which forms the preliminary basis of our method.

The Stable Diffusion is based on the latent diffusion model [30]. It operates diffusion and denoise processes in an autoencoder's latent space. Formally, it uses a pre-trained encoder $\mathcal{E}$ to compress images x into smaller "latent images" z for stabilized training. Then, the model features

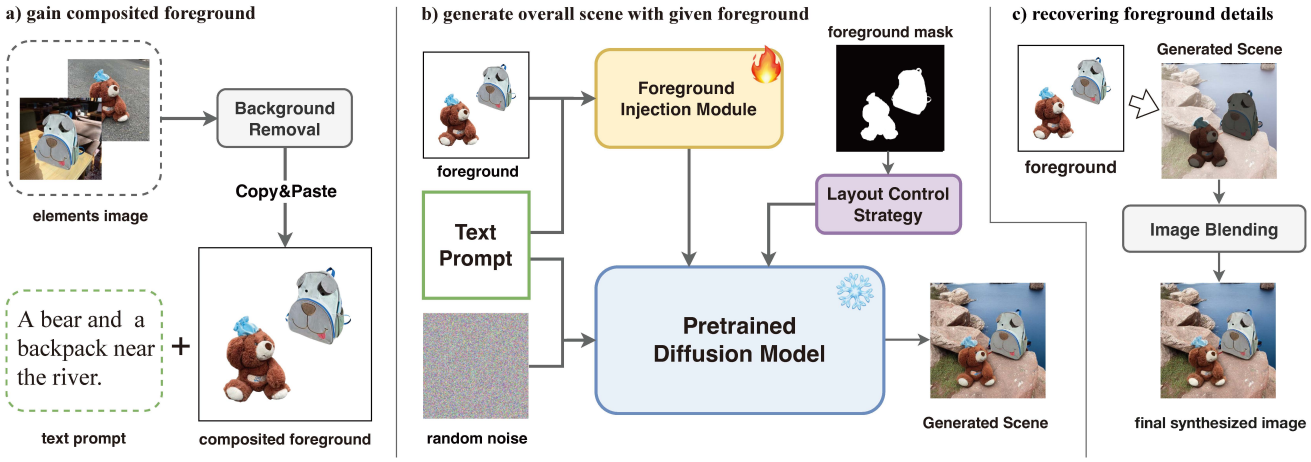

Figure 3. The overall framework of our proposed AnyScene, which is designed to synthesize diverse scenes with user-specified foreground elements. a) AnyScene begins with gaining a composited foreground and text prompts for the target scene. b) Then, the Foreground Inject Module(Sec. 4.1), alongside a Layout Control Strategy(Sec. 4.2), guides a pre-trained diffusion model to generate the scene contextually. c) Finally, the generated scene undergoes a Foreground Blending(Sec. 4.3) process to recover the foreground details, resulting in a harmoniously synthesized final image.

a denoising U-Net [31] $\mathbb{E}$ to predict the reverse diffusion process in the latent space. A key aspect of the model is its conditioning on textual inputs through a CLIP [26] text encoder τ_θ that translates the text condition y into an intermediate form $\tau_\theta(y)$, which is integrated into the model via a cross-attention mechanism in the U-Net layers:

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \qquad (1)$$

where $Q = W_Q \cdot \varphi(z_t), K = W_K \cdot \tau_\theta(y), V = W_V \cdot \tau_\theta(y)$, and $\varphi(z_t)$ represents the hidden states in U-Net with z_t as the diffused latent representation at time t. The training objective of the latent diffusion model is to predict the noise ϵ added to the latent image representation, as formulated in:

$$\mathcal{L}_{\text{denoising}} = \mathbb{E}_{z,t,y,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right], \quad (2)$$

where ϵ is the added noise for the diffusion process, ϵ_θ indicates the U-Net's parameters which are used to predict this noise conditioned with text y.

During inference, the model starts with a random Gaussian noise sample z_T and iteratively denoises it over T steps to arrive at the estimated initial code $\hat{z}_0$. Finally, the image decoder $\mathcal{D}(\cdot)$ is applied to reconstruct this latent code $\hat{z}_0$ into the final image $\hat{x}$ with high fidelity.

4. Methodology

This study focuses on the task of "*customizing diverse scenes for composited foreground*". As illustrated in Fig. 3 (a), this task has two inputs: 1) the specific foreground elements onto the blank canvas denoted c_f; 2) a text prompt of the target scene, denoted y. Given these inputs, the goal of our proposed task consists of the following requirements.

- Generate scenes that are visually cohesive with the provided foreground and textual description;
- Prevent severe distortion and redundant generation of foreground elements during scene creation;
- Preserve the precise details of the given foreground and blend it into the generated image naturally.

Based on these requirements, we introduce AnyScene, as illustrated in Fig. 3. We propose a comprehensive framework based on three modules designed to address each of the points above, ultimately synthesizing images that align with user input and uphold overall visual harmony and the accuracy of the original foreground details.

4.1. Foreground Injection Module

We propose the Foreground Injection Module F_c based on the ControlNet [45] architecture to introduce the foreground information into the pre-trained diffusion model. F_c employs the same encoder blocks as the diffusion U-Net to extract detailed information from c_f and learns the correlation between the input text y and the foreground elements. By adding the learned intermediate features into the diffusion model's latent space, F_c guides the frozen diffusion models to generate a cohesive overall scene with the input foreground and text prompt.

To facilitate the training of the Foreground Injection Module F_c, we construct the training data, utilizing the foreground image c_f with its target scene description y to reconstruct the original image x and learn the module's foreground integration capabilities. During training, we apply random color enhancements to the input foregrounds,

encouraging F_c to guide the diffusion model to generate visually harmonious scenes, even with variable foreground colors and lighting conditions. To ensure effective foreground integration, we apply the following three parts of training objectives for F_c.

Denoising Loss. We adopt the standard denoising loss from latent diffusion models [15, 30]. During the denoising training phase of latent diffusion models, the denoising loss aims to reconstruct the initial latent code z_0 from the diffused code z_t in timestep t by predicting the added noise. In our cases, the denoising loss can be formulated as:

$$\mathcal{L}_{denoising} = \mathbb{E}_{z_0,t,y,c_f,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y), F_c(c_f)\right)\|_2^2\right], \tag{3}$$

where the ϵ is the random sampling noise to simulate the diffusion process, we leverage the U-Net's parameters ϵ_θ to predict this added noise in conditions with y and c_f.

Content Loss. We apply a pixel-level foreground content loss to enhance the color and texture learning with the given foreground elements. At timestep t, we reverse the diffused latent z_t to their estimated initial state $\hat{z}_t$, which is formulated as: $\hat{z}_t = \left(z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(z_t, t, y, c_f\right)\right)/\sqrt{\bar{\alpha}_t}$, where $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s, \alpha_s = 1-\beta_s$ and β_s represents forward process variances [15]. From this, the reconstructed estimated image is denoted as $\hat{x}_t = \mathcal{D}(\hat{z}_t)$, where $\mathcal{D}(\cdot)$ is the image decoder as defined in Sec. 3. The Content Loss is then applied using the foreground binary mask M to align the reconstructed and original images within the foreground region:

$$\mathcal{L}_{content} = \|x_0 - \hat{x}_t\|_2^2 \odot M. \tag{4}$$

With this training objective, F_c can effectively learn the color and texture information of the given foreground. Particularly, incorporating the color enhancement applied on the foreground elements during training, the introduction of this loss helps adjust the foreground colors according to the overall generated scene in response to various lighting and color conditions, thereby improving the visual harmony of the generated scene.

Edge Gradient Loss. In conventional image processing, identifying foreground boundaries often relies on detecting changes in color gradients within the image. These changes are typically pronounced at the edges of foreground elements due to variations in lighting, color, and depth of field. Inspired by this, we propose an edge gradient loss, which encourages F_c to generate soft and realistic foreground boundaries by learning the gradient changes of foreground boundaries in real images. To this end, we generate an edge mask $M_{edge} = \text{dilate}(M) - \text{erode}(M)$, which is obtained by subtracting the eroded mask from the dilated mask to capture the morphological edges of the foreground. Subsequently, we utilize the Laplacian operator to calculate the color transition gradient in the RGB channels of the original image x_0 and the estimated image $\hat{x}_t$. The formula for aligning gradients at the edges between the two images is as follows:

$$\mathcal{L}_{\text{gradient}} = \|\nabla x_0 - \nabla \hat{x}_t\|_2^2 \odot M_{edge}, \tag{5}$$

where ∇ denotes the Laplacian operator. Incorporating this loss enables F_c to refine image generation from two aspects: 1) clarify the edges around the foreground elements; 2) align edge gradients to preserve the foreground shape, which could minimize distortions in the generated images.

Training Strategies. Building upon the previously introduced concepts, the Foreground Injection Module F_c is trained using the foundational denoising loss in conjunction with the newly proposed losses. Recognizing that the model cannot reconstruct the original image in the early stages of denoising, we adjust our training strategy accordingly. Specifically, we apply $L_{gradient}$ exclusively at timesteps $t < \alpha T$, where T denotes the total number of diffusion timesteps and $\alpha \in [0, 1]$ is an adjustment parameter, defaulting to 0.25. Consequently, the overall training objective of the network is formulated as follows:

$$L_{total} = \begin{cases} L_{denoising} + L_{content} + L_{gradient}, & \text{if } t < \alpha T \\ L_{denoising} + L_{content}, & \text{otherwise} \end{cases}. \tag{6}$$

A more detailed training strategy is discussed in the experimental section(Sec 5.1).

4.2. Foreground Layout Control

Through training the F_c in the proposed paradigm, it can achieve precise foreground injection into the generation process. However, due to the lack of semantic understanding of the foreground introduced in the frozen diffusion model, challenges such as distortion and redundant generation of the foreground elements may still arise. To address this challenge, we propose a foreground layout control strategy applied during the iterative denoising steps of model inference to avoid such errors.

As discussed in previous works [5, 12, 39], in cross-attention based text-to-image diffusion models, the cross-attention map between the word tokens and latent spatial patches is calculated over the visual features (Q) and word embeddings (K), as referenced in Eq.(1). In timestep t, for a text sequence of length N, the cross-attention map can be denoted as $A_t \in \mathbb{R}^{p\times p\times N}$, with p indicating the resolution of the map. Specifically, $A_t[i, j, n]$ denotes the probability that the n-th token conveys information to the (i, j)-th spatial patch of the intermediate feature map, reflecting the extent of influence each word has on the image patch. Therefore, observing cross-attention maps in incorrectly generated examples is crucial to addressing the issue of generation errors with foreground elements.

As illustrated in Fig. 4, during the default sampling process, it is observed that although the foreground information has been injected into the diffusion model, the attention

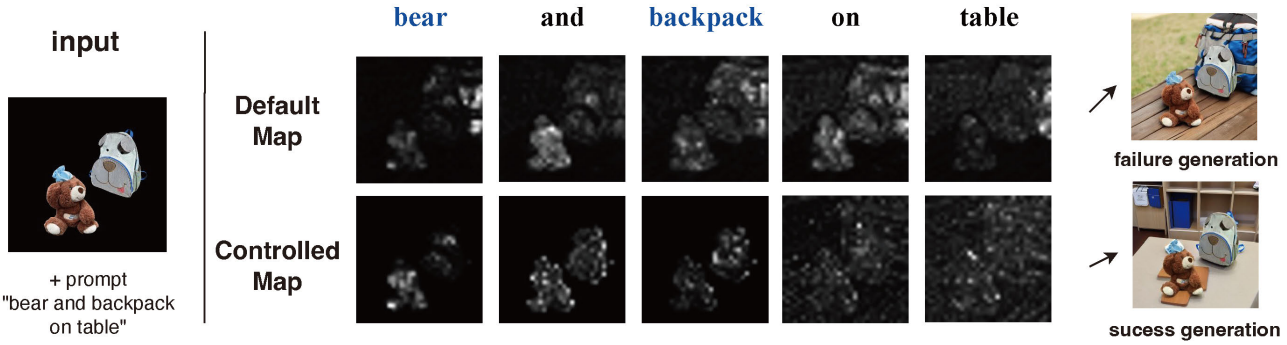

Figure 4. In the failure cases, the semantic information of "backpack" extends beyond the designated foreground region, resulting in the distortion of foreground elements and their redundant generation within the synthesized image. Our proposed foreground layout control strategy capably avoids such failures by effectively confining the foreground semantics to the appropriate region.

maps of words associated with foreground elements still probably extend beyond their intended region. To deal with this issue, we apply the foreground mask M to these expansive attention maps. Specifically, for words designated by the user as related to foreground elements, we identify their positions in the input prompt as $\mathcal{K} = \{k_1, k_2, \dots k_n\}$. The cross-attention editing strategy can then be written as:

$$A^*_{t_k} = A_t[:, :, k] \odot (M \downarrow p), k \in K \tag{7}$$

where $\downarrow$ indicates the down-sampling operation on the mask M to the resolution $p \in \{64, 32, 16\}$, matching the resolution of A_t at different network levels. This control strategy ensures that the semantic information of the foreground elements remains confined within the masked area, as shown in Fig. 4, thus facilitating controllable preservation of the foreground layout. Moreover, since this strategy operates during the sampling steps, it is independent of the model's training phase. This allows for its direct application to models already trained, providing an adaptable control mechanism for the practical application of our method.

4.3. Foreground Blending

To reintegrate the original foreground details into the generated scenes, current alternatives merely paste the foreground over the mask M, which may lead to issues with border harmony and disregard the scene's overall color tone. Addressing these limitations, a viable solution is the application of image blending techniques like Poisson Blending [25], which can smoothen intensity transitions in the pasting area, thus reducing artifacts. Poisson Blending achieves harmonized blending by formulating a guidance vector field as:

$$\min_f \iint_\Omega |\nabla f - \nabla g|^2 \text{ with } f|_{\partial\Omega} = f^*|_{\partial\Omega}\,. \tag{8}$$

where ∇ represents a gradient operator. The main purpose is to calculate the final blending image f under the condition that the foreground boundary $\partial\Omega$ of generated scene f^* remains unchanged while ensuring the gradient of f in the foreground region Ω closely approximates the gradient of foreground g. Utilizing an eroded mask M_{eroded}, we reintegrate the intricate inner details of the foreground into the generated image. Finally, it outputs the synthesized image with overall visual harmony and precise details of the foreground elements.

5. Experiment

5.1. Implementations Details

Training Dataset. We construct a training dataset comprising {foreground image, scene image, text prompt} data pairs for training the foreground injection module. The training images are sourced from FFHQ [16, 39], COCO [4, 22], and OpenImage [18] datasets. We utilize the segmentation annotations in these datasets to extract foreground objects from images, which were then recombined to form composited foregrounds. Additionally, we employ BLIP-2 [20] to generate scene descriptions for data lacking textual annotations. We collect 240K training pairs from various categories to train our network.

Data Augmentation. Throughout our training process, we apply random data augmentations to the foreground elements to enhance the ability of the Foreground Injection Module F_c to generate visually harmonious scenes under various lighting, color, and clarity conditions. Using the Albumentations library [3], we introduce random color(RGBshift), contrast, and gamma adjustments, creating diverse foreground augmentations.

Framework Training. We develop our framework using Stable Diffusion v1.5 [30] as the base model, with all model parameters frozen. Particularly, we start training our network from the weights of community ControlNet-Inpaint [45] checkpoint. Although it cannot be directly applied to the proposed scene generation task (it fails to inpaint large blank areas like the entire background), its scene completion capability trained on random masks can accel-

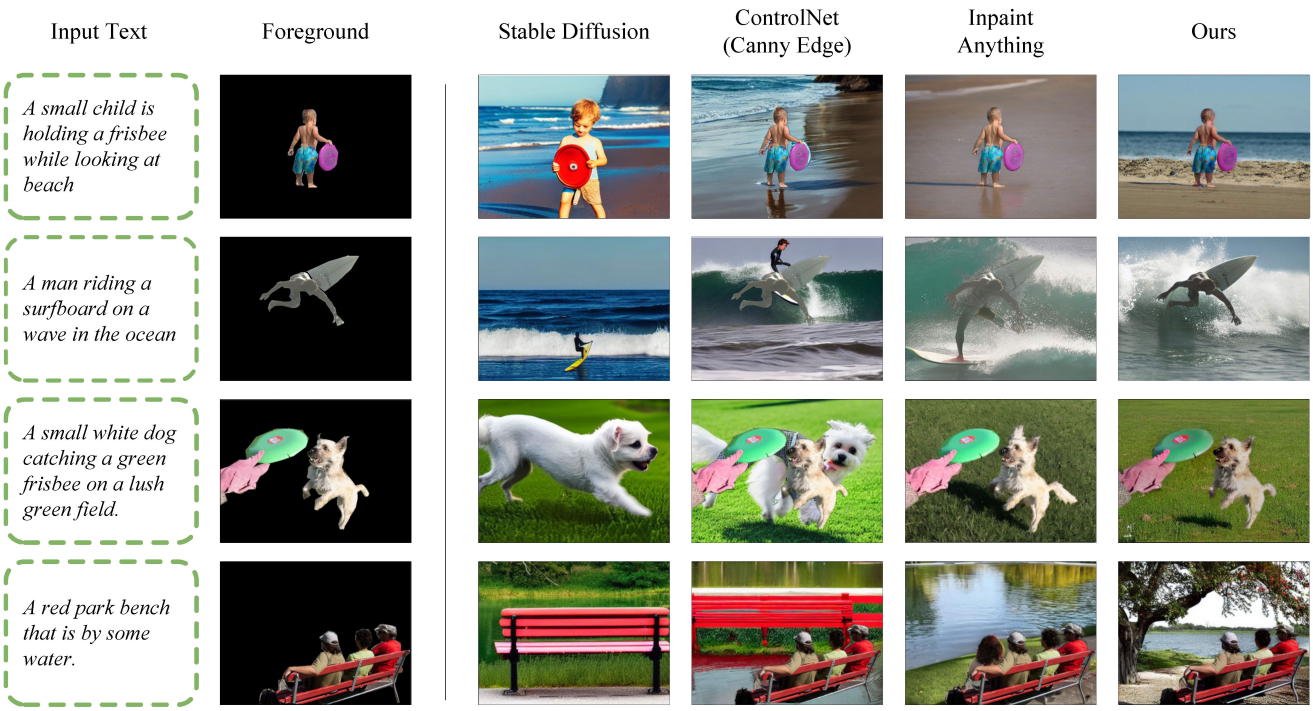

Figure 5. Qualitative comparisons on the proposed benchmark, which shows the generation qualities conditioned with the foreground from the real scene. To demonstrate the scene generation capabilities of the generative model without introducing a foreground, we also display and compare the results generated by default versions of Stable Diffusion.

erate the training efficiency of our proposed framework. Training was conducted on two Nvidia A800 GPUs, with a batch size of 16, at an image resolution of 512×512 and a learning rate of 1×10^{-5}. Employing the AdamW optimizer, the training was completed over 40 hours, finishing 100K training steps.

5.2. Evaluation Settings

To comprehensively evaluate the efficacy of the AnyScene on the task of *"customizing diverse scenes for composited foreground"*, we devise an evaluation benchmark. We use the same method as for constructing the training data to form 3,108 pairs of composited foreground as well as target scene descriptions on COCO [22], which is specifically curated to evaluate the ability of models to synthesize target scenes from given foregrounds and scene descriptions.

We employ three classic metrics to evaluate the quantitative quality of the synthesized images, including FID [13] (Fréchet Inception Distance) for visual quality, CLIP-Score [26] for text-visual consistency, and LAION Aesthetic [36] for the aesthetic quality of generated images.

5.3. Comparison Method

To comprehensively evaluate the generation quality and applications with our approach, comparative analyses were conducted against three distinct types of generative models:

Alternatives for the Proposed Task. Specifically, it includes two methods. 1) Canny Edge: the Canny Edge version of ControlNet [45]. We extract the canny edges of the input foreground and use them to guide the overall scene generation. 2) Inpainting Anything [30, 42]: a method that employs an inpainting version of Stable Diffusion to replace the background with a given foreground. We treat these methods as the primary comparative method to compare the generative capabilities through quantitative evaluation and visualization of results.

Reference-based Subject-Driven Method. Methods in this category use a single image as a visual reference to synthesize new scenes with text prompts. We employ the zero-shot setting of BLIP-Diffusion [19] for comparison to evaluate how our method compares to this technique in the proposed task.

Finetuning-based Subject-Driven Method. This approach entails finetuning the model with multiple images of the same subject to learn visual concepts and generate new scenes. We employed DreamBooth [32] and CustomDiffusion [17], finetuning each with 4 to 10 images for every foreground element, composing them to generate customized images.

Since the subject-driven methods can customize scenes for specific objects, we compare the image quality generated by these methods with our approach in the image qual-

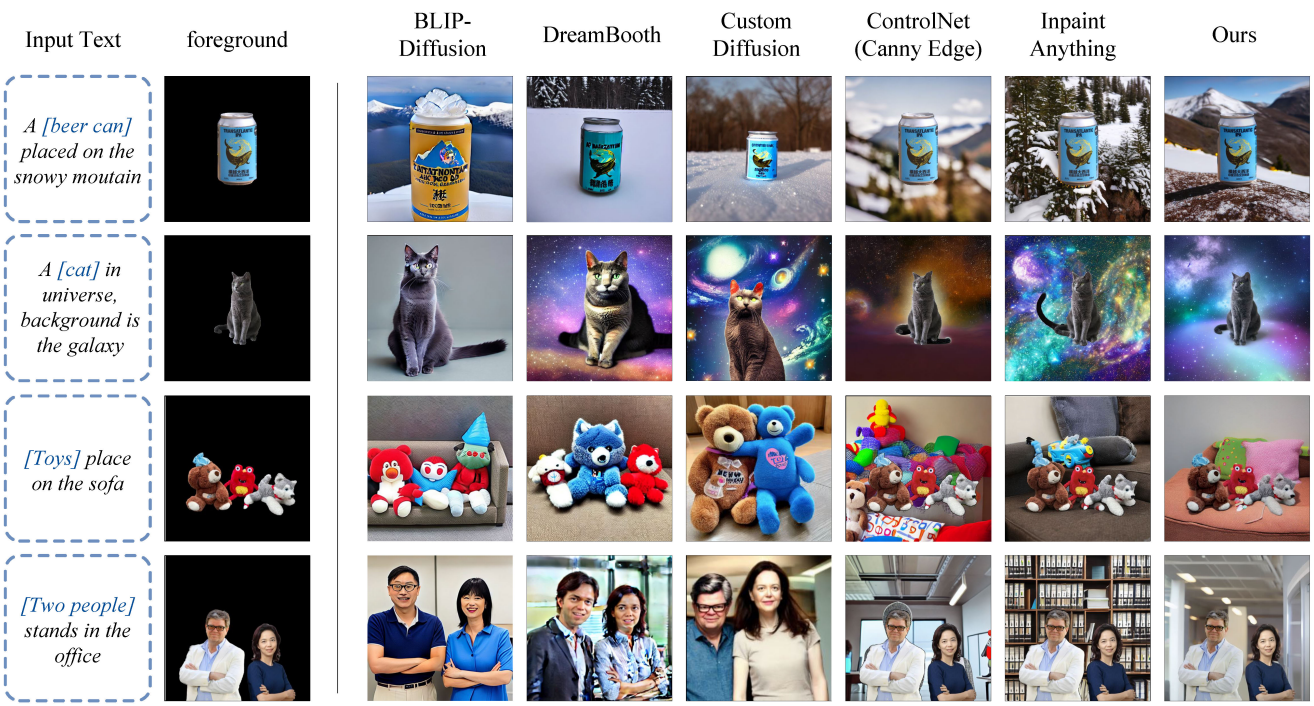

Figure 6. Qualitative comparisons by simulating the user's combination of specific objects to form the foreground, where the words marked blue in the input text are user-specified foreground-related words, and we use this information to control the foreground layout using the proposed foreground control strategy.

Table 1. The quantitative evaluation results with the proposed benchmark.

	FID (↓)	CLIP-S (↑)	Aesthetic (↑)
Stable Diffusion [30]	26.82	14.19	6.02
Canny Edge [45]	24.07	14.64	5.56
Inpaint Anything [42]	16.52	14.72	5.83
AnyScene (ours)	**16.09**	**15.18**	**5.94**

ity analysis with specific foreground objects and discuss their differences in applications.

5.4. Comparison Experiments

5.4.1 Quantitative Comparisons

The quantitative comparative results are detailed in Table.1. As a baseline, we include the performance of default Stable Diffusion, which generates images solely from text prompts. The results in the table indicate that AnyScene surpasses the alternative methods across both metrics. Regarding the aesthetic metric of the generated images, our results more closely align with the aesthetic evaluation result of the default Stable Diffusion, demonstrating our method's effective utilization of the pre-trained model's capabilities to synthesize high-quality images based on the given foreground.

5.4.2 Qualitative Comparisons

Fig. 5 showcases the generated images in our proposed benchmark, which evaluates the model's ability to customize images for real scenes. Meanwhile, in Fig. 6, we simulate user actions of compositing elements to obtain foregrounds and incorporate the foreground control strategy for generation.

Ours vs. alternative methods. As shown in the visualization results. ControlNet based on Canny Edge can only correctly generate scenes when the foreground elements are clear and only a single element. Inpainting Anything can produce images with roughly reasonable semantics but faces issues like foreground distortion, especially when the foreground elements are complex, leading to a significant decline in visual harmony. In contrast, our proposed method exhibits a more precise understanding of foreground semantics, thereby achieving superior visual effects and synthesizing scenes that are both accurate and of higher image quality.

Ours vs. subject-driven methods. As depicted in Fig. 6, we compare our method with subject-driven approaches to highlight differences in application focus. BLIP-Diffusion[19], relying only on a single reference image, struggles to preserve the details of the foreground in the generation results. Meanwhile, DreamBooth [32] and Cus-

Figure 7. By applying the proposed losses, the visual quality of the synthesized image can be enhanced in object texture fidelity and overall scene harmony.

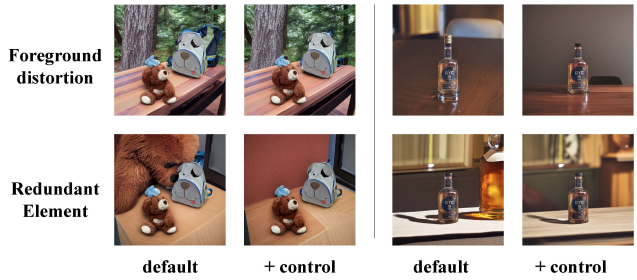

Figure 8. the visualization of applying the foreground layout control strategy to eliminate two types of errors.

tomDiffusion [17] achieve favorable visual quality by generating individual elements in customized scenes. However, they struggle to cohesively combine multiple visual concepts, resulting in loss of the foreground elements and significant detail distortion. Although our proposed framework does not provide extensive variability in foreground elements, its robust layout control and superior scene synthesis quality still make it well-suited for a wide range of practical applications.

5.5. Module Analysis

In this section, we discuss the core components of the proposed AnyScene to validate their respective effectiveness.

Training Losses. In this study, the Foreground Injection Module is trained to seamlessly integrate foreground information into the pre-trained diffusion model. We incorporate two additional losses to aid the network's training alongside the standard denoising loss. As depicted in Fig. 7, visual comparisons highlight the distinct outcomes when employing different loss functions. The baseline solely using the denoising loss shows slight edge and color distortions in the generated images. Incorporating content loss improves the network's capability to accurately capture texture details and preserve the true colors of the foreground textures. Besides, employing gradient loss contributes to generating more realistic foreground edges while curbing shape distortions in foreground elements by constraining gradient changes at their edges. Combining these two losses' advantages, AnyScene can generate structurally accurate and visually harmonious customized scene images.

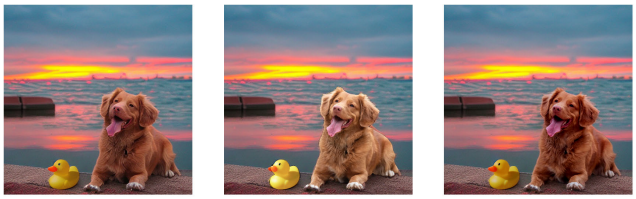

Figure 9. the visualization of the effectiveness of employing foreground blending.

Foreground Control Strategy. We conduct visualization experiments to assess the efficacy of the foreground control strategy, particularly its impact on the surrounding scenes after modifying the attention map. Fig. 8 shows the scene changes before and after strategy application with consistent input noise. Results indicate that our strategy eliminates severe foreground distortions and redundant elements without degrading overall scene quality. Instead, it enhances contextual coherence in the eliminated areas, improving the visual integrity of the generated images.

Effect of Foreground Blending. Fig. 9 demonstrates the impact of applying foreground blending. In this case, AnyScene receives a foreground image and the textual prompt "with sunset" to generate a preliminary corresponding scene. The foreground blending module is then utilized to integrate the details from foreground elements. As the figure indicates, directly using "Copy & Paste" results in noticeable color discrepancies and edge disharmony. In contrast, employing the image blending technique can consider the generated scene's color and lighting conditions and achieve a visually harmonious final synthesized image.

6. Conclusion

This paper presents AnyScene, a tailored image synthesis framework for customizing scenes to specific composited foregrounds. Firstly, we propose the foreground injection module, which effectively guides the pre-trained diffusion model to generate harmonious scenes with the given foreground. Then, we develop the foreground layout control strategy to ensure robust scene generation and the foreground blending mechanism to preserve foreground details. Our extensive experiments validate the effectiveness of AnyScene with the proposed task, highlighting its versatility and potential for a wide range of applications.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under grants U21B2024, 62202329, and 62272337.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[3] Alexander V. Buslaev, Alex Parinov, Eugene Khvedchenya, Vladimir I. Iglovikov, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *ArXiv*, abs/1809.06839, 2018. 5

[4] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2016. 5

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2, 4

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[7] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 2

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[10] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 4

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017. 6

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[17] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2, 6, 8

[18] Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128:1956 – 1981, 2018. 5

[19] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 1, 2, 6, 7

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5

[21] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5, 6

[23] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 1, 2

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[25] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582. 2023. 5

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.

Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[29] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 4, 5, 6, 7

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[32] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 2, 6, 7

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[35] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 6

[37] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[39] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2, 4, 5

[40] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[41] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2

[42] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 1, 2, 6, 7

[43] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[44] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 2

[45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 3, 5, 6, 7