MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing

Mingdeng Cao[1,2*] Xintao Wang[2✉] Zhongang Qi[2] Ying Shan[2] Xiaohu Qie[2] Yinqiang Zheng[1✉]
[1]The University of Tokyo [2]ARC Lab, Tencent PCG

https://github.com/TencentARC/MasaCtrl

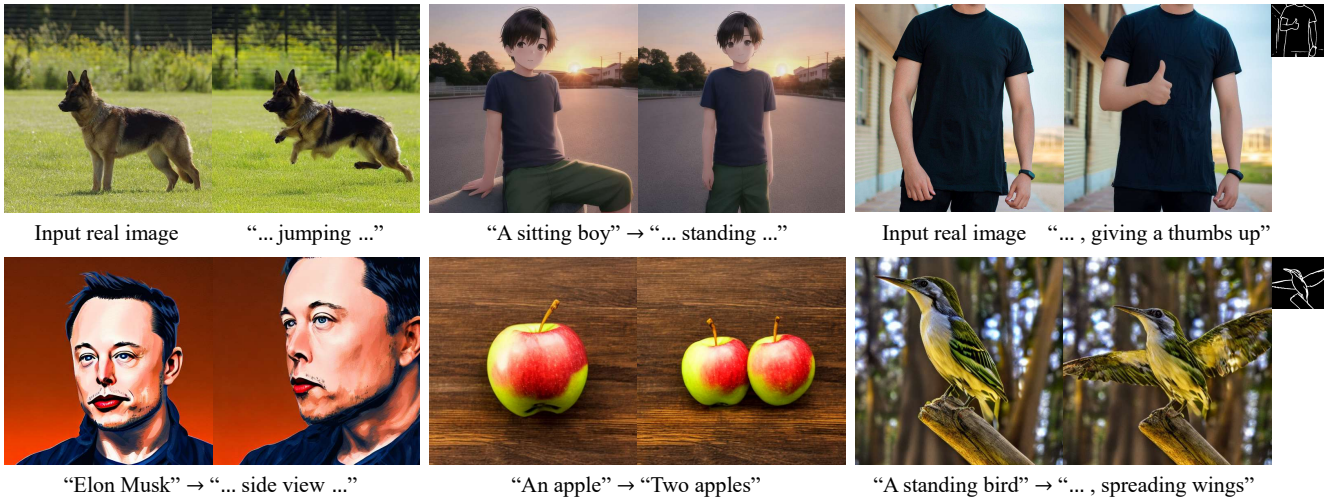


Figure 1: Our method **MasaCtrl** can perform text-based non-rigid image synthesis and real image editing without fine-tuning. Meanwhile, our method can be easily integrated into controllable diffusion models, like T2I-Adapter [17] or ControlNet [45], to perform more consistent and faithful synthesis and editing (last column).

Abstract

Despite the success in large-scale text-to-image generation and text-conditioned image editing, existing methods still struggle to produce consistent generation and editing results. For example, generation approaches usually fail to synthesize multiple images of the same objects/characters but with different views or poses. Meanwhile, existing editing methods either fail to achieve effective complex non-rigid editing while maintaining the overall textures and identity, or require time-consuming fine-tuning to capture the image-specific appearance. In this paper, we develop ***MasaCtrl****, a tuning-free method to achieve consistent image generation and complex non-rigid image editing simultaneously. Specifically, MasaCtrl converts existing self-attention in diffusion models into mutual self-attention, so that it can query correlated local contents and textures from source images for consistency. To further alleviate the query confusion between foreground and background, we propose a mask-guided mutual self-attention strategy, where the mask can be easily extracted from the cross-attention maps. Extensive experiments show that the proposed MasaCtrl can produce impressive results in both consistent image generation and complex non-rigid real image editing.*

1. Introduction

Recent advances in text-to-image (T2I) generation [25, 19, 41, 24, 27] have achieved great success. Those large-scale T2I models, such as Stable Diffusion [27], can generate diverse and high-quality images conforming to given text prompts. When leveraging the T2I models, we can also perform promising text-conditioned image editing [19, 10, 35, 21]. However, there is still a large gap between our needs and existing methods in terms of *consistent* generation and editing.

*Work done during an internship at ARC Lab, Tencent PCG.

For the text-to-image generation, we usually want to generate several images of the same objects/characters but with different views or complex non-rigid variances (*e.g.*, the changes of posture). Such capabilities are urgently needed for creating comic books and generating short videos using existing powerful T2I image models. However, this requirement is highly challenging. Even if we fix the input random noise and use very similar prompts (*e.g.*, 'a sitting cat' *vs*. 'a laying cat' shown in Fig. 2), the generated images vary in both structures and identity.

For text-conditioned image editing, existing methods [10, 35, 21] achieve impressive editing effects in image translation, stylization, and appearance replacement while keeping the input structure and scene layout unchanged. However, those methods usually fail to change poses or views while maintaining the overall textures and identity, leading to inconsistent editing results. The latter editing way is a more complicated *non-rigid editing* for practical use. Imagic [12] is then proposed to address this challenge. It allows complex non-rigid edits while preserving its original characteristics. It can make a standing dog sit down, cause a bird to spread its wings, *etc*. Nevertheless, it requires fine-tuning the entire T2I diffusion model and optimizing the textual embedding to capture the image-specific appearance for each edit, which is time-consuming and impractical for real-world applications.

In this paper, we aim to develop a *tuning-free* method to address the above challenges, enabling a more consistent generation of multiple images and complex non-rigid editing without fine-tuning. The core challenge is how to keep consistent. Unlike previous works [10, 21, 3] that usually operate on cross-attention in T2I models, we propose to convert existing *self-attention* to mutual self-attention, so that it can query correlated local structures and textures from a source image for consistency. Specifically, we first generate an image from a random (or inverted a real image) noise, resulting in the denoising process (DP1) for the source image synthesis. In the new denoising process (DP2) of generating a new image or editing an existing one, we can use the *Query* features in DP2 self-attentions to query the corresponding *Key* and *Value* features in DP1 self-attentions. In other words, we transform the existing self-attention into 'cross-attention', where the crossing operation happens in the self-attentions of two related denoising processes, rather than between the U-Net features and text embeddings. We call this 'crossing self-attention as *mutual self-attention*. However, directly applying this strategy can only generate images almost identical to the source image and cannot comply with the target text prompt (as analyzed in Fig. 9). Thus we further control the denoising timestep and the layer position in U-Net for performing mutual self-attention to achieve consistent synthesis and editing. More analyses are in Sec. 4.1 and Sec. 5.5. In this way, we can use contents in the source image as the generation material to better maintain the texture and identity. Meanwhile, its structure, pose, and non-rigid variances can be controlled by target text prompt, and guided by recent controllable T2I-Adapters [17] or ControlNet [45]. Fig. 1 shows some synthesis and editing examples.

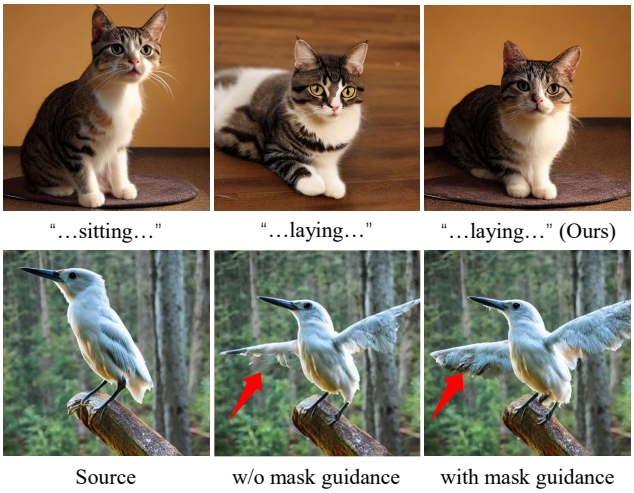


Figure 2: First row: images synthesized from fixed random seed (middle, changed identity) and our method (right, maintained identity). Second row: image synthesized without mask guidance (middle) and with mask guidance (right).

The proposed mutual self-attention control can work well for images with disentangled foreground and background, but may fail in the synthesis and editing process where the foreground and background have similar patterns and colors. In these cases, mutual self-attention tends to confuse the foreground and background, leading to a messy result (2nd row in Fig. 2). To address this problem, we further propose a mask-guided mutual self-attention. Specifically, we first utilize the cross attention in T2I diffusion models to extract a mask associated with the main object in the image. This mask can successfully separate foreground and background, and restrict the target foreground/background features to query only foreground/background features of the source image, respectively. Such an operation can effectively alleviate the query confusion between the foreground and background.

Our contributions can be summarized as follows. **1)** We propose a tuning-free method, namely, MasaCtrl, to simultaneously achieve consistent image synthesis and complex non-rigid image editing. **2)** An effective mutual self-attention mechanism with delicate designs is proposed to change pose, view, structures, and non-rigid variances while maintaining the characteristics, texture, and identity. **3)** To alleviate the query confusion between foreground and background, we further propose a masked-guided mutual self-attention, where the mask can be easily computed from the

cross-attentions. **4)** Experimental results have shown the effectiveness of our proposed MasaCtrl in both consistent image generation and complex non-rigid real image editing.

2. Related Work

2.1. Text-to-Image Generation

Early image generation methods conditioned on text description mainly based on GANs [26, 43, 44, 39, 14, 2, 46, 40, 42, 34], due to their powerful capability of high-fidelity image synthesis. These models try to align the text descriptions and synthesized image contents via multi-modal vision-language learning and have achieved cheerful synthesis results on domain-specific datasets. Text-to-image generation with auto-regressive and diffusion models has obtained impressive diversity results. DALL·E [25], CogView [6] and Parti [41], large-scale text-to-image models trained with a large amount of data, enable generating images from open-domain text descriptions. However, the auto-regressive generation nature leads to the slow generation process defect. Most recently, diffusion models [32, 11, 20, 5] have shown superior generative power and achieved state-of-the-art synthesis results in terms of image quality and diversity than previous GAN-based and auto-regressive image generation models. By conditioning the text prompt into the diffusion model, various text-to-image diffusion models GLIDE [19], VQ-Diffusion [8], LDM [27], DALL·E 2 [24], and Imagen [30] have been developed. They can synthesize high-quality images that highly comply with the given text description.

2.2. Text-guided Image Editing.

Text-guided image editing is a challenging task that aims to manipulate images according to natural language descriptions. Previous methods based on generative adversarial networks (GANs) [18, 15, 38, 22] have achieved some success on domain-specific datasets (*e.g.*, face datasets), but they have limited applicability and generality. A recent approach based on auto-regressive models, VQGAN-CLIP [4], combines VQGAN [7] and CLIP [23] to produce high-quality images and precise edits with diverse and controllable results. However, this approach suffers from slow generation speed and high computational cost.

Different from previous methods based on GANs or auto-regressive models, diffusion models offer a fast and efficient way to synthesize and edit images conditioned on text prompts. However, existing diffusion-based methods have some limitations regarding local and global editing. For example, the works [19, 1] require extra masks to edit local regions of the image; [13] can edit global aspects of the image by changing the text prompt directly, but cannot modify local details; [10, 35] use cross-attention or spatial features to edit both global and local aspects of the image by changing the text prompt directly. Still, they tend to preserve the original layout of the source image and fail to handle non-rigid transformations (*e.g.*, changing object pose). In contrast, we propose a novel approach that leverages the self-attention mechanism to achieve consistent and complex non-rigid image synthesis and editing. Our approach can modify various object attributes (*e.g.*, pose, shape) by changing the text prompt accordingly. The most related work to ours is Imagic [12], which also enables various prompt-based non-rigid image editing. However, unlike our approach can edit images on the fly, Imagic requires careful optimization of the textual embedding and fine-tuning of the model, which is time-consuming and unfriendly for ordinary users.

3. Preliminaries

3.1. Latent Diffusion Models

Diffusion models [11, 31, 20] are generative models that can synthesize desired data samples from Gaussian noise via iterative denoising. Our method is based on the recent state-of-the-art text-conditioned model Stable Diffusion (SD) [27], which performs the diffusion-denoising process in the latent rather than image space. Specifically, a pretrained image autoencoder first encodes the image x into latent representations z. Then the denoising network ϵ_θ (a time-conditional U-Net [28]) is trained in this latent space. After being trained, we can sample a random noise $z_T \sim \mathcal{N}(0, 1)$ and perform the latent denoising process. The denoised latent representation z_0 can be decoded into an image using the pretrained autoencoder.

3.2. Attention Mechanism in Stable Diffusion

The denoising U-Net ϵ_θ in the SD model, consists of a series of basic blocks, and each basic block contains a residual block [9], a self-attention module, and a cross-attention [36] module. At denoising step t, the features from the previous $(l-1)$-th basic block first pass through the residual block to generate intermediate features f_t^l; then they are reorganized by the self-attention layer, and receive textual information from the text prompt P by the following cross-attention layer. The attention mechanism can be formulated as follows:

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

where Q is the query features projected from the spatial features, and K, V are the key and value features projected from the spatial features (in self-attention layers) or the textual embedding (in cross-attention layers) with corresponding projection matrices. $A = \text{Softmax}(\frac{QK^T}{\sqrt{d}})$ is the attention map used to aggregate the value V.

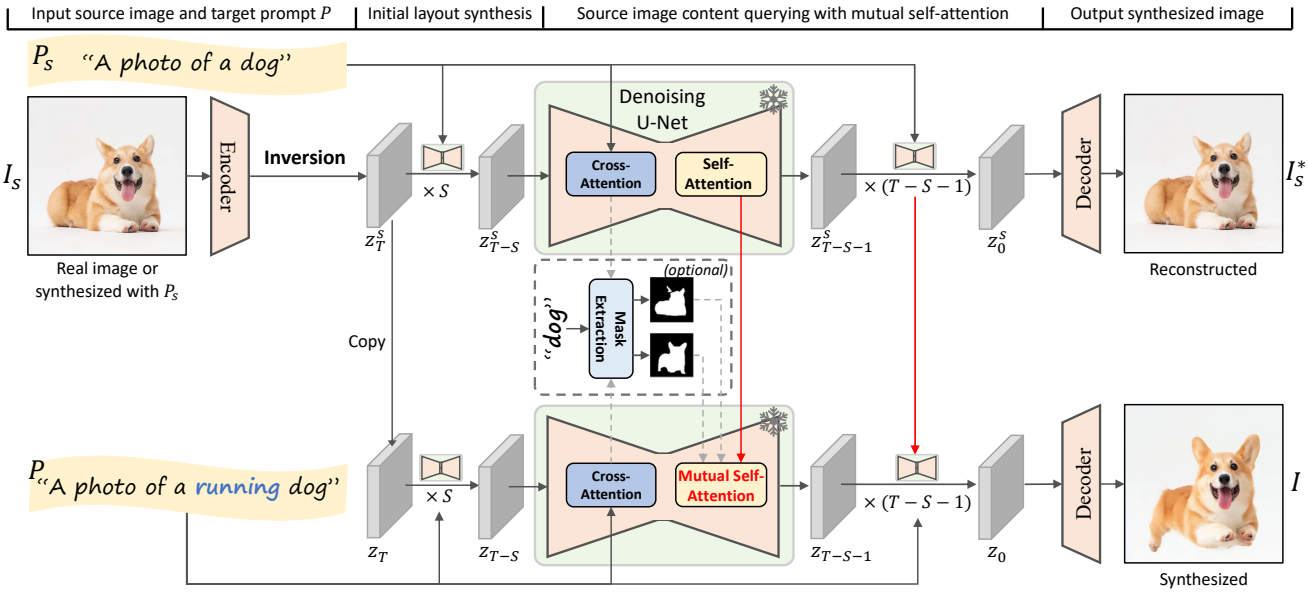


Figure 3: Pipeline of the proposed MasaCtrl. Our method tries to perform complex non-rigid image editing and synthesize content-consistent images. The source image is either real or synthesized with source text prompt P_s. During the denoising process for image synthesis, we convert the self-attention into mutual self-attention to query image contents from source image I_s, so that we can synthesize content-consist images under the modified target prompt P.

These attention layers in the SD model contain much information for the overall structure/layout and content formation of the synthesized image [10, 35]. The internal cross-attention maps are high-dimensional tensors that bind the spatial pixels and textual embedding [10], and they are explored for image editing [10] and faithful image synthesis [3]. In addition, the features in the self-attention layer are employed as plug-and-play features to be injected into specified U-Net layers to perform image translation. However, these controls cannot perform non-rigid editing (*e.g.*, pose change) since they maintain the semantic layout and structures. Inspired by the phenomenon that performing self-attention across batches can generate similar image contents, which is also observed in Tune-A-Video [37], we adapt the self-attention mechanism in the T2I model to query contents from source images with delicate designs. Thus we can perform consistent synthesis and non-rigid editing that change the layout and structure of the source image while preserving image contents.

4. Tuning-Free Mutual Self-Attention Control

Given a source image I_s and the corresponding text prompt P_s, our goal is to synthesize the desired image I that complies with the target edited text prompt P (directly modified from P_s). Note that the edited target image I is spatially edited from I_s and should preserve the object contents (*e.g.*, textures and identity) in I_s. For instance, consider a photo (corresponding to P_s) where a dog is sitting, and we want the dog to be in the running pose with the edited text prompt P that adds the 'running' into the source prompt P_s (see Fig. 3).

Our core idea is to combine the semantic layout synthesized with the target prompt P and the contents in the source image I_s to synthesize the desired image I. To achieve so, we propose MasaCtrl, which adapts the self-attention mechanism in the SD model into a crossing one to query semantically similar contents from the source image. Consequently, the target image I can be synthesized by querying the contents from I_s with the modified self-attention mechanism during the denoising process. We can achieve so for the following reasons: **1)** the image layout is formed in the early denoising steps (shown in Fig. 4(a)); **2)** in addition, as shown in Fig. 4(b), the encoded query features in the self-attention layer are semantically corresponded (*e.g.*, the horses are in the same color), thus one can query content information from another.

The overall architecture of the proposed pipeline to perform synthesis and editing is shown in Fig. 3, and the algorithm is summarized in Alg. 1. The input source image I_s is either a real image or a generated one from the SD model with text prompt P_s [1]. During each denoising step t of synthesizing target image I, we assemble the inputs of

[1]When I_s is a real image, we set the text prompt P_s as null and utilize the deterministic DDIM inversion [31] to obtain the initial noise map. When I_s is an image synthesized with prompt P_s, the initial noise map is the same as the one used to synthesize I_s.

Algorithm 1 MasaCtrl: Tuning-Free Mutual Self-Attention Control

Input: A source prompt P_s, a modified prompt P, the source and target initial latent noise maps z_T^s and z_T.

Output: Latent map z_0^s, edited latent map z_0 corresponding to P_s and P.

for $t = T, T-1, ..., 1$ **do**
  $\epsilon_s, \{Q_s, K_s, V_s\} \leftarrow \epsilon_\theta(z_t^s, P_s, t)$;
  $z_{t-1}^s \leftarrow \text{Sample}(z_t^s, \epsilon_s)$;
  $\{Q, K, V\} \leftarrow \epsilon_\theta(z_t, P, t)$;
  $\{Q^*, K^*, V^*\} \leftarrow \text{EDIT}(\{Q, K, V\}, \{Q_s, K_s, V_s\})$;
  $\epsilon = \epsilon_\theta(z_t, P, t; \{Q^*, K^*, V^*\})$;
  $z_{t-1} \leftarrow \text{Sample}(z_t, \epsilon)$;
end for

Return z_0^s, z_0

the self-attention by **1)** keeping the current Query features Q unchanged, and **2)** obtaining the Key and Value features K_s, V_s from the self-attention layer in the process of synthesizing source image I_s. We dub this strategy mutual self-attention, and more details are in Sec. 4.1.

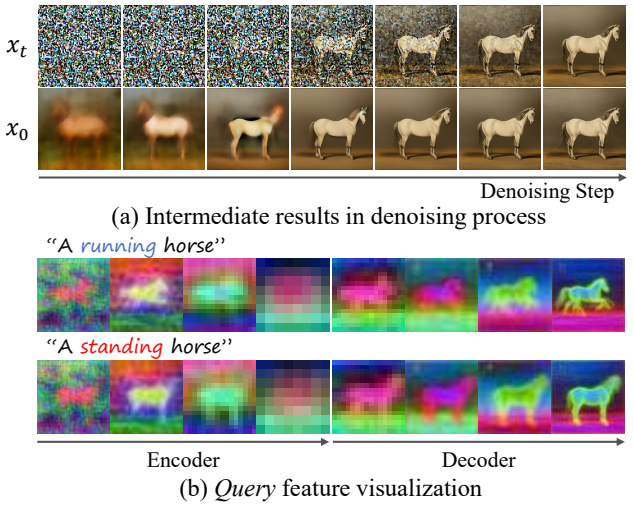
Meanwhile, we also observe the edited image often suffers from the problem of confusion between the foreground objects and background. Thus, we propose a mask-aware mutual self-attention strategy guided by the masks obtained from the cross-attention mechanism. The object mask is automatically generated from the cross-attention maps of the text token associated with the foreground object. Please refer to Sec. 4.2 for more details.

In addition, since the edited prompt P may not yield desired spatial layouts due to the inner limitations of the SD model, MasaCtrl can be easily integrated into existing controllable image synthesis method (*e.g.*, T2I-Adapter [17] and ControlNet [45]) for more faithful non-rigid image editing. Please refer to Sec. 4.3 for more details.

4.1. Mutual Self-Attention

As shown in the left part of Fig. 5(a), at denoising step t and layer l, the query features are defined as the projected query features Q^l in the self-attention module, and the content features are the key features K_s^l and value features V_s^l from the corresponding self-attention layer in the process of reconstructing the source image I_s. After that, we perform attention according to Eq. 1 to aggregate the contents from the source image.

However, intuitively performing such attention on all layers among all denoising steps will result in a failed image I that is nearly the same as the reconstructed image I_s. We argue the reason is that performing self-attention control in the early steps can disrupt the layout formation of the target image. In the premature denoising steps, the target image layout has not yet been formed (shown in Fig. 4(a)).



(a) Intermediate results in denoising process

(b) *Query* feature visualization

Figure 4: (a) The intermediate results during the iterative denoising process, and (b) visualization of the projected *Query* features Q in the self-attention layers of the U-Net at the 15th sampling step.

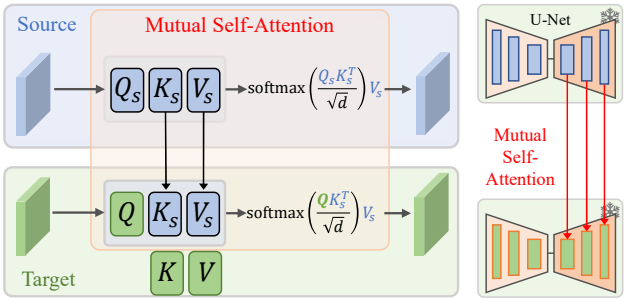
We further observe the *Query* features in the shallow layers of U-Net (*e.g.*, encoder part) cannot obtain clear layout and structure corresponding to the modified prompt (shown in Fig. 4(b)). Thus we cannot obtain the image with the desired spatial layout.

Therefore, we propose to control the mutual self-attention *only in the decoder part of the U-Net after several denoising steps*, due to the formed clear target image layout and semantically similar features (see Fig. 4). We can change the original layout into the target one with edited prompt P and keep the main objects unchanged with proper starting denoising step S and layer L for synthesis and editing. Thus the EDIT function in Alg. 1 can be formulated as follows:

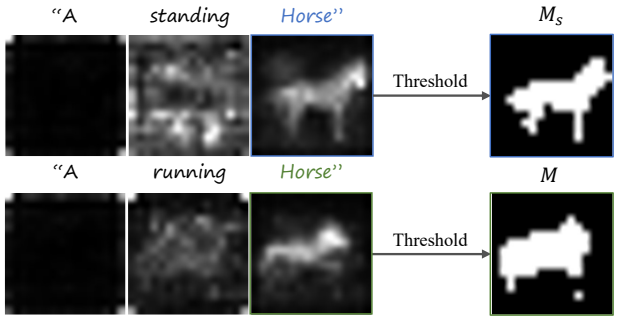
$$\text{EDIT} := \begin{cases} \{Q, K_s, V_s\}, & \text{if} \quad T - t > S \quad \text{and} \quad l > L, \\ \{Q, K, V\}, & \text{otherwise}, \end{cases} \tag{2}$$

where S and L are the timestep and layer index to start attention control, respectively. T is the total denoising steps.

In the early steps, the composition and shape of the object can be roughly generated, complying with the target prompt P. Then the content information from the source image I_s is queried by the mutual self-attention mechanism to fill the generated layout of target image I. After iterative denoising, we can obtain the synthesized image with similar contents in the source image and structure of I^* that complied with the input prompt. Note that our algorithm does not require fine-tuning or optimization, bringing many conveniences for ordinary users for content creation.



(a) Mutual self-attention control



(b) Mask extraction from cross-attention maps

Figure 5: (a) The mutual self-attention mechanism and control strategy in the decoder part of denoising U-Net to query contents from the source image; and (b) mask extraction strategy from cross-attention maps.

4.2. Mask-Guided Mutual Self-Attention

We also observed the above synthesis/editing would fail when the object and background are too similar to be confused. One feasible way to tackle this problem is to segment the image into the foreground and background parts and query contents only from the corresponding parts. Inspired by previous work [10, 33], the cross-attention maps correlating to the prompt tokens contain most information of the shape and structure. Therefore, we utilize the semantic cross-attention maps to create a mask to distinguish the foreground and background in both source and target images I_s and I.

Specifically, at step t, we first perform forward passes with source prompt P_s and edited prompt P, respectively, to generate intermediate cross-attention maps. Then we average the cross-attention maps across all heads and layers with the spatial resolution 16×16. The resulting cross-attention maps are denoted as $A_t^c \in \mathbb{R}^{16\times16\times N}$, where N is the number of the textual tokens. We then obtain the averaged cross-attention map for the token correlated to the foreground object. We denote M_s and M as masks extracted for the foreground objects in I_s and I, respectively. With these masks, we can restrict the object in I to query contents information only from the object region in I_s:

$$f_o^l = \text{Attention}(Q^l, K_s^l, V_s^l; M_s), \tag{3}$$

$$f_b^l = \text{Attention}(Q^l, K_s^l, V_s^l; 1 - M_s), \tag{4}$$

$$\bar{f}^l = f_o^l * M + f_b^l * (1 - M), \tag{5}$$

where $\bar{f}^l$ is the final attention output. The object and background regions in the target image query the contents from corresponding restricted areas rather than all source image features. Note that this strategy is optional and used when confusion occurs.

4.3. Integration to Controllable Diffusion Models

Our method can be integrated into existing controllable diffusion models (*e.g.*, T2I-Adapter [17] and ControlNet [45]) for more faithful non-rigid image synthesis and editing. These methods enable the original Stable Diffusion model to be more controllable (*e.g.*, pose, sketch, segmentation map) in image synthesis. Thus we can use them to synthesize images with desired poses and shapes. Since our method can query image contents (*e.g.*, textures) from a reference image, we can easily integrate our approach into these models to generate more consistent images.

Specifically, we follow the same process depicted in Alg. 1, and the desired target image synthesis process is changed from the original SD model to using these controllable models instead. In the following experiment section, we demonstrate the effectiveness of such a combination that can synthesize images consistently.

5. Experiments

Setup. We apply the proposed method to the state-of-the-art text-to-image Stable Diffusion [27] model with publicly available checkpoints v1.4. We also validate the proposed method on the pre-trained anime-style model Anything-V4. Meanwhile, we perform editing on both synthetic images and real images. For real image editing, we first invert the image into the initial noise map with DDIM deterministic inversion [31]. Note that we set the starting noise map the same for source prompt P_s and the desired prompt P unless otherwise specified. During sampling, we perform DDIM sampling [31] with 50 denoising steps, and the classifier-free guidance is set to 7.5. The step and layer to start attention control is set to $S = 4, L = 10$ as default. Note that it may be changed for specific checkpoints.

5.1. Comparisons with Previous Works

We mainly compare the proposed tuning-free method to the current prompt-based editing methods with diffusion models, including tuning-free methods SDEdit [16], P2P [10], PnP [35], and Imagic [12]. We use their open-sourced codes to produce the editing results [2].

[2]We adopt the community version of Imagic as the official source code is not open-sourced

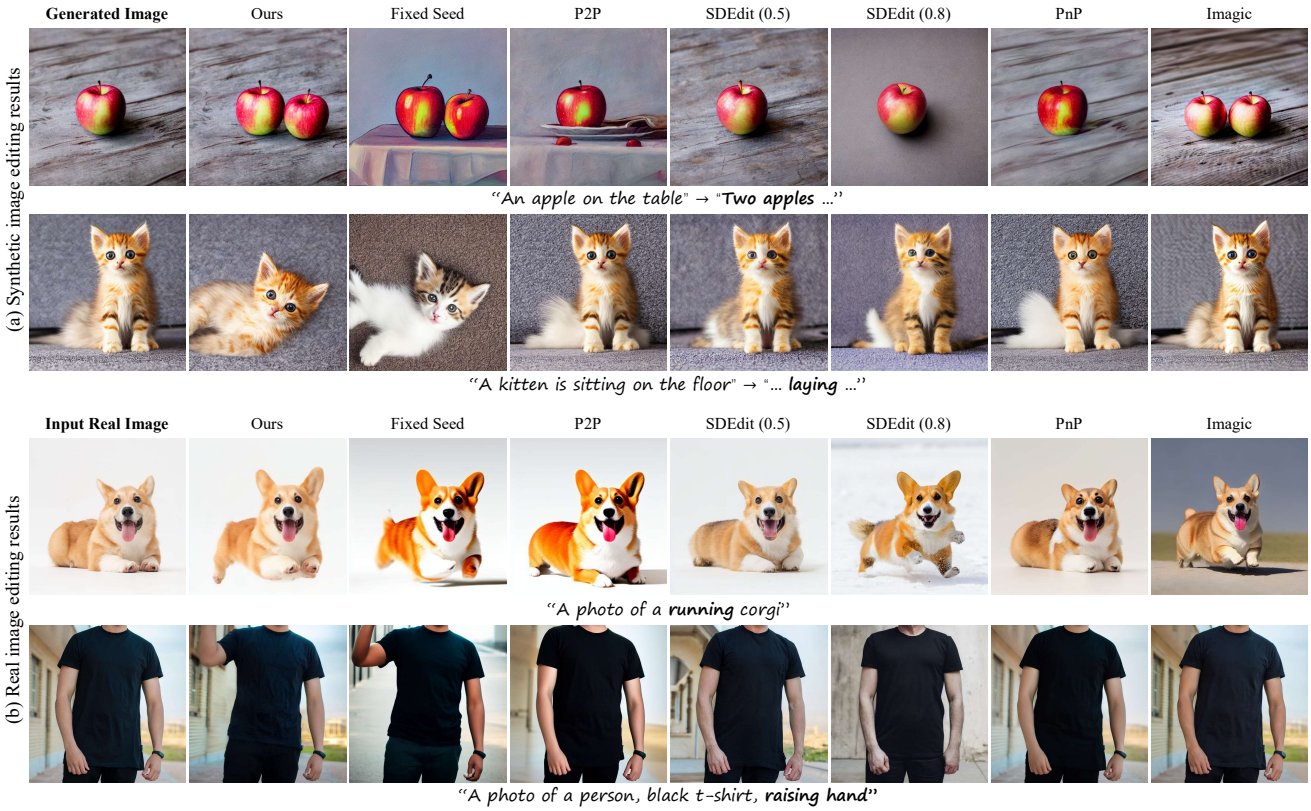


Figure 6: Editing results of different methods on the synthetic images (a) and real images (b). Our method enables consistent synthesis by combining the layout of the target prompt and the contents of source generated image. From left to right: the source generated source image with source prompt, synthesis results with the proposed MasaCtrl method, synthesis results from target prompt with the same random noise of source image, synthesis/editing results with existing methods P2P [10], SDEdit [16], PnP [35], and Imagic [12].

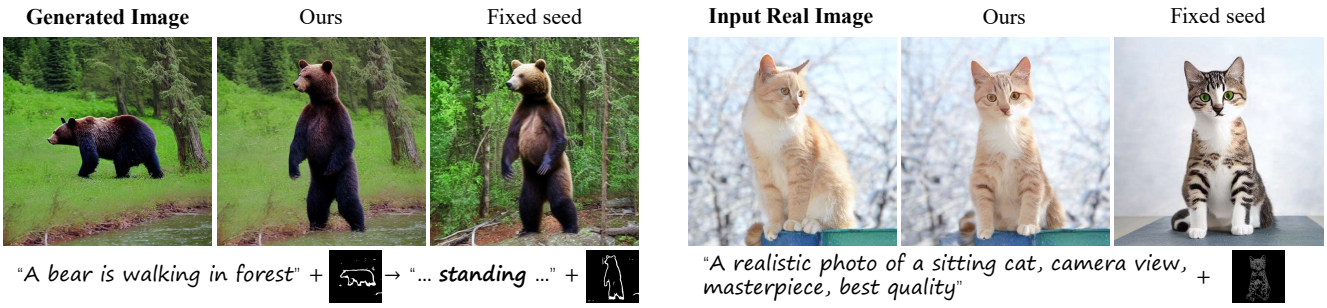


Figure 7: Editing results of synthetic images (left part) and real images (right part) with MasaCtrl integrated into T2I-Adapter [17].

Qualitative comparison. The consistent synthesis results are shown in Fig. 6(a). By directly modifying the text prompt, our method can synthesize content-consistent images. These synthesized images (1) contain contents (foreground objects and background) that are highly similar to those in the generated source images (first column of Fig. 6) and (2) highly comply with the target prompt P (third column of Fig. 6). While existing methods fail to synthesize desired images, either conforming to the target text prompt or inconsistent with the source images. Our method also achieves promising results in editing real images shown in Fig. 6(b). The edited image is consistent with the target prompt and maintains source image content. These results demonstrate the effectiveness of the proposed method.

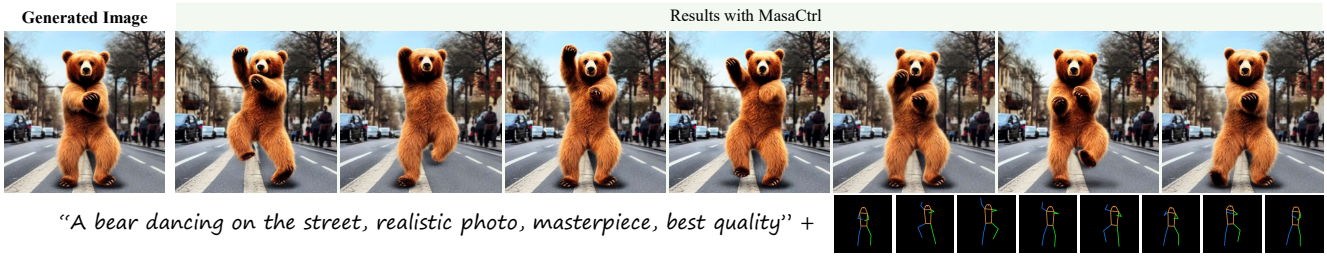


Figure 8: Video synthesis results of proposed MasaCtrl with T2I-Adapter [17] (with key-pose guidance).

Quantitative comparison. For quantitative evaluation, we used 20 different source images (10 are synthetic and 10 are real) and edited them with P2P, PnP, Imagic, and MasaCtrl. We evaluate the *text-alignment* between the target prompt and edited image, and the *image-alignment* between the source and edited images in the CLIP feature space. Tab. 1 shows that MasaCtrl can synthesize images that comply with the target prompt while maintaining content consistency more than other methods. We also conducted a user study and collected 700 answers from professional participants. As shown in the *preference* column of Tab. 1, 73.5% of participants preferred our method. Meanwhile, our method is efficient since no fine-tuning and optimization are required.

We further analyze the reasons attributed to the failure of existing methods (*i.e.*, P2P, SDEdit, and PnP). These methods try to keep the original layout or object shape and pose unchanged by leveraging the layout information encoded in the cross-attention maps (P2P), features (PnP), and original input images (SDEdit). Meanwhile, the contents in the formed image mainly come from the encoded text embedding. As a result, the images synthesized by these methods have similar layouts to the source image but have different contents. In our proposed method MasaCtrl, the structure of the desired image is first determined by the former iteration in the denoising process, and the final image is formed by obtaining the image content from the source image. Thus we can perform various types of non-rigid image editing.

5.2. Results with T2I-Adapter

The initial layout controlled by modifying the text prompt usually fails due to the inherent drawbacks of the Stable Diffusion model. Therefore, we further integrate our method into existing controllable synthesis pipelines to obtain stable synthesis and editing results. The synthetic and real image editing results of MasaCtrl with T2I-Adapter [17] are shown in Fig. 7. We see that T2I-Adapter can generate an image with desired target layout, yet with different contents in the source image. In contrast, our method can effectively combine the layout synthesized by T2I-Adapter with the target prompt and the contents in the source image. Therefore, more faithful and fine-grained synthesis and editing results can be obtained. Note that we may change the attention strategy of starting denoising step S and U-Net layer L to obtain faithful results ($S = 2$, $L = 8$ in our experiment), since the target layout is strongly controlled by extra guidance. Therefore, we can perform attention control in early steps and layers to query faithful contents in the source image, refer to the ablation study Sec. 5.5 for the analysis.

Method	*Text-alignment*	*Image-alignment*	*Preference*	*Runtime*
P2P [10]	0.2691	0.8793	3.0%	15s
PnP [35]	0.2589	0.8902	2.5%	60s
Imagic [12]	0.2688	0.9159	21.0%	14min
Ours	**0.2793**	**0.9286**	**73.5%**	16s

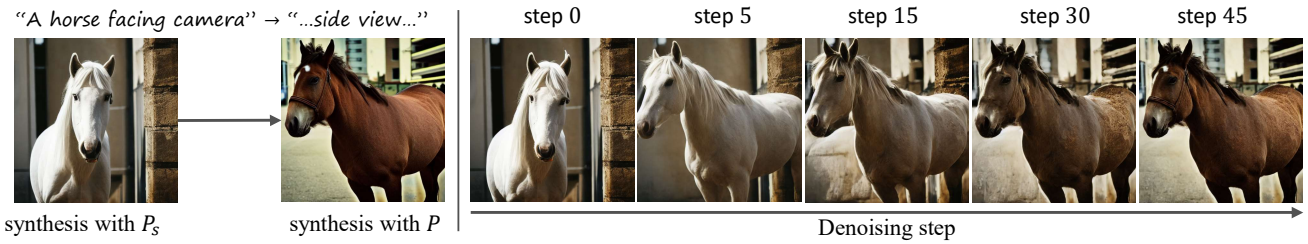
Table 1: Quantitative and user study results.

5.3. Robustness to Other Models

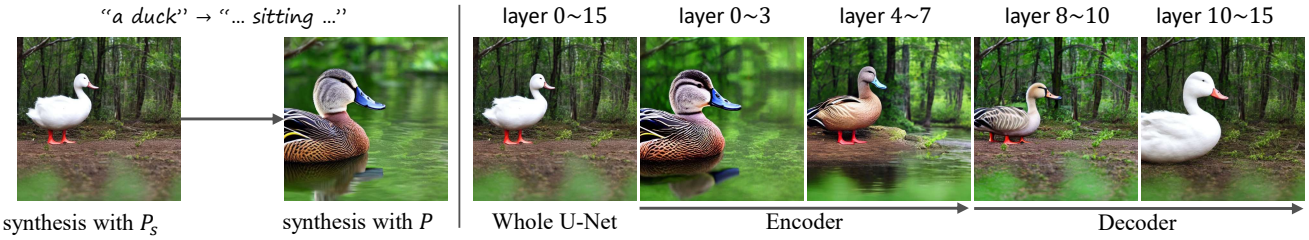
Our method can be applied to other versions of Stable Diffusion models (*e.g.*, v1.5), domain-specific models (*e.g.*, the amine-style model Anything-v4), and customized models (*e.g.*, models fine-tuned with DreamBooth [29]) to achieve consistent image synthesis and editing. Please refer to the supplementary materials for the visual results.

5.4. Extension to Video Synthesis

Although our method is designed for consistent image synthesis and editing, we can easily extend our method for the video synthesis task using T2I-Adapter and ControlNet with a series of dense temporal-coherent guidance (*e.g.*, pose, edge, depth). Specifically, we first generate a source image with the prompt P_s and guidance as the canonical frame (shown in the first column of Fig. 8). The other frames are synthesized separately with MasaCtrl. Then all the synthesized frames can be concatenated into a video since they are content-consistent with the source frame. Fig. 8 shows the video synthesis results with coherent dense guidance. The proposed method successfully synthesizes consistent frames with highly similar content (more video results are available on the project page and supplementary materials). However, our method can only animate the foreground objects (such as the bear in Fig. 8) and hardly bring the background alive. Therefore, video-based approaches still need to be explored since the current MasaCtrl has



(a) Results of mutual self-attention control starting from different denoising steps



(b) Results of mutual self-attention control in different U-Net layers

Figure 9: Results of mutual self-attention control in different denoising steps (a) and different U-Net layers (b). We see that only performing mutual self-attention control after several denoising steps (*e.g.*, step 5), and in the decoder part (*e.g.*, layer $10 \sim 15$) can preserve the shape and structure information from target prompt P and query contents from the source image with prompt P_s.

a significant limitation in synthesizing scenes with background dynamics.

5.5. Ablation Study

The results of both synthetic and real image editing can demonstrate the effectiveness of the proposed mutual self-attention control. We further analyze the control strategy in terms of different starting steps at the denoising process and the layers in the denoising U-Net. From Fig. 9(a), we see that performing mutual self-attention in the premature step can only synthesize an image identical to the source image, conveying all source image contents and ignoring the layout from the target prompt. As the step increases, the synthesized desired image can maintain the layout from the target prompt and the contents from the source image. While the image would gradually lose the source image contents and eventually becomes the image synthesized images without mutual self-attention control. We also observe a similar phenomenon when performing control in different U-Net layers shown in Fig. 9(b). Performing control among all layers can only generate an image identical to the source image. Performing control in low-resolution layers (*i.e.*, layer $4 \sim 10$) cannot preserve the source image contents and target layouts. While in high-resolution layers (*i.e.*, layer $0 \sim 3$, $10 \sim 15$), the target layout can be maintained, and the source image contents can only be transformed when controlled in the decoder part. As a result, the proposed method performs control in the decoder part of U-Net after several denoising steps.

6. Limitations and Discussion

Our method inherits most of the limitations of the Stable Diffusion model in generating desired images with text prompts. Please refer to the supplementary materials for detailed analysis and discussion. In addition, when editing real images, the DDIM inversion may fail to reconstruct the source images. In this case, MasaCtrl would fail to edit these images.

7. Conclusion

We propose MasaCtrl, a tuning-free mutual self-attention control method applied to T2I diffusion models for non-rigid consistent image synthesis and editing. We convert the self-attention mechanism in diffusion models into a cross one, dubbed mutual self-attention, enabling effective structure and appearance query from the source image when applied to specific denoising steps and U-Net layers. We further consider the confusion problem of the foreground objects and background during the querying process and alleviate it with a mask-guided strategy. Meanwhile, our method can be easily integrated into recently proposed controllable strategies over diffusion models and perform consistent image synthesis and editing without fine-tuning the model and textural embedding. We believe such a method provides ordinary users with a convenient and effective way for content creation under text description.

Acknowledgement. This research was partly supported by JSPS KAKENHI Grant Numbers 22H00529, 20H05951.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022.

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

[3] 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. *arXiv preprint arXiv:2301.13826*, 2023.

[4] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *ECCV*, pages 88–105. Springer, 2022.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

[6] 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. *NeurIPS*, 34:19822–19835, 2021.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021.

[8] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] 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.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[12] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

[13] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2426–2435, 2022.

[14] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *NeurIPS*, 32, 2019.

[15] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *CVPR*, pages 7880–7889, 2020.

[16] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[17] 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.

[18] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *NeurIPS*, 31, 2018.

[19] 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.

[20] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021.

[21] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023.

[22] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094, 2021.

[23] 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 *ICML*, pages 8748–8763. PMLR, 2021.

[24] 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.

[25] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021.

[26] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069. PMLR, 2016.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[29] 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. *arXiv preprint arXiv:2208.12242*, 2022.

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[32] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32, 2019.

[33] Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022.

[34] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *CVPR*, pages 16515–16525, 2022.

[35] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[37] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

[38] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, pages 2256–2265, 2021.

[39] 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 *CVPR*, pages 1316–1324, 2018.

[40] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021.

[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*, 2022.

[42] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, pages 833–842, 2021.

[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 *ICCV*, pages 5907–5915, 2017.

[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 TPAMI*, 41(8):1947–1962, 2018.

[45] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[46] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, pages 5802–5810, 2019.