

AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks

Tao Xu^{*1}, Pengchuan Zhang², Qiuyuan Huang²,
Han Zhang³, Zhe Gan⁴, Xiaolei Huang¹, Xiaodong He²

¹Lehigh University ²Microsoft Research ³Rutgers University ⁴Duke University
 {tax313, xih206}@lehigh.edu, {penzhan, qihua, xiaohc}@microsoft.com
 han.zhang@cs.rutgers.edu, zhe.gan@duke.edu

Abstract

In this paper, we propose an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. With a novel attentional generative network, the AttnGAN can synthesize fine-grained details at different sub-regions of the image by paying attentions to the relevant words in the natural language description. In addition, a deep attentional multimodal similarity model is proposed to compute a fine-grained image-text matching loss for training the generator. The proposed AttnGAN significantly outperforms the previous state of the art, boosting the best reported inception score by 14.14% on the CUB dataset and 170.25% on the more challenging COCO dataset. A detailed analysis is also performed by visualizing the attention layers of the AttnGAN. It for the first time shows that the layered attentional GAN is able to automatically select the condition at the word level for generating different parts of the image.

1. Introduction

Automatically generating images according to natural language descriptions is a fundamental problem in many applications, such as art generation and computer-aided design. It also drives research progress in multimodal learning and inference across vision and language, which is one of the most active research areas in recent years [20, 18, 31, 19, 4, 29, 5, 1, 30]

Most recently proposed text-to-image synthesis methods are based on Generative Adversarial Networks (GANs) [6]. A commonly used approach is to encode the whole text description into a global sentence vector as the condition for GAN-based image generation [20, 18, 31, 32]. Although



Figure 1. Example results of the proposed AttnGAN. The first row gives the low-to-high resolution images generated by G_0 , G_1 and G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively. Here, images of G_0 and G_1 are bilinearly upsampled to have the same size as that of G_2 for better visualization.

impressive results have been presented, conditioning GAN only on the global sentence vector lacks important fine-grained information at the word level, and prevents the generation of high quality images. This problem becomes even more severe when generating complex scenes such as those in the COCO dataset [14].

To address this issue, we propose an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. The overall architecture of the AttnGAN is illustrated in Figure 2. The model consists of two novel components. The first component is an attentional gener-

^{*}work was performed when was an intern with Microsoft Research

ative network, in which an attention mechanism is developed for the generator to draw different sub-regions of the image by focusing on words that are most relevant to the sub-region being drawn (see Figure 1). More specifically, besides encoding the natural language description into a global sentence vector, each word in the sentence is also encoded into a word vector. The generative network utilizes the global sentence vector to generate a low-resolution image in the first stage. In the following stages, it uses the image vector in each sub-region to query word vectors by using an attention layer to form a word-context vector. It then combines the regional image vector and the corresponding word-context vector to form a multimodal context vector, based on which the model generates new image features in the surrounding sub-regions. This effectively yields a higher resolution picture with more details at each stage. The other component in the AttnGAN is a Deep Attentional Multimodal Similarity Model (DAMSM). With an attention mechanism, the DAMSM is able to compute the similarity between the generated image and the sentence using both the global sentence level information and the fine-grained word level information. Thus, the DAMSM provides an additional fine-grained image-text matching loss for training the generator.

The contribution of our method is threefold. (*i*) An Attentional Generative Adversarial Network is proposed for synthesizing images from text descriptions. Specifically, two novel components are proposed in the AttnGAN, including the attentional generative network and the DAMSM. (*ii*) Comprehensive study is carried out to empirically evaluate the proposed AttnGAN. Experimental results show that the AttnGAN significantly outperforms previous state-of-the-art GAN models. (*iii*) A detailed analysis is performed through visualizing the attention layers of the AttnGAN. For the first time, it is demonstrated that the layered conditional GAN is able to automatically attend to relevant words to form the condition for image generation.

2. Related Work

Generating high resolution images from text descriptions, though very challenging, is important for many practical applications such as art generation and computer-aided design. Recently, great progress has been achieved in this direction with the emergence of deep generative models [12, 26, 6]. Mansimov *et al.* [15] built the alignDRAW model, extending the Deep Recurrent Attention Writer (DRAW) [7] to iteratively draw image patches while attending to the relevant words in the caption. Nguyen *et al.* [16] proposed an approximate Langevin approach to generate images from captions. Reed *et al.* [21] used conditional PixelCNN [26] to synthesize images from text with a multi-scale model structure. Compared with other deep generative models, Generative Adversarial Networks

(GANs) [6] have shown great performance for generating sharper samples [17, 3, 23, 13, 10]. Reed *et al.* [20] first showed that the conditional GAN was capable of synthesizing plausible images from text descriptions. Their follow-up work [18] also demonstrated that GAN was able to generate better samples by incorporating additional conditions (*e.g.*, object locations). Zhang *et al.* [31, 32] stacked several GANs for text-to-image synthesis and used different GANs to generate images of different sizes. However, all of their GANs are conditioned on the global sentence vector, missing fine-grained word level information for image generation.

The attention mechanism has recently become an integral part of sequence transduction models. It has been successfully used in modeling multi-level dependencies in image captioning [29], image question answering [30] and machine translation [2]. Vaswani *et al.* [27] also demonstrated that machine translation models could achieve state-of-the-art results by solely using an attention model. In spite of these progress, the attention mechanism has not been explored in GANs for text-to-image synthesis yet. It is worth mentioning that the alignDRAW [15] also used LAPGAN [3] to scale the image to a higher resolution. However, the GAN in their framework was only utilized as a post-processing step without attention. To our knowledge, the proposed AttnGAN for the first time develops an attention mechanism that enables GANs to generate fine-grained high quality images via multi-level (*e.g.*, word level and sentence level) conditioning.

3. Attentional Generative Adversarial Network

As shown in Figure 2, the proposed Attentional Generative Adversarial Network (AttnGAN) has two novel components: the attentional generative network and the deep attentional multimodal similarity model. We will elaborate each of them in the rest of this section.

3.1. Attentional Generative Network

Current GAN-based models for text-to-image generation [20, 18, 31, 32] typically encode the whole-sentence text description into a single vector as the condition for image generation, but lack fine-grained word level information. In this section, we propose a novel attention model that enables the generative network to draw different sub-regions of the image conditioned on words that are most relevant to those sub-regions.

As shown in Figure 2, the proposed attentional generative network has m generators (G_0, G_1, \dots, G_{m-1}), which take the hidden states (h_0, h_1, \dots, h_{m-1}) as input and generate images of small-to-large scales ($\hat{x}_0, \hat{x}_1, \dots, \hat{x}_{m-1}$).

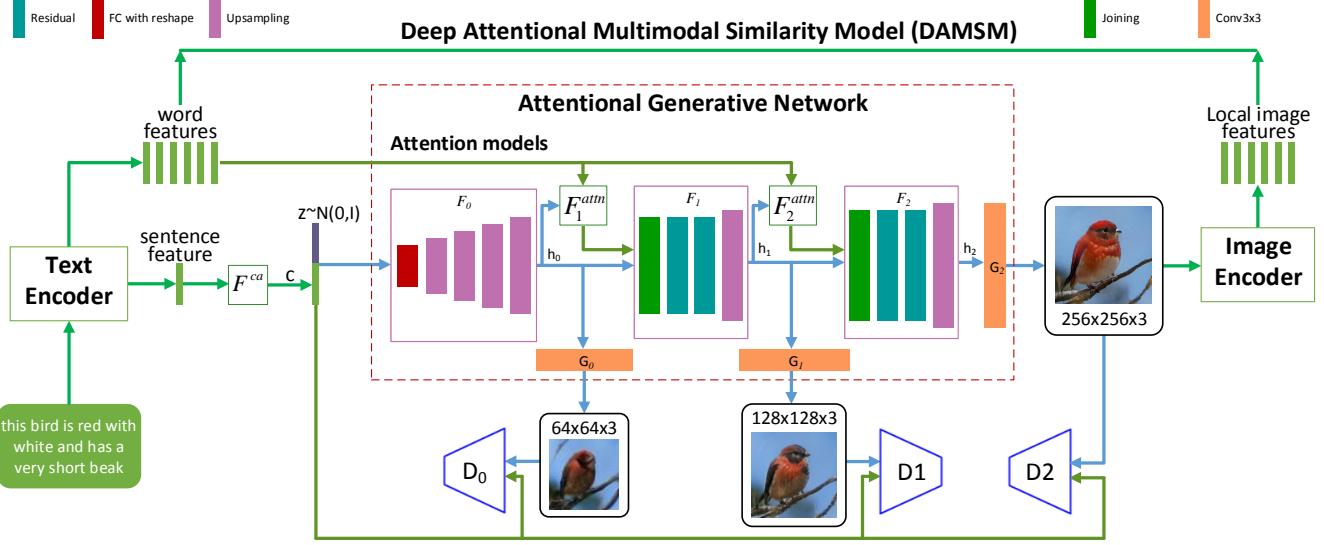


Figure 2. The architecture of the proposed AttnGAN. Each attention model automatically retrieves the conditions (*i.e.*, the most relevant word vectors) for generating different sub-regions of the image; the DAMSM provides the fine-grained image-text matching loss for the generative network.

Specifically,

$$\begin{aligned}
 h_0 &= F_0(z, F^{ca}(\bar{e})); \\
 h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, \dots, m-1; \\
 \hat{x}_i &= G_i(h_i).
 \end{aligned} \tag{1}$$

Here, z is a noise vector usually sampled from a standard normal distribution. \bar{e} is a global sentence vector, and e is the matrix of word vectors. F^{ca} represents the Conditioning Augmentation [31] that converts the sentence vector \bar{e} to the conditioning vector. F_i^{attn} is the proposed attention model at the i^{th} stage of the AttnGAN. F^{ca} , F_i^{attn} , F_i , and G_i are modeled as neural networks.

The attention model $F^{attn}(e, h)$ has two inputs: the word features $e \in \mathbb{R}^{D \times T}$ and the image features from the previous hidden layer $h \in \mathbb{R}^{\hat{D} \times N}$. The word features are first converted into the common semantic space of the image features by adding a new perceptron layer, *i.e.*, $e' = Ue$, where $U \in \mathbb{R}^{\hat{D} \times D}$. Then, a word-context vector is computed for each sub-region of the image based on its hidden features h (query). Each column of h is a feature vector of a sub-region of the image. For the j^{th} sub-region, its word-context vector is a dynamic representation of word vectors relevant to h_j , which is calculated by

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \text{ where } \beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1} \exp(s'_{j,k})}, \tag{2}$$

$s'_{j,i} = h_j^T e'_i$, and $\beta_{j,i}$ indicates the weight the model attends to the i^{th} word when generating the j^{th} sub-region of the image. We then denote the word-context matrix for image feature set h by $F^{attn}(e, h) = (c_0, c_1, \dots, c_{N-1}) \in \mathbb{R}^{\hat{D} \times N}$.

Finally, image features and the corresponding word-context features are combined to generate images at the next stage.

To generate realistic images with multiple levels (*i.e.*, sentence level and word level) of conditions, the final objective function of the attentional generative network is defined as

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \text{ where } \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}. \tag{3}$$

Here, λ is a hyperparameter to balance the two terms of Eq. (3). The first term is the GAN loss that jointly approximates conditional and unconditional distributions [32]. At the i^{th} stage of the AttnGAN, the generator G_i has a corresponding discriminator D_i . The adversarial loss for G_i is defined as

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(D_i(\hat{x}_i))] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(1 - D_i(\hat{x}_i, \bar{e})]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(1 - D_i(\hat{x}_i, \bar{e})]}_{\text{conditional loss}}, \tag{4}$$

where the unconditional loss determines whether the image is real or fake while the conditional loss determines whether the image and the sentence match or not.

Alternately to the training of G_i , each discriminator D_i is trained to classify the input into the class of real or fake by minimizing the cross-entropy loss defined by

$$\begin{aligned}
 \mathcal{L}_{D_i} = & \underbrace{-\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i)] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(1 - D_i(\hat{x}_i))]}_{\text{unconditional loss}} + \\
 & \underbrace{-\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i, \bar{e})] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(1 - D_i(\hat{x}_i, \bar{e})]}_{\text{conditional loss}},
 \end{aligned} \tag{5}$$

where x_i is from the true image distribution p_{data_i} at the i^{th} scale, and \hat{x}_i is from the model distribution p_{G_i} at the

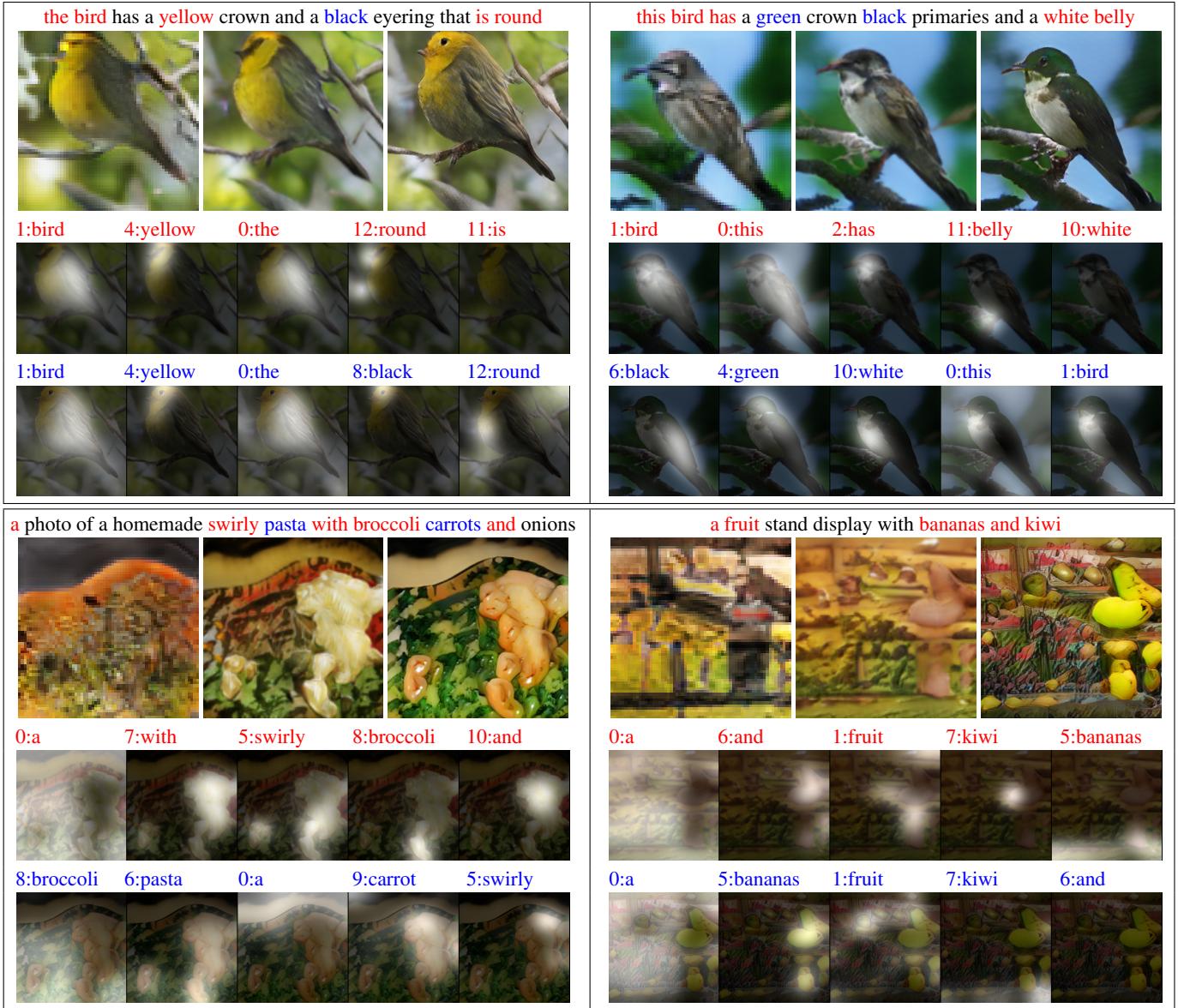


Figure 4. Intermediate results of our AttnGAN on CUB (top) and COCO (bottom) test sets. In each block, the first row gives 64×64 images by G_0 , 128×128 images by G_1 and 256×256 images by G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively. Refer to the supplementary material for more examples.

Dataset	GAN-INT-CLS [20]	GAWWN [18]	StackGAN [31]	StackGAN-v2 [32]	PPGN [16]	Our AttnGAN
CUB	$2.88 \pm .04$	$3.62 \pm .07$	$3.70 \pm .04$	$3.82 \pm .06$	/	$4.36 \pm .03$
COCO	$7.88 \pm .07$	/	$8.45 \pm .03$	/	$9.58 \pm .21$	$25.89 \pm .47$

Table 3. Inception scores by state-of-the-art GAN models [20, 18, 31, 32, 16] and our AttnGAN on CUB and COCO test sets.

ing complex scenarios like those in the COCO dataset.

To better understand what has been learned by the AttnGAN, we visualize its intermediate results with attention. As shown in Figure 4, the first stage of the AttnGAN (G_0) just sketches the primitive shape and colors of objects and generates low resolution images. Since only the global sentence vectors are utilized in this stage, the generated images

lack details described by exact words, *e.g.*, the beak and eyes of a bird. Based on word vectors, the following stages (G_1 and G_2) learn to rectify defects in results of the previous stage and add more details to generate higher-resolution images. Some sub-regions/pixels of G_1 or G_2 images can be inferred directly from images generated by the previous stage. For those sub-regions, the attention is equally allo-

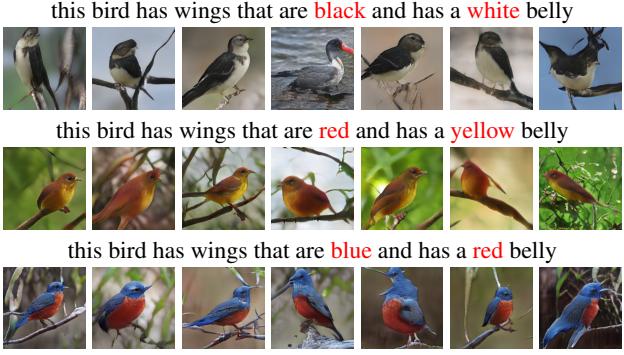


Figure 5. Example results of our AttnGAN model trained on CUB while changing some most attended words in the text descriptions.



Figure 6. 256×256 images generated from descriptions of novel scenarios using the AttnGAN model trained on COCO. (Intermediate results are given in the supplementary material.)



Figure 7. Novel images by our AttnGAN on the CUB test set.

cated to all words and shown to be black in the attention map (see Figure 4). For other sub-regions, which usually have semantic meaning expressed in the text description such as the attributes of objects, the attention is allocated to their most relevant words (bright regions in Figure 4). Thus, those regions are inferred from both word-context features and previous image features of those regions. As shown in Figure 4, on the CUB dataset, the words *the*, *this*, *bird* are usually attended by the F^{attn} models for locating the object; the words describing object attributes, such as colors and parts of birds, are also attended for correcting defects and drawing details. On the COCO dataset, we have similar observations. Since there are usually more than one object in each COCO image, it is more visible that the words describing different objects are attended by different sub-regions of the image, *e.g.*, *bananas*, *kiwi* in the bottom-right block of Figure 4. Those observations demonstrate that the AttnGAN learns to understand the detailed semantic meaning expressed in the text description of an image. Another

observation is that our second attention model F_2^{attn} is able to attend to some new words that were omitted by the first attention model F_1^{attn} (see Figure 4). It demonstrates that, to provide richer information for generating higher resolution images at latter stages of the AttnGAN, the corresponding attention models learn to recover objects and attributes omitted at previous stages.

Generalization ability. Our experimental results above have quantitatively and qualitatively shown the generalization ability of the AttnGAN by generating images from unseen text descriptions. Here we further test how sensitive the outputs are to changes in the input sentences by changing some most attended words in the text descriptions. Some examples are shown in Figure 5. It illustrates that the generated images are modified according to the changes in the input sentences, showing that the model can catch subtle semantic differences in the text description. Moreover, as shown in Figure 6, our AttnGAN can generate images to reflect the semantic meaning of descriptions of novel scenarios that are not likely to happen in the real world, *e.g.*, *a stop sign is floating on top of a lake*. On the other hand, we also observe that the AttnGAN sometimes generates images which are sharp and detailed, but are not likely realistic. As examples shown in Figure 7, the AttnGAN creates birds with multiple heads, eyes or tails, which only exist in fairy tales. This indicates that our current method is still not perfect in capturing global coherent structures, which leaves room to improve. To sum up, observations shown in Figure 5, Figure 6 and Figure 7 further demonstrate the generalization ability of the AttnGAN.

4.2. Comparison with previous methods

We compare our AttnGAN with previous state-of-the-art GAN models for text-to-image generation on CUB and COCO test sets. As shown in Table 3, on the CUB dataset, our AttnGAN achieves 4.36 inception score, which significantly outperforms the previous best inception score of 3.82. More impressively, our AttnGAN boosts the best reported inception score on the COCO dataset from 9.58 to 25.89, a 170.25% improvement relatively. The COCO dataset is known to be much more challenging than the CUB dataset because it consists of images with more complex scenarios. Existing methods struggle in generating realistic high-resolution images on this dataset. Examples in Figure 4 and Figure 6 illustrate that our AttnGAN succeeds in generating 256×256 images for various scenarios on the COCO dataset, although those generated images of the COCO dataset are not as photo-realistic as that of the CUB dataset. The experimental results show that, compared to previous state-of-the-art approaches, the AttnGAN is more effective for generating complex scenes due to its novel attention mechanism that catches fine-grained word level and sub-region level information in text-to-image generation.

