DreamArtist: Towards Controllable One-Shot Text-to-Image Generation via Contrastive Prompt-Tuning

Ziyi Dong
Sun Yat-Sen University
Guangzhou, China
dongzy6@mail2.sysu.edu.cn

Pengxu Wei
Sun Yat-Sen University
Guangzhou, China
weipx3@mail.sysu.edu.cn

Liang Lin
Sun Yat-Sen University
Guangzhou, China
linliang@ieee.org

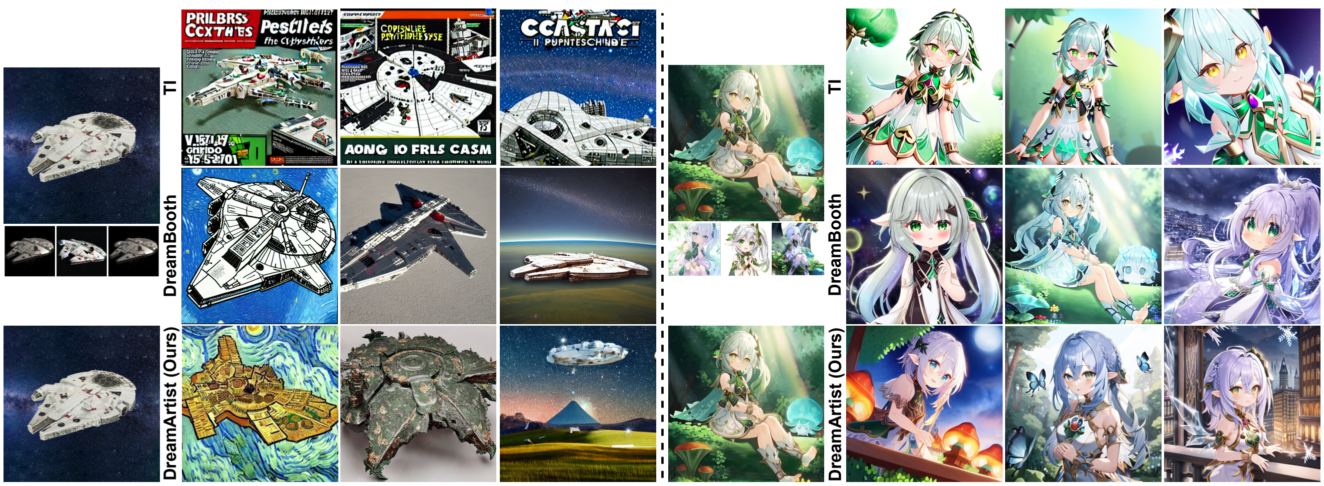

Figure 1. Given only one reference image (left), our DreamAritst can generate diverse images with the guidance of a text prompt, reproducing the image subjects in different contexts, styles, materials, details, *etc*. DreamAritst learns a pair of positive and negative pseudo-words (S_*^p and S_*^n) to represent the characteristics in that image by contrastive prompt-tuning. Instead, existing methods [9, 29] need a reference set with 3-5 images and still present a poor controllability to specifically and exactly embody the given text prompt. For instance, DreamBooth [29] is prone to generating images too similar to the training image and even with composition collapse.

Abstract

Large-scale text-to-image generation models with an exponential evolution can currently synthesize high-resolution, feature-rich, high-quality images based on text guidance. However, they are often overwhelmed by words of new concepts, styles, or object entities that always emerge. Although there are some recent attempts to use fine-tuning or prompt-tuning methods to teach the model a new concept as a new pseudo-word from a given reference image set, these methods are not only still difficult to synthesize diverse and high-quality images without distortion and artifacts, but also suffer from low controllability.

To address these problems, we propose a DreamArtist method that employs a learning strategy of contrastive prompt-tuning, which introduces both positive and negative embeddings as pseudo-words and trains them jointly. The positive embedding aggressively learns characteristics in the reference image to drive the model diversified generation, while the negative embedding introspects in a self-supervised manner to rectify the mistakes and inadequacies from positive embedding in reverse. It learns not only what is correct but also what should be avoided. Extensive experiments on image quality and diversity analysis, controllability analysis, model learning analysis and task expansion have demonstrated that our model learns not only concept but also form, content and context. Pseudo-words of DreamArtist have similar properties as true words to generate high-quality images.

1. Introduction

"Everyone is an artist." — Joseph Beuys.

Being an artist means a fundamental ability to create and be creative, with productive imaginations, specialized experiences and fantastic inspirations. Rome wasn't built in a day, but your childhood dreams can be! With the exponential evolution of generative models [2,6,14,15,22,28,30,34,35], the focus of research on Text-to-Image synthesis has gradually shifted from GAN to Diffusion [4, 5, 13, 17, 21–23, 26, 36], which probably realizes our artist dreams of producing creative, attractive, and fantastic image creations. More inspiring, only given the texts with classifier [20] or classifier-free [12] guidance, large-scale text-to-image models [8, 25, 31, 37, 38], *e.g.*, stable diffusion model [27], make it possible to synthesize high-resolution images with rich details and various characteristics, meeting our diverse personalized requirements.

Despite yielding impressive images, those models have absorbed large-scale images with prohibitive computation cost, and even need many words to depict a desirable image. In particular, they are possibly overwhelmed by words of new concepts, styles or object entities that always emerge; unfortunately, it is not a piece of cake to re-train those large-scale models.

Related work. To alleviate this problem, there are only two recent attempts Textual Inversion (TI) [9] and DreamBooth [29] (in Fig. 3). They try to teach the pre-trained large-scale text-to-image models a new concept as new words from 3-5 images with text guidance. DreamBooth [29] employs the fine-tuning strategy on a pre-trained text-to-image model and learns to bind a unique class-specific pseudo-word with that new concept. TI [9] learns an embedding as pseudo-words S_* to represent the concepts in the input images by prompt-tuning. Even though they present a great compelling potential for image generation, they also suffer from some typical limitations (Tab. 1). The fine-tuning strategy in DreamBooth not only requires tremendous computing (over 40GB VRAM), but also severely over-fits the training set with catastrophic forgetting. Thus, it suffers from context-appearance entanglement and the generated images are monotonous (in Fig. 3). Alternatively, the prompt-tuning in TI is energy-saving, but its generated images suffer from heavy artifacts and distortions, with low diversity. In the training phase, they require that any input noise guided by the pseudo-word should generate images as similar as possible to the training images. This makes the pseudo-word attract too much attention, so that it obscures almost all additional control signals, which leads to poor controllability and low diversity. Besides, both of them have to collect 3-5 input images as reference, which are usually with similar features and depict the same object. But those are usually multi-view images for the same object. It is not so convenient to prepare them and even the selected images are too much diverse, inviting unexpected challenges to the models.

Our work. In this paper, we propose a task of one-shot text-to-image generation, which uses only one image to teach the model to learn to represent its characteristics with a pseudo-word. And this word should generate diverse and highly controllable images like the model's other original words. It effectively produces high-quality and diverse images with one given reference image (not a reference set) and embraces various new renditions in different contexts. DreamArtist employs a learning strategy of contrastive prompt-tuning (CPT). Rather than describing the entire image with just positive embeddings in conventional prompt-tuning, our CPT jointly trains a paired positive and negative embedding (S_*^p and S_*^n) with pre-trained fixed text encoder $\mathcal{B}$ and denoising u-net ϵ_θ. S_*^p aggressively to learn the characteristics of the reference image to drive the model diversified generation, while S_*^n introspects in a self-supervised manner and rectifies the inadequacies of positive prompt in reverse. Due to the introduction of S_*^n, S_*^p does not need to be forcibly aligned with the input image. Thus S_*^p does not need to attract as much attention as TI to align the generated and training images, which brings diversity and high controllability. Without excessive attention to S_*^p, the characteristics included in the additional descriptions will not be obscured and will be clearly rendered in the synthesized images. Moreover, due to the rectification from S_*^n, these additional characteristics will be rendered as harmonious as those from S_*^p.

Method	TI [9]	DreamBooth [29]	Ours
Given image number	3-5	3-5	1
Parameters	2K	983M	5K
Image quality	artifacts, mosaic	over-smooth, artifacts	vivid
Diversity	poor	very poor	highly diverse
Controllability	poor	poor	high

Table 1. Comparison with current state-of-the-art methods.

Extensive experiments on generative models trained on the natural dataset LAION-5B [33] and the anime dataset Danbooru [1] demonstrate that our method substantially outperforms existing methods. Various experiments have indicated that our method outperforms existing methods in several aspects, including the quality and diversity of the generated images, the style similarity to the reference image, and the controllability subject to additional descriptions.

In summary, our main contributions are as follows:

1. Empirically, we have introduced the one-shot text-to-image generation task, which requires the model to learn controllable characteristics (form, content, context, style, semantics, etc.) from a single image, without forgetting the model's original generative abilities.

2. Technically, we propose a DreamArtist method, which introduces contrastive prompt tuning to train positive and negative pseudo-words jointly, allowing the model to learn high quality, diverse, and highly controllable characteristics from a single image.

3. Experimentally, extensive qualitative and quantitative

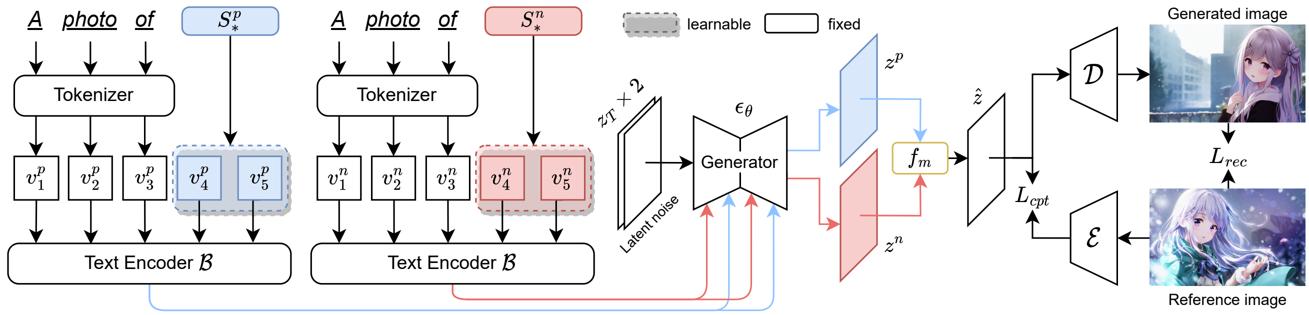

Figure 2. The framework of our DreamArtist. Only the embedding corresponding to positive and negative pseudo-words (S_*^p and S_*^n) can be learned, and the rest of the parameters are fixed, where f_m is the fusion function of z^p and z^n in Eq. (2).

experiments on both natural data and anime data domains have demonstrated that our method substantially outperforms existing methods in content quality, style similarity, and detail quality. Our approach can render highly harmonious and realistic images even combined with complex additional descriptions. The generated images are difficult to distinguish from those created by real human beings.

2. Methodology

To overcome the limitations of existing methods aforementioned, enabling the model to synthesize highly realistic and diversity images with high controllability through just one user-given image, our DreamArtist is proposed, shown in Fig. 2. DreamArtist introduces both positive and negative embeddings and jointly trains them with contrastive prompt-tuning through introspection in a self-supervised manner. Allowing embeddings to describe not only the characteristics we need, but also those that need to be excluded.

2.1. Latent Diffusion Model

With the remarkable capacity of image generation, the Latent Diffusion Model (LDM) [27] is utilized as the base model. Different from the conventional DDPM [11, 34] that performs denoising operations in the image space, LDM conducts this in the image space. This can apply diffusion operations to the feature space. Formally, firstly, an input image x is encoded into the feature space by an AutoEncoder $z = \mathcal{E}(x), \hat{x} = \mathcal{D}(z)$ (with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$) pre-trained with a large number of images, and then a Denoising U-Net ϵ_θ which consists of the Transformer is used to perform denoising on the feature map $z_{t-1} = \epsilon_\theta(z_t, t)$. The text-guided conditional image generation with text S and encoded text feature $y = \mathcal{B}(S)$ is implemented by the cross attention mechanism, using the transformed image features as query $W_Q^{(i)} \cdot \varphi_i(z_t)$ and the transformed text features as key and value $W_K^{(i)} \cdot \tau_\theta(y)$ and $W_V^{(i)} \cdot \tau_\theta(y)$, then its training loss can be expressed as:

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

where t represents the time step, z_t is the diffusion feature map of z at step t, and ϵ is the unscaled noise. In this training phase, AutoEncoder is fixed and only ϵ_θ is learnable.

2.2. Contrastive Prompt-Tuning

In essential, conventional prompt-tuning [3, 7, 16, 18, 19, 32, 40, 41] optimistically considers only a positive prompt. Namely, it simply aligns it with the downstream task and constructs the mapping from the prompt to the training set. However, this easily leads to collapse and over-fitting, when there are few samples given in the training stage. Especially, for one-shot text-to-image generation, it generates images with obvious artifacts and very low diversity. Accordingly, we propose contrastive prompt-tuning to avoid these problems. That is, it enables the model not only aggressively to learn the characteristics in the image, but also to rectify its mistakes following an introspection mechanism.

This is also analogous to human beings that also learn to paint through introspection with self-supervision. Keep trying and analyzing the differences with the masterpiece, and record and introspect the shortcomings to avoid repeating them in the future. After thousands of repetitions and refinements, a new master is yielded. If one just puts a Mona Lisa there and simply imitates it, without any introspection. Then, it easily has a consequence that this person can only draw some clumsy copy of the Mona Lisa, limiting the imagination for creations.

Our contrastive prompt-tuning also learns to paint through introspection in a self-supervised manner. Given two identical noise maps z_t, they can be guided separately by positive texts and negative texts (e.g. "a photo of a dog" and "a photo of a cat") and yield two different feature maps z^p and z^n. z^p contains the characteristics we need, while z^n contains the characteristics we prefer to be excluded in the generated image. According to the conclusion stated in [22], we can make the model extrapolated in the direction of z^p and away

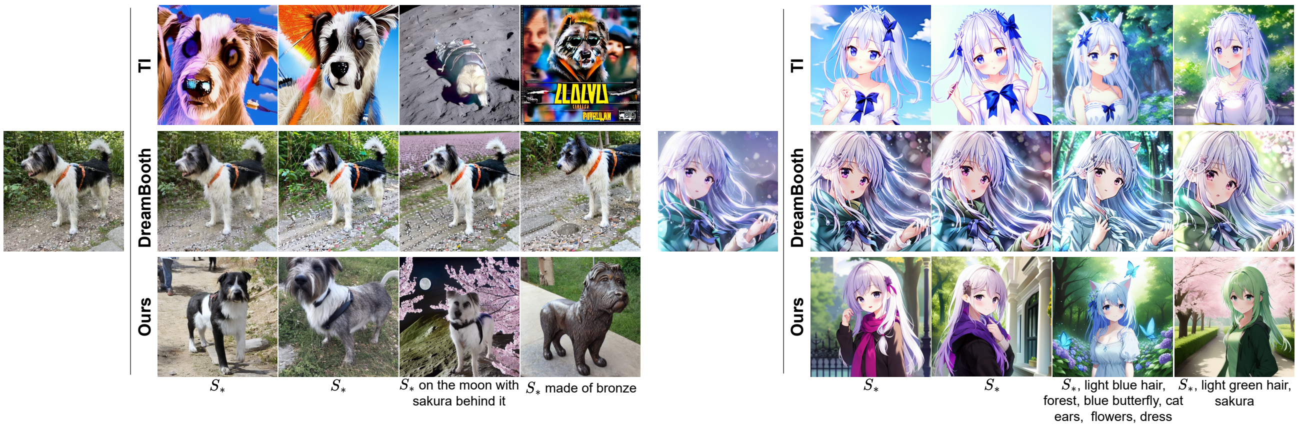

Figure 3. Comparison of our method with recent methods for one-shot text-to-image generation.

from z^n, with the following guiding strategy:

$$\hat{z} = f_m(z^p, z^n) = z^n + \gamma(z^p - z^n) \tag{2}$$

We add learnable pseudo-words (S_*^p and S_*^n) to both positive and negative texts, respectively, and fix pre-trained text-encoder and diffusion models. Then, we can realize contrastive prompt-tuning. We jointly learn two different embeddings, representing characteristics that are relevant to the training image and characteristics that are away from the training image or are found to be avoided from introspection. Learning to guide the model not only extrapolated in the desired direction, but also avoid inappropriate directions. Then the contrastive prompt-tuning loss is:

$$\begin{aligned} z^p = \epsilon_\theta(z_t, t, \tau_\theta(\mathcal{B}(S_*^p))), z^n = \epsilon_\theta(z_t, t, \tau_\theta(\mathcal{B}(S_*^n))) \\ \mathcal{L}_{cpt} = \| f_m(z^p, z^n) - \mathcal{E}(x) \|_2^2 \end{aligned} \tag{3}$$

where x is the image for training and $\| \cdot \|_2^2$ is the $\ell 2$ loss.

S_*^p portrays the primary forms of objects and contexts, while S_*^n rectifies the inadequacies of S_*^p in reverse. With the involvement of S_*^n, S_*^p no longer needs to attract excessive attention to force all the z_T to be guided to the given image x. This allows S_*^p to drive diversified image generation and does not attract too much attention, which could overwhelm additional control signals leading to poor controllability. Even when combined with some complex descriptions, DreamArtist is able to render high-quality and diverse images with the characteristics of these descriptions in the learned style harmoniously, which benefits from the rectifying effect from S_*^n.

2.3. Reconstruction Constraint for Detail Enhancement

Constraints in the feature space only, would make the generated images be smoothness and even with some deficiencies in details and colors. Thus, we add an additional pixel-level reconstruction constraint to enhance the embedding's ability on describing details and colors. After $\hat{z}$ is computed via Eq. (2), it is decoded by the decoder $\mathcal{D}$ and transformed into the image space $\hat{x} = \mathcal{D}(\hat{z})$. Make the decoded image $\hat{x}$ as consistent as possible with the training image x at each pixel, thus enhancing the learning of details. Accordingly, the reconstruction loss can be written as:

$$\mathcal{L}_{rec} = \| \mathcal{D}(f_m(z^p, z^n)) - x \| \tag{4}$$

where $\| \cdot \|$ is the $\ell 1$ loss.

3. Experiments

3.1. Experimental Settings

Dataset. Similar to TI and DreamBooth, the LAION-5B dataset [33] is used for natural image generation. Additionally, an anime dataset, Danbooru [1], is added for a popular interest on many applications, *e.g.*, games and animes.

Implementation details. The experiments on all domains were trained using one image with the learning rate of 0.0025 and the γ of 5. The training was performed on an RTX2080ti using a batch size of 1 with about 2k-8k iterations. We use an embedding occupying 6 words for TI, while we use 3 words for both positive and negative embeddings for our method. DreamBooth then follows the settings in the paper and employs prior-preservation loss to reduce over-fitting. All methods are trained with the one-shot setting.

Metrics. For quantitative analysis, we use LPIPS [39] and style loss [10] to measure feature similarity and style similarity with the training image, respectively. Another two evaluation metrics, CLIP detail score (CDS) and CSTD, are defined based on the CLIP [24] model. CDS uses CLIP to get the probability of an image belonging to "detailed" in B the set of ["little detail", "detailed"]. CSTD calculates the standard deviation of the feature map of the image encoded by the image encoder of CLIP.

For controllability evaluation, we define a CLIP feature score (CFS). The description of a feature is input to CLIP together with multiple descriptions that are similar to its

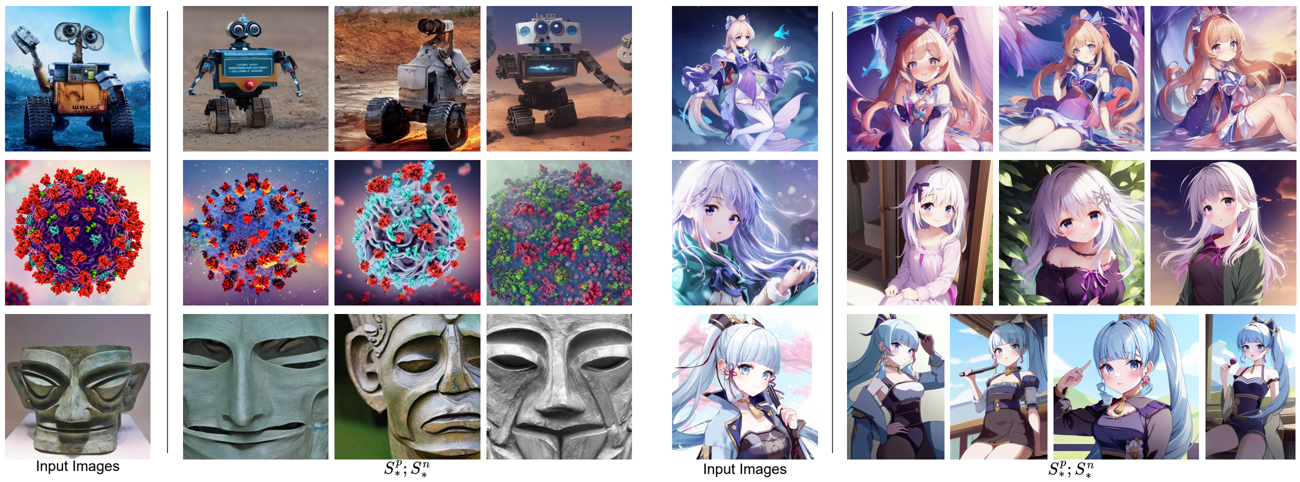

Figure 4. One-shot text-to-image generation with only learned pseudo-words for DreamArtist. It can learn content and context from a single image without adding additional text descriptions, generating diversity and high-quality images in both natural and anime scenes.

Method	LPIPS↓	Style loss↓	CDS↑	CSTD↑
Natural Image Generation				
TI	0.71	24.47	0.73	**1.79**
DreamBooth	**0.33**	**5.12**	0.63	0.69
Ours	0.62	9.84	**0.74**	1.53
Anime Image Generation				
TI	0.63	7.47	0.41	0.87
DreamBooth	**0.49**	1.16	0.33	0.72
Ours	0.60	**0.69**	**0.60**	**1.28**

Table 2. Quantitative comparison of our DreamArtist with existing methods for one-shot text-to-image generation.

Method	Natural Image Generation				Anime Image Generation			
	CFS↑	CSTD↑	Style loss↓	CDS↑	CFS↑	CSTD↑	Style loss↓	CDS↑
TI	0.37	1.46	17.26	0.40	0.23	0.98	5.52	0.46
DreamBooth	0.24	1.19	**1.36**	**0.69**	0.28	0.81	**1.28**	0.31
Ours	**0.89**	**1.55**	8.03	0.57	**0.63**	**1.15**	2.71	**0.58**

Table 3. Quantitative analysis of our DreamArtist compared with existing methods on feature controllability. The feature controllability of DreamArtist substantially exceeds existing methods.

semantic category but different. And for each feature on multiple images, the average of the probability that CLIP considers the image to belong to the semantic category we describe is calculated.

3.2. One-Shot Text-guided Image Synthesis

We compare our DreamArtist with two existing works, including TI [9] and DreamBooth [29] for one-shot text-to-image generation. All methods are trained with only one image given as a reference for a fair comparison. Next, we will elaborate the comparison results from image quality, diversity, characteristics and style similarities.

Image Quality and Diversity. From the qualitative analysis, it is shown in Fig. 3 that the images generated by the TI have serious artifacts and distortions. And the diversity is also low for generated anime images and few meaningful details are presented, while most are artifacts. DreamBooth generates images with few artifacts, but the diversity is incredibly low on both natural and anime scenes. It generates images overly similar to the reference image, which evidences an over-fitting issue. From Fig. 3 and 4, our DreamArtist can alleviate these problems and not only generates highly realistic images with remarkable light, shadow and detail, but also keeps the generated images highly diverse.

Quantitative analysis in Tab. 2 can also reach the same conclusion. According to the results of style similarity and qualitative analysis, it is an illusion that the CSTD metrics of TI are high. This is possibly caused by generated artifacts. The generated images by DreamBooth usually have extremely low diversity and low image quality. Our method, instead, performs well in both natural and anime scenes in terms of image details, quality, and diversity.

Generates New Concepts From Fig. 3, it is observed that, for TI, the style in the generated image differs greatly from the input reference image in natural scenes; in anime scenes, not only the style differs greatly, but also the coat of the input image is learned as a dress and the eyes color is incorrect. This indicates that TI is limited in learning the characteristics of the input image, which cannot learn the content and style effectively. Although DreamBooth can well learn the characteristics in the input image, the over-fitting is too serious because it tends to simply remember the whole image. Our method generates images that are highly stylized and consistent with the input image; the form, content, and context are also well learned with aesthetics.

The quantitative results in Tab. 2 also support the above observations. Style loss of TI is extremely high and LPIPS is also not low, indicating that it really cannot learn the features of the input image effectively. Our method, on the other hand, is able to show a high style and content similarity while maintaining considerable performance in all other aspects.

Style Cloning. Compared to the abstract style like Vincent

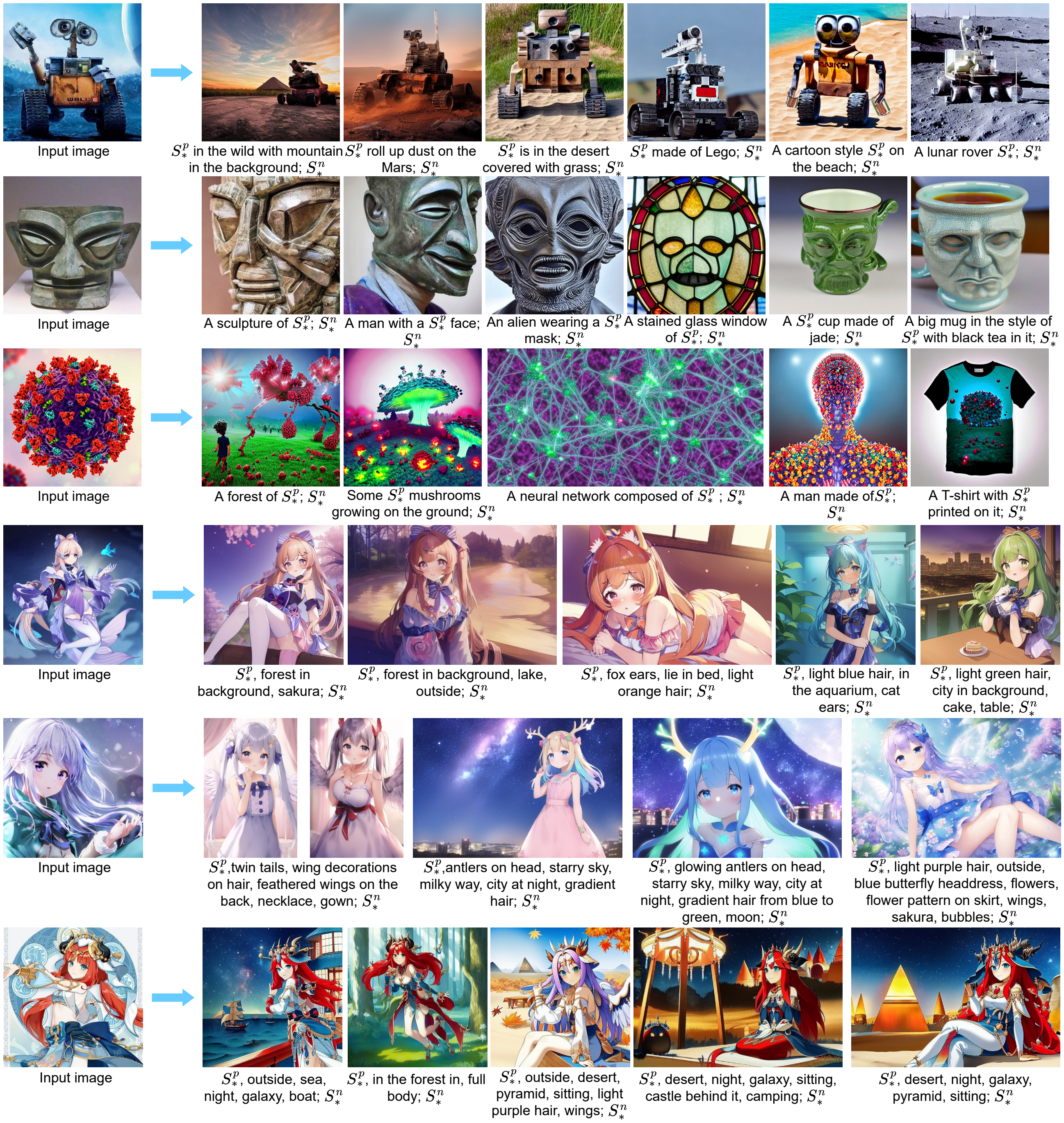

Figure 5. The generated images of DreamArtist with the guidance of additional complex texts. DreamArtist exhibits a superiors capability of controllable generation: even with few words in the text guidance, diverse and faithful images are generated; with more words, vivid images with rich details are generated. More importantly, DreamArtist can successfully render almost all the given words.

Van Gogh's "The Starry Night". More often than not, users need some more practical styles, for example, the style of a movie or game scene, the style of Cyberpunk or Steampunk, the style of a Chinese Brush Painting or paper-cut, or the painting style of your favorite artist.

Existing methods can only learn some highly abstract styles and the generated images are difficult to show much meaningful content, which is more like some textures. As can be seen in Fig. 6, our method can learn some practical and highly refined styles pretty well. The generated images are identical to the training images in terms of colors and texture features, cloning the style of the training images remarkably.

Figure 6. Style cloning via DreamArtist, for example, stytles of wash painting, paper-cut art, Cyberpunk, comic of caricaturists and road map in a game (from left to right).

Thus, our model substantially outperforms existing methods.

In anime scenes, different artists have different painting styles of different brushwork, composition, light processing, color processing, scenery, and many other details. The different painting styles are not as diverse as the different styles in natural scenes, but they will give the reader a completely different impression. Our method can learn a painting style fairly well. It is even possible to create images that are highly similar to other works by the same artist based on the text description, which is difficult for existing methods.

3.3. Method Evaluation and Analysis

Controllability and Compatibility Analysis. A successful one-shot text-guided image synthesis method should not only be able to learn the characteristics in the reference image, but also should enable these characteristics to be controlled by additional descriptions. From the Fig. 3, we observe that the pseudo-words learned by the TI method are difficult to combine with some additional features. For instance, in the natural scene "on the moon with sakura behind it", only "on the moon" works, while "made of bronze" does not work at all. DreamBooth has slightly better compatibility with the additional descriptions, but the generated images look almost the same due to severe overfitting, and some of the described features are still not rendered.

Our method, as an alternative, can effectively solve these problems. As can be seen in Fig. 3 and 5 and the CFS in Tab. 3 that DreamArtist can not only be easily compatible with additional complex descriptions, but also generate highly harmonious and diverse images with those descriptions and learned features. For instance, the second row of the learned embedding of a mask can render richly diverse and highly realistic images in controlled by various complex descriptions. Besides, the described features can be rendered, even if there is a conflict between the additional description and the learned features. For example, for the last one in the fourth row of Fig. 5, the training image has pink hair with a pure background, while our method can generate descriptions that require characters with light green hair and with a city in the background. DreamAritst can really follow the user's description to create and be creative, with productive imaginations, specialized experiences and fantastic inspirations.

Model Ablative Analysis. To verify the roles of S_*^p and S_*^n, we visualize the generated images guided by S_*^p and S_*^n, respectively, shown in Fig. 9. While S_*^p portrays the basic layout and form, it lacks in features, style, and details. S_*^n shows some distortions and unreasonable styles to rectify inadequacies in S_*^p. For example, the second row of S_*^p generates a rough painting with the wrong style. S_*^n points out and rectifies these defects very well. Combining them can generate images that are not only rich in details with reasonable characteristics, but also highly consistent in style with the input image.

3.4. Human Evaluation

To demonstrate that our method can synthesize high-quality realistic images, we have conducted a user study following the rules of the Turing test with 700 subjects for TI and our methods, respectively. There are 12 synthetic and 8 real images in the test. The TI method makes subjects have a failure rate of 26.6% in the test. While our method makes the subjects failure rate reach 34.5%, which has significantly exceeded the Turing test requirement of 30%. This shows that the images generated by our method are fairly realistic and difficult to be distinguished from the real images.

Besides, in the study of judging which creation has higher quality, 52.31% and 83.13% of the subjects from various

Figure 7. Results of concept compositions via DreamArtist. It presents a promising generation potential via the pseodu text guidance from the arbitrary combination of the learnt pseodu-words.

Figure 8. Text-guided image editing via DreamArtist.

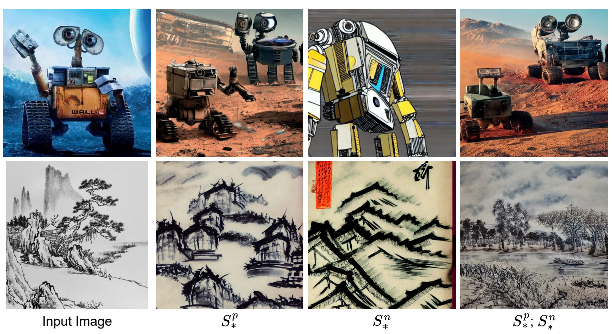

Figure 9. Ablation study on S^p_* and S^n_* in DreamArtist. It shows the generated images with the guidance of S^p_*, S^n_* and both, respectively.

walks of life and even including some professional anime artists, have chosen the DreamArtist synthesized creation in the nature and anime scenes, respectively.

3.5. Extended Task 1: Concept Compositions

Our method can easily combine multiple learned pseudo-words, not only limited to combining objects and styles, but also using both objects or style, and generating some reasonable images. In combining these pseudo-words, it is necessary to add both parts of the embedding in the positive and negative prompts. As illustrated in Fig. 7, combining multiple pseudo-words trained with our method show excellent results in both natural and anime scenes. Each component of the pseudo-words can be rendered in the generated image, even combining two radically different objects or styles. For example, we can have a robot painted in the style of an ancient painting, or make a dog have a robot body. These are difficult to realize for existing methods. For example, in the work of TI, it mentions that TI is struggling to combine multiple pseudo-words [9].

3.6. Extended Task 2: Prompt-Guided Image Editing

As seen from Fig. 8, the pseudo-words learned with our method works well for text-guided image editing that follows the paradigm of LDM [27]. The modified areas not only show the learned form and content, but also integrate well into the environment, which looks harmonious. The performance of image editing with the learned features of our method is as effective as that of the original features from LDM.

4. Conclusions

We introduce a one-shot text-to-image generation task, using only one reference image to teach a text-to-image model to learn some new characteristics as a pseudo-word. According to the shortcomings of existing methods, we propose a DreamArtist method. DreamArtist employs a learning strategy of contrastive prompt-tuning in a self-supervised manner, enabling the model to learn from introspection, which no longer forces the positive pseudo-word to align with the reference image. With contrastive prompt-tuning, pseudo-words can not only make the model generates high-quality and diverse images, but also can be easily controlled by additional descriptions. DreamArtist not only learns concepts in images, but also form, content and context. Extensive qualitative and quantitative experimental analyses have demonstrated that our method substantially outperforms existing methods in various aspects. Moreover, our DreamArtist method is highly controllable and can be used in combination with complex descriptions without losing any components, presenting a promising flexibility for deploying other models.

References

[1] Anonymous, Danbooru community, and Gwern Branwen. Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset. https://www.gwern.net/Danbooru2021, 2022. 2, 4

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 2

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[4] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10908–10917, 2020. 2

[5] 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 *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII*, volume 13697, pages 88–105, 2022. 2

[6] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. 2

[7] Kun Ding, Ying Wang, Pengzhang Liu, Qiang Yu, Haojian Zhang, Shiming Xiang, and Chunhong Pan. Prompt tuning with soft context sharing for vision-language models. *CoRR*, abs/2208.13474, 2022. 3

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 19822–19835, 2021. 2

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *CoRR*, abs/2208.01618, 2022. 1, 2, 5, 8

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2414–2423, 2016. 4

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[13] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 857–866, 2022. 2

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 852–863, 2021. 2

[15] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10236–10245, 2018. 2

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059, 2021. 3

[17] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2063–2073, 2019. 2

[18] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597, 2021. 3

[19] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602, 2021. 3

[20] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach,

and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *CoRR*, abs/2112.05744, 2021. 2

[21] Yi Liu, Jialiang Peng, James Jian Qiao Yu, and Yi Wu. PPGAN: privacy-preserving generative adversarial network. In *25th IEEE International Conference on Parallel and Distributed Systems, ICPADS 2019, Tianjin, China, December 4-6, 2019*, pages 985–989, 2019. 2

[22] Alexander Quinn 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. In *International Conference on Machine Learning, ICML 2022*, volume 162, pages 16784–16804, 2022. 2, 3

[23] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1505–1514. Computer Vision Foundation / IEEE, 2019. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139, pages 8748–8763, 2021. 4

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 2

[26] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48, pages 1060–1069, 2016. 2

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685, 2022. 2, 3, 8

[28] 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*, volume 9351, pages 234–241, 2015. 2

[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. *CoRR*, abs/2208.12242, 2022. 1, 2, 5

[30] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 15:1–15:10, 2022. 2

[31] 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, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *CoRR*, abs/2205.11487, 2022. 2

[32] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269, 2021. 3

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. *CoRR*, abs/2210.08402, 2022. 2, 4

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2, 3

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2

[36] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Fei Wu, and Xiao-Yuan Jing. DF-GAN: deep fusion generative adversarial networks for text-to-image synthesis. *CoRR*, abs/2008.05865, 2020. 2

[37] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *CoRR*, abs/2206.10789, 2022. 2

[38] Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *CoRR*, abs/2112.15283, 2021. 2

[39] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595, 2018. 4

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804, 2022. 3

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 3