DreamArtist: Controllable One-Shot Text-to-Image Generation via Positive-Negative Adapter

Ziyi Dong[1,2], Pengxu Wei[1,2], Liang Lin[1,2*]

[1]School of Computer Science and Technology, Sun Yat-sen Unviersity, Guangzhou, China.
[2]Peng Cheng Laboratory, Shenzhen, China.

Abstract

State-of-the-arts text-to-image generation models such as Imagen [1] and Stable Diffusion Model [2] have succeed remarkable progresses in synthesizing high-quality, feature-rich images with high resolution guided by human text prompts. Since certain characteristics of image content *e.g.*, very specific object entities or styles, are very hard to be accurately described by text, some example-based image generation approaches have been proposed, *i.e.* generating new concepts based on absorbing the salient features of a few input references. Despite of acknowledged successes, these methods have struggled on accurately capturing the reference examples' characteristics while keeping diverse and high-quality image generation, particularly in the one-shot scenario (*i.e.* given only one reference). To tackle this problem, we propose a simple yet effective framework, namely DreamArtist, which adopts a novel positive-negative prompt-tuning learning strategy on the pre-trained diffusion model, and it has shown to well handle the trade-off between the accurate controllability and fidelity of image generation with only one reference example. Specifically, our proposed framework incorporates both positive and negative embeddings or adapters and optimizes them in a joint manner. The positive part aggressively captures the salient characteristics of the reference image to drive diversified generation and the negative part rectifies inadequacies from the positive part. We have conducted extensive experiments and evaluated the proposed method from image similarity (fidelity) and diversity, generation controllability, and style cloning. And our DreamArtist has achieved a superior generation performance over existing methods. Besides, our additional evaluation on extended tasks, including concept compositions and prompt-guided image editing, demonstrates its effectiveness for more applications.
DreamArtist project page: https://www.sysu-hcp.net/projects/dreamartist/index.html

Keywords: Text-to-Image Generation, Diffusion Models, One-Shot Learning, Controllable Image Generation, Prompt Learning

1 Introduction

"Imagination will take you everywhere."
—Albert Einstein.

With productive imaginations and fantastic inspirations, everyone can be an artist, creating and being creative, which is the goal of the recently rising visual content generation research [2, 1, 3]. Thanks to the exponential evolution of generative models [4, 5, 6, 7, 8, 9, 10, 11], we have witnessed the rapid progress of GAN and diffusion models on Text-to-Image synthesis [12, 13, 14, 15, 16, 17, 18, 19, 20]. Even more inspiring, given only texts with classifier [21] or classifier-free [22] guidance, large-scale text-to-image models [3, 23, 1, 24, 25], such as LDM [2], enable the synthesis of high-resolution images with rich details and various characteristics, fulfilling our diverse *personalized* requirements. Despite yielding impressive images, these models require numerous words to depict a desirable complex image. Furthermore, they may

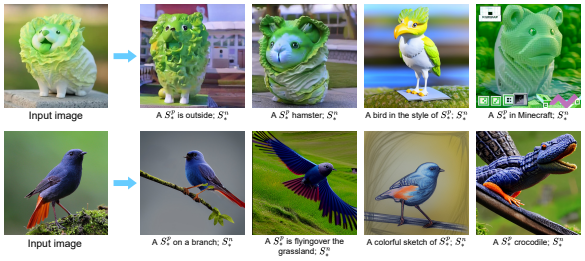

Fig. 1 DreamArtist excels in learning to generate relevant, high-quality, diverse, and highly controllable images from only one reference image. Furthermore, it has the ability to incorporate certain abstract features from the reference image to create novel visual compositions.

struggle with words describing new concepts, styles, or object entities for image generation.

To alleviate this problem, few attempts have been made, with Textual Inversion (TI) [27] and DreamBooth [28] being two notable examples. These methods aim to teach a pre-trained large-scale text-to-image model a new concept from a limited set of 3-5 images and associate it with a new pseudo-word. This newly learned pseudo-word can then be incorporated into the text-to-image generation process. Specifically, DreamBooth [28] employs a *fine-tuning* strategy on a pre-trained model, learning to bind a unique class-specific pseudo-word with the new concept. In contrast, TI [27] learns an embedding as a pseudo-word S_* to represent concepts in input images through *prompt-tuning*.

Despite the significant potential for image generation offered by these methods, they also show several limitations. DreamBooth necessitates fine-tuning the pre-trained model, i.e., optimizing a vast number of parameters (see Table 1), using only a few reference images. This approach often results in generated images that are monotonous and lack diversity, as illustrated in Figures 3 and 5. On the other hand, TI employs an energy-efficient prompt-tuning technique by optimizing only a limited number of parameters. Nevertheless, a notable drawback of this method is its usually failure to faithfully generate image details specified by prompt keywords. As depicted in Figure 2, crucial elements such as "flowers," "lake," and "city" are absent in the images generated by TI. Even with an increase in the number of reference images (from one-shot to 3-5 instances), this issue persists, as evidenced in Figure 3. The primary reason for these shortcomings lies in the absence of a more effective learning strategy to enhance generation controllability during prompt-tuning with a minimal number of reference images. Furthermore, incorporating multiple reference images may introduce unexpected ambiguities during the generation process, while gathering a collection of instances for a specific object would require additional effort. Therefore, it is desirable that the method can generate image content to be precisely aligned with the original texts, while emphasizing the key elements extracted from the reference images.

In this work, we introduce *DreamArtist*, a simple yet effective approach for one-shot text-to-image generation. *DreamArtist* is designed to learn a new concept from a single reference image (not a reference set), utilizing a pre-trained diffusion model. This method incorporates a *positive-negative adapter* learning strategy, which adeptly balances the preservation of the reference's specific characteristics with text-guided generation controllability. Contrary to traditional methods such as prompt-tuning or fine-tuning (including adapters like LoRA [29]), which learn an embedding (S_*) or a model (ϵ_θ) through positive guidance, *DreamArtist* acquires both positive and negative embeddings (S_*^p and S_*^n) as well as adapters (ϕ^p and ϕ^n). These embeddings and adapters are based on the pre-trained text encoder $\mathcal{B}$ and denoising U-Net ϵ_θ. Specifically, the positive components (S_*^p and ϕ^p) aggressively capture the characteristics of the reference image, thereby promoting diverse generation. Conversely, the negative components (S_*^n and ϕ^n) introspect in a self-induced manner to rectify the limitations inherent in the positive components. The inclusion of negative components introduces corrective information and facilitate rectifying the new concept from a differentiated (negative) aspect. The integration of these negative components relaxes the optimization objective, moving beyond a rigid adherence to the input reference. This shift facilitates the generation of highly diverse images and significantly enhances controllability, allowing the seamless integration of newly learned concepts with the original visual content. Importantly, since S_*^n serves to rectifies S_*^p specifically, the disturb on other textual features is minimal. We have conducted extensive experiments on the natural image dataset LAION-2B [30] and the anime dataset Danbooru [31]. The results demonstrate that our *DreamArtist* method achieves a substantial improvement over existing techniques.

Overall, our contributions are summarized as follows:

First, we propose a novel one-shot text-to-image approach, DreamArtist, which enables users to express their creativity in painting based on one reference image.

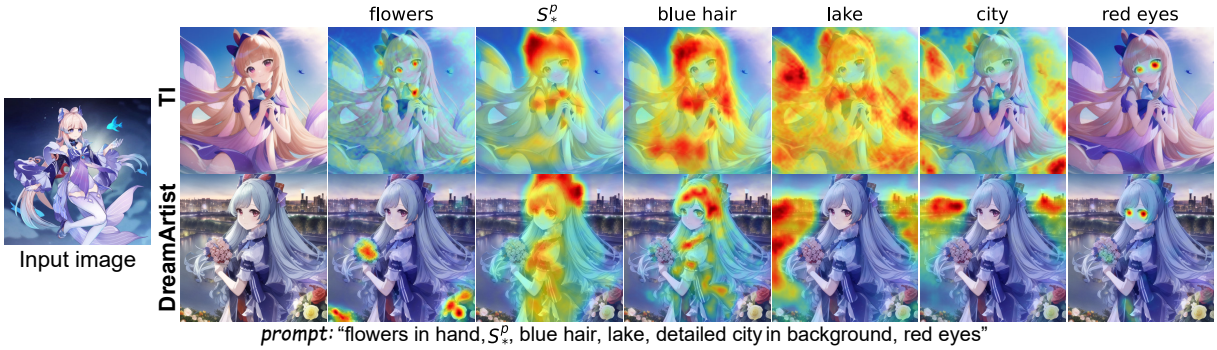

Fig. 2 Attention maps (DAAM [26]) of each token of the given text guidance in one-shot text-to-image generation. As prompt-tuning based methods, TI [27] and our DreamArtist learn pseudo-word S_* from the input image. In comparison to TI, DreamArtist demonstrates a superior ability to accurately render the visual content specified by the given text descriptions.

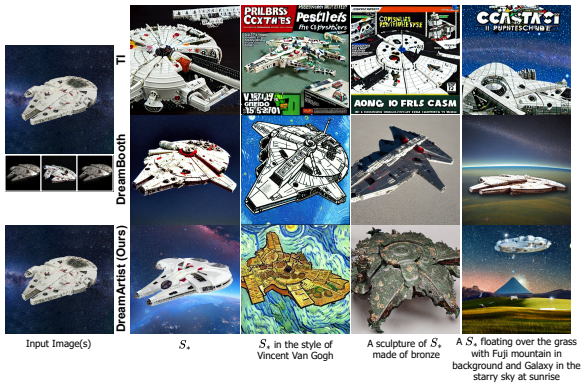

Fig. 3 Given a reference set with 3-5 images, TI [27] and DreamBooth [28] generate different images under different text guidance via learning a pseudo-word S_*. For comparison, given only **one** reference image, our DreamArtist can generate diverse images in different contexts, styles, materials, or others that are specified by given texts, presenting a better controllability.

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

Method	TI [27]	DreamBooth [28]	Ours
Given image number	3-5	3-5	1
Parameters	2K	983M	5K
Image quality	fair, mosaic	vivid	vivid
Diversity	limited	limited	highly diverse
Controllability	limited	limited	high

Second, we propose a general positive-negative adapter, facilitating diverse and highly controllable image generation from the positive guidance with the complementary negative rectification.

At last, we have conducted extensive qualitative and quantitative experiments on both natural and anime data, demonstrating that our method substantially outperforms existing methods in content quality, style similarity, diversity, and detail quality.

2 Related Work

2.1 Diffusion-Based Generative Models

In recent times, diffusion-based image generation models have achieved remarkable success. The first approach, known as the DDPM, proposed a method to iteratively denoise a noisy image and generate the image progressively. Compared to GAN-based methods, DDPM is more stable in training and generates more diverse images. However, the generation process of DDPM requires thousands of iterations, making it impractical to reality scenarios. To address this limitation, several algorithms such as DDIM, K-Diffusion, and DPM solver have been developed to accelerate the sampling and denoising process of DDPM.

Another aspect, [21] introduced a classifier guidance mechanism, which utilize the gradient of the log likelihood of an auxiliary classifier model $p_\theta(z_t|c)$ to guide the denoising process with condition c:

$$\tilde{\epsilon}_\theta\left(z_t, c\right) = \epsilon_\theta\left(z_t, c\right) - \gamma\sigma_t\nabla_{z_t}\log p_\theta\left(c \mid z_t\right). \quad (1)$$

It enhances the conditional controllability of the diffusion model's generation and, for the first time, surpasses the performance of GAN-based methods. Furthermore, [22] proposed the classifier-free guidance (CFG) method, enabling the diffusion model to simultaneously learn both conditional ($p_\theta(z_t|c)$) and unconditional ($p_\theta(z_t) = p_\theta(z_t|\emptyset)$) image generation. By constructing an implicit classifier $p_\theta(z_t|c) \propto \frac{p_\theta(z_t|c)}{p_\theta(z_t)}$, the performance of the model can be significantly enhanced.

2.2 Text-to-Image Synthesis

With the exponential evolution of generative models, the focus of research on Text-to-Image synthesis

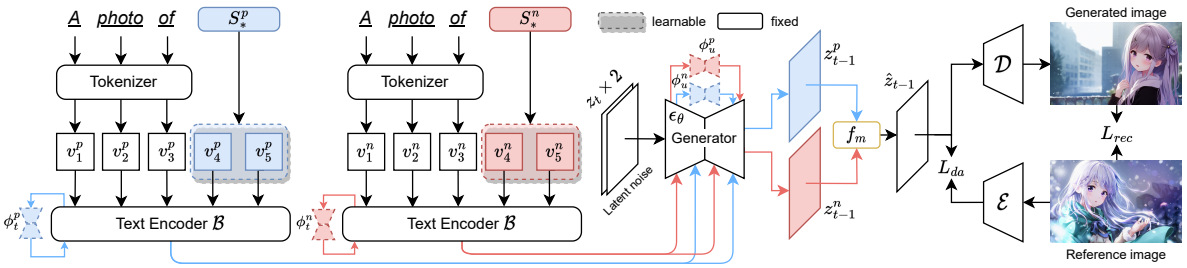

Fig. 4 Framework of our DreamArtist. Only the embeddings corresponding to positive and negative pseudo-words (S_*^p and S_*^n) need to be learned, and the rest of the parameters are fixed. $f_m = z^n + \gamma(z^p - z^n)$ is the classifier-free guidance of z^p and z^n.

has gradually shifted from GAN to Diffusion [12, 13, 14, 15, 16, 17]. Some large-scale text-to-image models [3, 23, 24, 2] have made highly accurate and fine-grained controllable semantic generation. The recently proposed stable diffusion [2] unprecedented making high-resolution and high-quality large-scale text-to-image models become reality. These Diffusion models usually employ classifier guidance [21] or classifier-free [22] guidance to generate images with text guiding. [32, 33] attempts to add spatial conditioning controls pretrained text-to-image diffusion models, while [34] employs both image and text as prompts to control the diffusion models with an additional cross attention. However, it is difficult for these methods to generate images following user-given patterns. And complex descriptions are required to generate high quality images.

2.3 Few-Shot Text-to-Image Generation

The diffusion-based Text-to-Image generation models have shown the ability to accurately and controllably generate high-quality images based on natural language descriptions. However, these models are not applicable to entirely novel concepts provided by users. Therefore, some attempts try to tune diffusion model with a small image set, enabling the model to guide the denoising process toward those specific features. [27] proposes the TI method trying to find a pseudo-words in the text vector space to represent the personalized object via prompt-tuning. [28], instead, proposes DreamBooth attempt to fine-tuning the entire model with a small image set under the premise of known personalized object categories. [29] employs a reparameterizable Adapter to fine-tune a small subset of model parameters, which can mitigate overfitting to some extent. [35] combining the advantages of DreamBooth and TI, trains only the k and v layer of the cross attention and introduces prompt-tuning together.

While these methods are able to learn the features of an object from few images, these methods suffer from overfitting and poor controllability. These methods still require 3-5 images, while our method requires only 1 image to generate highly controllable personalized features, which can be easily used with complex descriptions.

3 Preliminary

3.1 Latent Diffusion Model

With the remarkable capacity of image generation, Latent Diffusion Model (LDM) [2] is utilized as the base model. Different from DDPM [36, 5] that performs denoising operations in the image space, LDM conducts this in the feature space. This readily facilitates the diffusion operations in the feature space. Formally, firstly, an input image x is encoded into the feature space by an AutoEncoder (with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, *i.e.*, $z = \mathcal{E}(x), \hat{x} = \mathcal{D}(z)$), pretrained with a large number of images [2]. t indicates the time step, and z_t is the diffusion feature map of z at the t-th step. At the t-th denoising step, a Denoising U-Net ϵ_θ equipped with transformer blocks is used to perform denoising on the feature map $z_{t-1} = \epsilon_\theta(z_t, t)$. For text-guided conditional image generation, LDM utilizes a pre-trained text encoder $\mathcal{B}$ for given texts S and has its text feature $y = \mathcal{B}(S)$. It employs the cross-attention mechanism with the image feature as query and two transformations of the text feature as key and value. Its training loss is formulated as

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

where ϵ is the unscaled noise [2] and $\|\cdot\|_2^2$ is the $\ell 2$ loss. In this training phase, AutoEncoder is fixed and only ϵ_θ is learnable.

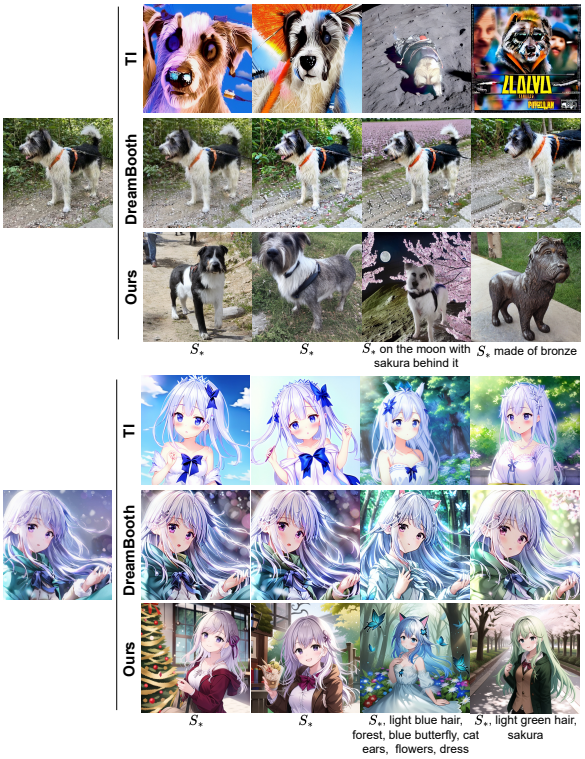

Fig. 5 Comparison with existing methods for one-shot text-to-image generation.

3.2 Deficiencies in Existing Methods

The majority of existing text-to-image Diffusion models, including LDM, incorporate the Conditional Fine-Grained (CFG) mechanism to enhance performance. More specifically, CFG utilizes an implicit classifier to perform guidance with $\frac{p(z_t|c)}{p(z_t)}$. However, the influence of this mechanism on the training process has been overlooked in previous few-shot text-to-image generation methods, leading to issues such as poor controllability or low-quality generated images.

The one/few-shot generation method learns from the given reference image(s) $w \sim p(w)$, where $p(w)$ is a distribution different from the distribution of the pre-training data $p(z)$. TI uses prompt tuning to learn a pseudo word S^p and maximizes log-likelihood of $p_\theta(w_t|S^p)$. With the fact that w and z will eventually diffuse to $\mathcal{N}(0, \mathcal{I})$ at timestamp T, we have w_T=z_T; then $p_\theta(w_{T-1}|w_T, S^p) = p_\theta(w_{T-1}|z_T, S^p)$. Accordingly, TI generates images with $\frac{p_\theta(w_t|S^p)}{p_\theta(z_t)}$. However, $p(w)$ and $p(z)$ are different distributions, TI only fits $p_\theta(w_t|S^p)$ while ignoring the differences between $p_\theta(w_t)$ and $p_\theta(z_t)$, leading to incorrect guidance directions and generating unexpected features and artifacts. Moreover, in the one-shot learning tasks, when $p(w)$ forms a single point, simply maximizing the log-likelihood of $p_\theta(w_t|S^p)$ can lead to severe overfitting and limited diversity. This problem is the same for fine-tuning-based methods like DB or Adapter-based methods like lora. These approaches aim to maximize the log-likelihood of $p_\theta(w_t|S^p)$ by learning θ, and similarly ignoring the distinctions between $p_\theta(w_t)$ and $p_\theta(z_t)$.

4 Methodology

To overcome the limitations of existing methods, our proposed DreamArtist framework, depicted in Fig. 4, synthesizes highly realistic and diverse images with enhanced controllability from a single user-provided image. Without bells and whistles, DreamArtist just introduces a novel approach that incorporates both positive and negative adapters. These adapters are concurrently optimized through positive-negative tuning, which remarkably improves generation quality, even within the challenging constraints of a one-shot scenario.

4.1 Positive-Negative Prompt-Tuning

In essential, conventional prompt-tuning [37, 38, 39, 40, 41, 42, 43, 44] optimistically considers only one prompt. Namely, it directly aligns it with the downstream task and learns a mapping from the prompt to the training set. However, based on our aforementioned analysis, applying these methods directly to the diffusion model may easily lead to collapse and overfitting, particularly when the training stage involves a limited samples. Especially, for one-shot text-to-image generation, this prompt-tuning has a remarkable overfitting to the reference image and generates images with limited diversity and even artifacts. Accordingly, we propose positive-negative prompt-tuning to address these problems, which disentangles the conventional prompt tuning into two components and learns to generate in a self-induced manner.

Specifically, based on the CFG mechanism, given a noise map z_t, it can be guided respectively by positive and negative text prompts and yields two different feature maps z^p_{t-1} and z^n_{t-1}. It is expected that z^p contains the desired characteristics of the reference image, while z^n contains the characteristics we prefer to be excluded from the generated image for rectification. Our approach involves the simultaneous learnable positive and negative pseudo-words (S^p_* and S^n_*) and maximizes the log-likelihood of $\frac{p_\theta(w_t|S^p_*)}{p_\theta(w_t|S^n_*)}$. Therefore,

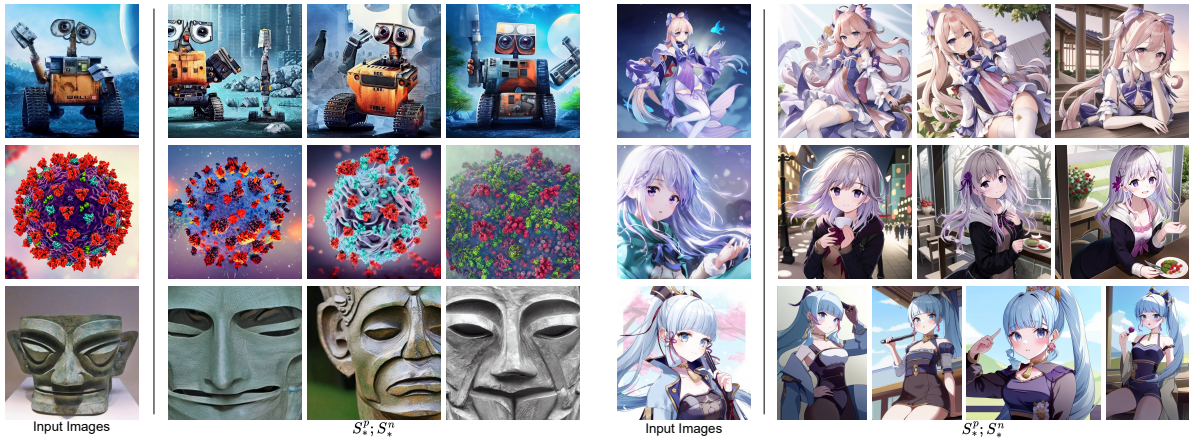

Fig. 6 One-shot text-to-image generation with only learned pseudo-words for DreamArtist. It can learn content and context from a single input image without adding additional text descriptions, generating diversity and high-quality images in both natural and anime scenes.

the loss function of DreamArtist is defined as follows:

$$\mathcal{L}_{da} = \|\epsilon - (\epsilon_\theta(z_t, S_*^n) + \gamma\,(\epsilon_\theta(z_t, S_*^p) - \epsilon_\theta(z_t, S_*^n)))\|^2 \tag{3}$$

DreamArtist learns S_*^p and S_*^n jointly, which ensures that the training and inference are guided in the same direction, thus eliminating artifacts and improving the quality of image generation and controllability of features. It avoids the problem of inaccurate feature generation or uncontrollable outcomes caused by incorrect guidance.

In Eq. (3), our essential optimization objective is to align the distribution $p(z_t|S_*^p)(\frac{p(z_t|S_*^p)}{p(z_t|S_*^n)})^\gamma = \frac{p(z_t|S_*^p)^{\gamma+1}}{p(z_t|S_*^n)^\gamma}$ with the distribution of reference image $\tilde{p}(z_t|c)$. Given that $p(z_t|S_*^p) < 1$ and $p(z_t|S_*^n) < 1$, we have:

$$\left(\frac{p(z_t|S_*^p)}{p(z_t|S_*^n)}\right)^\gamma \geq \frac{p(z_t|S_*^p)^{\gamma+1}}{p(z_t|S_*^n)^\gamma} \approx \tilde{p}(z_t|c). \tag{4}$$

When there is only one reference image, if only the positive branch is used, the optimization objective is to align the distribution $p(z_t|S_*^p)$ with the distribution $\tilde{p}(z_t|c)$. However, because there is only one reference image, $\tilde{p}(z_t|c) = 1$, which would lead to severe overfitting and cause the model to lose controllability. In our method, the positive branch $p(z_t|S_*^p)$ does not directly align with $\tilde{p}(z_t|c)$; instead, the constraint on the positive branch is relaxed through $p(z_t|S_*^n)$. Moreover, as γ increases, $\frac{p(z_t|S_*^p)}{p(z_t|S_*^n)}$ can become smaller, resulting in a greater degree of relaxation. Therefore, a larger γ will lead to a lower fitting and provide higher controllability.

4.2 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 for the generated image. In this context, we use approximate z_0 for calculating the loss through the noise $\hat{\epsilon}$, which is predicted by the model based on z_t through the CFG mechanism $\hat{\epsilon} = \epsilon_\theta(z_t, S_*^n) + \gamma\,(\epsilon_\theta(z_t, S_*^p) - \epsilon_\theta(z_t, S_*^n))$. Then the approximate z_0 is:

$$z_{t\to 0} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_t - \hat{\epsilon}\sqrt{1-\bar{\alpha}_t}\right), \tag{5}$$

Accordingly, the reconstruction loss can be expressed as:

$$\mathcal{L}_{rec} = \|\mathcal{D}(z_{t\to 0}) - x\|, \tag{6}$$

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

4.3 Positive-Negative Adapter

Due to the high dimensionality of image data, which differs from the characteristics of natural language, and the fact that humans are sensitive to the perception of details in images. Relying solely on prompt tuning for learning fine-grained features still falls short. It often struggles to faithfully reproduce the detailed features of entities present in the reference images, leading some individuals to subjectively perceive discrepancies. Additionally, limitations in controllability may also arise from insufficient comprehension of features, resulting in associations errors or difficulties in controlling details (e.g., inaccurate modification of accessories or altering hair color affecting eyes color). Moreover, if

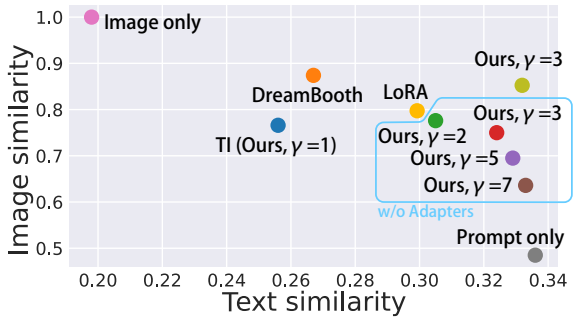

Fig. 7 Comparison of our DreamAritist with different values of γ and existing methods, following the same evaluation method to TI.

Table 2 Quantitative comparison of our DreamArtist with existing methods for one-shot text-to-image generation[1].

Method	LPIPS↓	Style loss↓	CDS↑	CFV↑
Natural Image Generation				
TI	0.71	6.12	0.73	**1.79**
DreamBooth	**0.33**	1.31	0.63	0.69
LoRA	0.47	1.60	0.48	0.89
Ours(DreamArtist w/o Adapter)	0.62	2.46	0.74	1.53
Ours(DreamArtist)	0.46	**1.29**	**0.75**	1.46
Anime Image Generation				
TI	0.63	7.47	0.41	0.87
DreamBooth	0.49	1.16	0.33	0.72
LoRA	**0.48**	0.71	0.22	0.66
Ours(DreamArtist w/o Adapter)	0.60	0.69	**0.60**	**1.28**
Ours(DreamArtist)	**0.48**	**0.38**	0.59	1.03

the distribution of features in the reference images significantly deviates from those in the pre-trained model, the model may struggle to learn effectively and fail to adequately characterize these features.

In response to these challenges, we propose the LoRA-based Positive-Negative Adapter method, which incorporates the Adapter. The Generator part's Adapters (ϕ_u^p and ϕ_u^n) serve a dual purpose: enhancing the model's ability to preserve fine-grained details from the reference images and enabling the model to learn features more faithful that are absent in the pre-trained image data. Adapters in the Text Encoder part (ϕ_t^p and ϕ_t^n) further elevate the model's controllability, allowing the features in the reference images to be precisely and finely controlled through textual descriptions. Our DreamArtist method combines the Positive-Negative Adapter approach with prompt tuning, enabling the model to learn the generation of images with high controllability, strong diversity, intricate details, and high quality through only one reference image as input.

Table 3 Quantitative analysis of our DreamArtist compared with existing methods on feature controllability. The feature controllability of DreamArtist substantially exceeds existing methods[1].

Method	Natural Image Generation				Anime Image Generation			
	CAS↑	CFV↑	Style loss↓	CDS↑	CAS↑	CFV↑	Style loss↓	CDS↑
TI	0.37	1.46	4.31	0.40	0.23	0.98	5.52	0.46
DreamBooth	0.24	1.19	**0.40**	**0.69**	0.28	0.81	1.28	0.31
LoRA	0.30	1.10	1.02	0.53	0.37	0.80	1.03	0.13
Ours(DreamArtist w/o Adapter)	0.83	**1.55**	2.01	0.57	0.63	**1.15**	2.71	**0.58**
Ours(DreamArtist)	**0.89**	1.43	0.98	0.58	**0.65**	1.07	**0.91**	0.57

Table 4 Evaluation on γ and $\mathcal{L}_{rec}$[1].

Method	LPIPS↓	Style loss↓	CDS↑	CFV↑	CAS↑
γ=3; w/o $\mathcal{L}_{rec}$	0.629	1.41	0.68	1.33	0.76
γ=2	**0.601**	**1.13**	0.59	1.02	0.67
γ=3	0.613	1.17	**0.72**	1.41	0.79
γ=5	0.640	1.85	0.48	**1.66**	0.87
γ=7	0.653	1.44	0.67	1.03	**0.92**

5 Experiments

5.1 Experimental Settings

Dataset. Following the existing experimental settings [27, 28], the LAION-2B dataset [30] is used for natural image generation. Additionally, we add an anime dataset, Danbooru [31] for the popular interest in many applications, *e.g.*, games and animes.

Implementation details. In DreamArtist, the learning rate is 0.0025 and γ is 3 (5 for style cloning). It is trained on one RTX2080ti using a batch size of 1 with about 2k-8k iterations. In TI, the length of an embedding for a pseudo-word is set to equal that of 6 words, while DreamArtist uses 3 words for both positive and negative embeddings. Positive prompt is initialized using similar words similar to TI (initialized with some similar words), while negative prompt is initialized using EOS token (referring to empty text $\emptyset$) with random noise. In the CFG configuration, the negative prompt is by default set to empty text. The negative prompt is derived from $p(z_t)$, so we use empty text with a small noise to initialize the negative embedding.

Metrics. 1) For quantitative evaluation, we follow the same evaluation metrics to TI, namely calculating *image similarity* to the **given** reference images and *text similarity* to the **given** texts. But, this presents a **measure bias** to the overfitting model to the **given** reference images and texts, especially for one/few-shot generation (Sec. 5.2). **2)** Accordingly, we also adopt different metrics from three aspects[2]: image/style similarity (LPIPS [45]/style loss [46]), image diversity (CLIP

[1] Values in **bold** are the best results and those in blue are the second best.
[2] More details on CDS, CFV and CAS are described in Supplementary.

detail score, CDS; CLIP image feature variance, CFV), and generation controllability (CLIP average score, CAS). *Notably, there is a tradeoff between the image diversity and similarity to given reference images.* High **similarity** indicates a potential overfitting bias and a limited ability of generating **diverse** images, and vice versa.

CDS employs the CLIP model to evaluate the richness of the details of the generated image content. More precisely, it represents the probability of the image being categorized as "detailed" within the set of [little detail", detailed"] using CLIP.

CFV is used to evaluate the diversity of generated images. It calculates the standard deviation of the feature maps of the generated images. Those feature maps are encoded through the image encoder of CLIP.

CAS is used to evaluate the generation controllability, to check whether additional text descriptions are accurately rendered in the generated image. Firstly, we extract all noun phrases from the text descriptions. Secondly, each of these noun phrases is fed into the CLIP model together with a randomly selected set of noun phrases from the database. Notably, those selected noun phrases do not exist in the text descriptions. Then, the average of the probabilities that CLIP classify the generated image to the noun phrases in the descriptions is CAS.

5.2 One-Shot Text-guided Image Synthesis

We compare our DreamArtist with two existing works, including TI [27], DreamBooth [28] and LoRA [29] for one-shot text-to-image generation[3]. All methods are trained with only one image given as reference for a fair comparison in all the experiments. Next, we will elaborate the comparison results from image similarity and diversity, generation controllability, and style cloning.

Image Similarity and Diversity. Qualitatively, Fig. 5 shows that images generated by TI have limited diversity and details. DreamBooth generates images with fairly quality, but the diversity is also limited for both natural and anime image generation. It generates images overly similar to the reference image, which evidences an over-fitting issue. From Fig. 5 and 6, our DreamArtist can alleviate these problems and not only generates highly realistic images with more promising and reasonable details, but also keeps the generated images highly diverse, *e.g.*, dogs in different contexts (Fig. 5) and WALL-Es with difference appearances (Fig. 6).

Quantitatively, we first follow the same evaluation method of TI and provide the comparison in Fig. 7. Our DreamArtist (Ours, $\gamma = 3$) has a superior balance of image generation between image similarity and text similarity, in comparison with TI and DreamBooth. This is also demonstrated in Tab. 2 and Tab. 3. Though DreamBooth and LoRA performs better on LPIPS and style loss due to over-fitting, it has a poor diversity (CFV). TI has a higher CFV for the diversity, but suffers from severe artifacts that make an illusion of high diversity metrics, as demonstrated in Fig. 5.

Generation Controllability. As shown in Fig. 5, TI has a limited generation controllability that it fails to render some of other words, *e.g.* green hair. Text similarity in Fig. 7 and CAS in Tab. 3 also demonstrate that. Although DreamBooth can render some additional texts, the generated images are too homogeneous in structure and extremely poor in diversity. This also indicates the limited controllability of DreamBooth. In contrast, DreamArtist can keep high controllability and diversity while maintaining sufficient similarity to the reference image. Besides, from CAS in Tab. 3, DreamArtist also performs best with a significant improvement (0.89 [DreamArtist] vs. 0.24/0.37 [DreamBooth/TI]).

Style Cloning. In Fig. 9, it is observed that DreamArtist can learn different image styles, *e.g.*, Chinese brush painting, paper-cut painting and Cyberpunk. In anime cases, 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 audience a 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 very promising for the creation of anime work. Besides, DreamArtist also manages to generate different game maps that seem reasonable in Fig. 9.

6 Method Evaluation and Analysis

6.1 Evaluation on γ

γ controls a generation trade-off between the image similarity and controllability in comparison with the

[3]Their generated images given 3-5 images are shown in Fig. 3 and more results are provided in Supplementary.

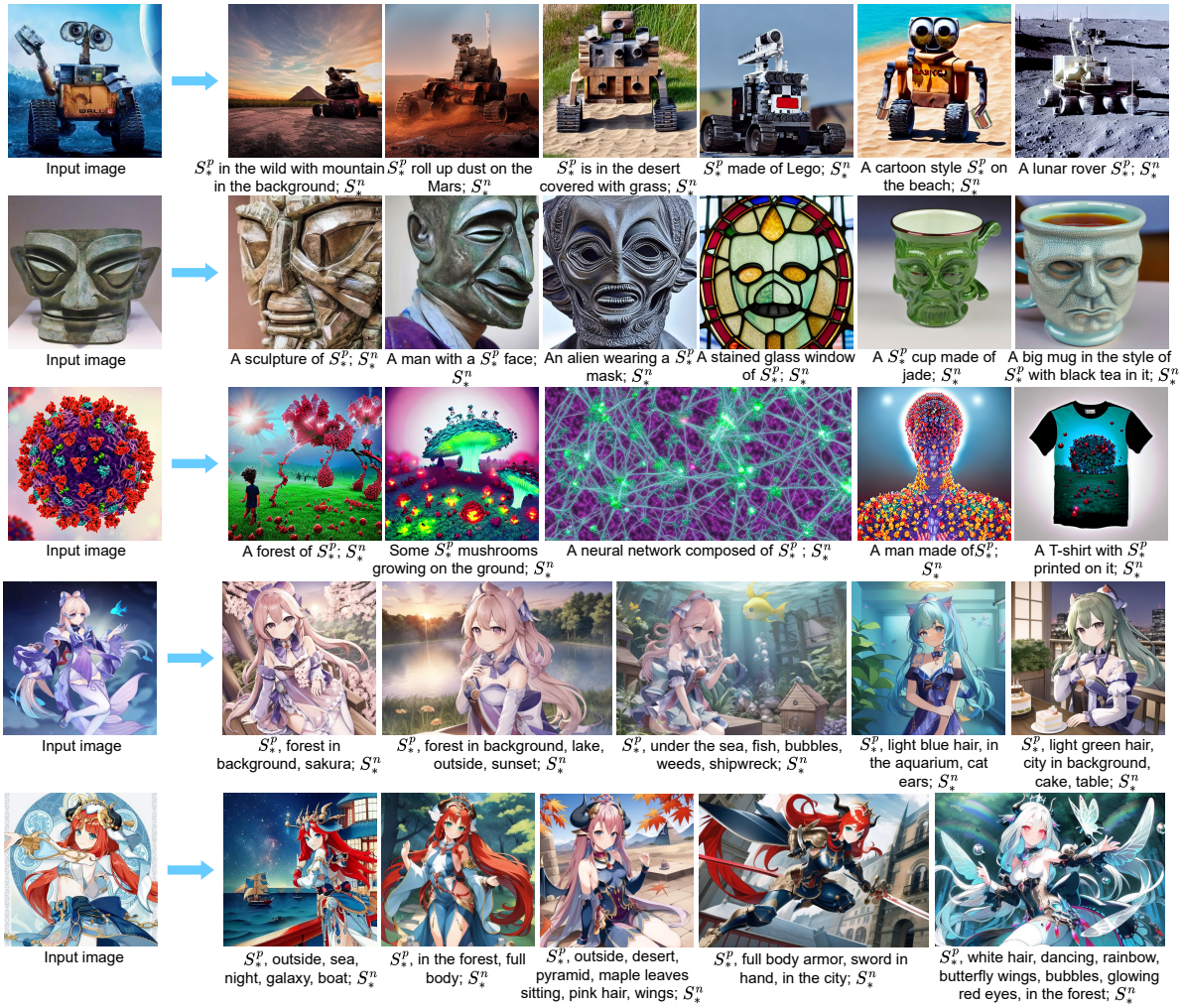

Fig. 8 One-shot text-to-image generation with the guidance of additional complex texts for DreamArtist. DreamArtist exhibits a superior capability of controllable generation: even with few words in the text guidance, diverse and faithful images are generated; with more words, vivid images with rich details are generated. More importantly, DreamArtist can successfully render almost all the given words.

reference image. In Tab. 4, the smaller the γ, the better DreamArtist performs for its LPIPS and style loss. This indicates the generated images are more faithful to the input image, but also underlies the inferior generation controllability and image diversity. This is demonstrated by its lower performance of CFV and CAS. On the contrary, the smaller the γ, the better DreamArtist performs for its CFV and CAS. This verifies the improvement of its generation controllability and image diversity. In our paper, we use $\gamma = 3$ to report empirical results ($\gamma = 5$ for learning image styles).

6.2 Ablative Study

DreamArtist is very simple but effective. Its main component is positive-negative adapter, without which TI can be regarded as its base model. Thus its ablation study is identical to the comparison of TI and DreamArtist, discussed in Sec. 5.2. Moreover, we also provide the ablation study on the reconstruction constraint (Sec. 4.2) in Tab. 4 and Fig. 13. We observe that, without it, the model has difficulty learning low-level features such as color and details (*e.g.*, the background color of WALL-E and the hair color of some anime characters. The bird's tail color is wrongly diffused to the body, which is different from the reference image.) when using only the feature space loss.

Ablation on Positive-Negative Adapters We also compared the performance between DreamArtist without Adapters, which utilizes only prompt tuning, and DreamArtist, which incorporates with adapters. As

Fig. 9 Style cloning via DreamArtist, for example, styles of wash painting, paper-cut art, Cyberpunk, comic of caricaturists, and road map in a game (from left to right).

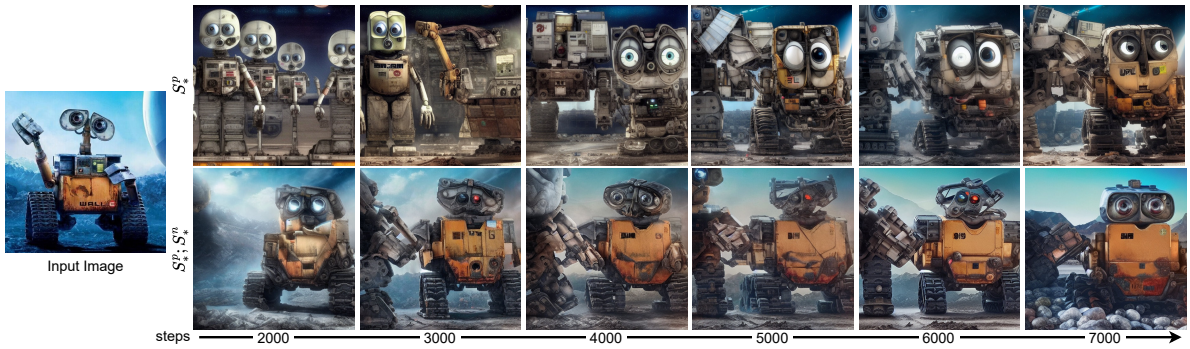

Fig. 10 Generation results via only S_*^p or (S_*^p;S_*^n) at different steps. UP: using S_*^p alone. Down: using an S_*^p after 7000 steps together with S_*^n from different steps. It is observed that S_*^n rectifies the generation based on S_*^p to improve the image quality.

depicted in Fig. 12, DreamArtist exhibits a significant improvement in its ability to learn fine details from the reference image and preserve the primary features of the reference image well. Furthermore, both image quality and feature controllability experience a noticeable enhancement. For instance, modifications made to features such as hair color do not affect other elements like clothing. Despite achieving better preservation of reference image features, our DreamArtist does not demonstrate signs of overfitting, maintaining commendable diversity and controllability in the generated images. The visual attributes and characteristics of the entities depicted in the images remain controllable through textual descriptions.

6.3 Analysis on Negative Branch

To further explain the mechanisms of S_*^p and S_*^n, we separately visualize the images generated by only S_*^p at different steps, and the images generated by a fixed S_*^p (trained for 7000 steps) together with S_*^n at different steps. In Fig. 10, it is observed that S_*^p steadily learns the salient features of the reference image. But the quality of the images generated by S_*^p is not so satisfactory, *e.g.*, lacking in style and details. When resorting to the help of S_*^n, the model rectifies deficiencies of S_*^p and progressively improves the image quantity. This indicates the effectiveness of PNPT.

Analysis on Controllability and Compatibility. As discussed in Sec. 5.2 and demonstrated in Fig. 2, it is

Fig. 11 The visualization results of training with different values of γ.

Fig. 12 Visualization results DreamArtist with LoRA. Compared to DreamArtist, which relies only on prompt tuning only, the detail and faithful of the generated image to the reference image is significantly improved by adding an adapter.

challenging that the learned pseudo-words of TI harmoniously combine with other words. Namely, TI would ignore some of words in the generated images. DreamBooth has slightly better compatibility but suffers from overfitting of image content and thus has inferior image diversity. Our method, as an alternative, can effectively address these issues. DreamArtist is highly compatible with complex descriptions and can generate diverse and harmonious images using learned features. The learned embedding, for instance, a mask in the second row, can produce highly realistic and diverse images based on various complex descriptions. Our method can also handle conflicts between the additional description and the learned features, as demonstrated in the fourth row of Fig. 8, where DreamArtist generates characters with light green hair and a city background despite the training image having pink hair and a pure background.

6.4 Analysis on Concept Decoupling

DreamArtist can learn not only the entities in the reference image, but also many different characteristics (concepts) such as light, material, and style (in Fig. 17). Namely, the learned pseudo words with adapters in different text contexts can be rendered with different desired characteristics of the reference image. This indicates the implicit decoupling of characteristics or concepts for promising generation controllability.

Fig. 13 Visualization results of ablation study on L_{rec}.

Fig. 14 Generation controllability evaluation. Given different texts on clothing styles, our DreamArtist can produce an image of a girl with different styles of clothes.

6.5 Human Evaluation

To demonstrate that our method can synthesize high-quality realistic images, we have conducted a user study on TI and our methods following the rules of the Turing test from 700 participants, respectively. For testing one method, its 12 generated images and other 8 real images are provided to participants, to select which images are real, not generated. TI achieves a failure rate of 26.6%. Instead, our method is 34.5%, which significantly exceeds 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, to evaluate which creation has higher quality, 52.31% and 83.13% of participants from various walks of life (even including professional anime

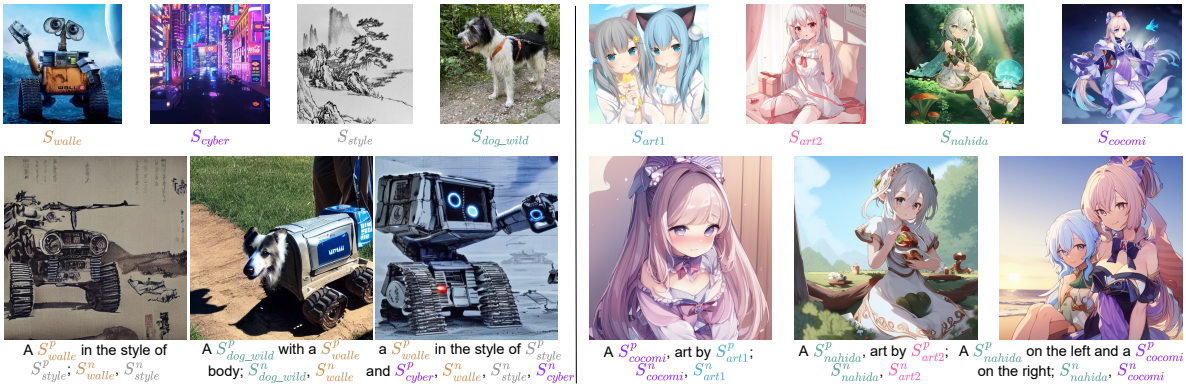

Fig. 15 Results of concept compositions via DreamArtist. It presents a promising generation potential via the text guidance from the combination of the learned pseudo-words.

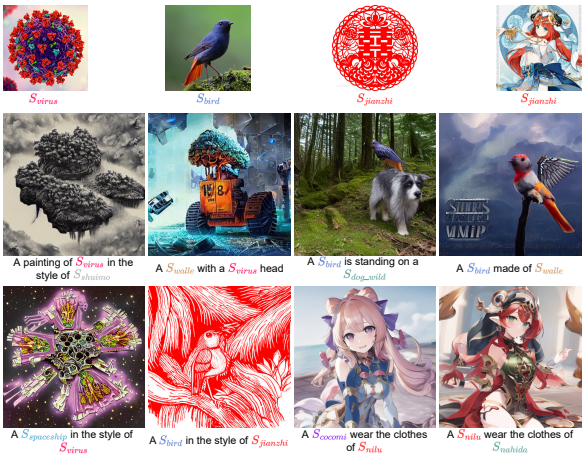

Fig. 16 More results of concept compositions via DreamArtist.

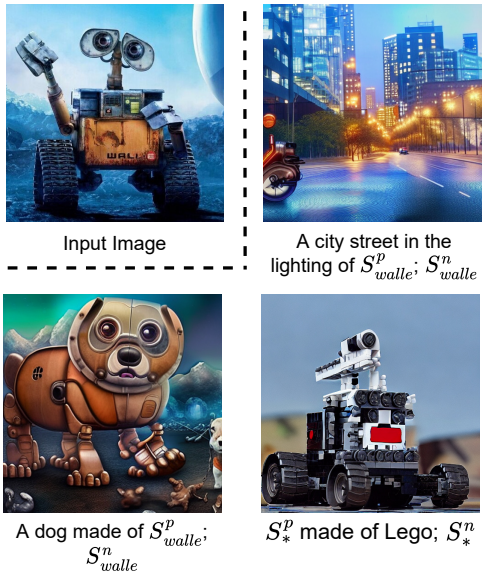

Fig. 17 Visualization results of concept decoupling. DreamArtist can apply abstract semantic concepts from reference images to a given context.

artists) prefer the synthesized images of DreamArtist for natural image cases or anime cases, respectively.

6.6 Analysis on Embedding Length

As illustrated in Fig. 18, on one hand, S^p_* with a length of 1 captures only coarse features, which limits the model's ability to learn finer details such as color and texture. Increasing the S^p_* length to 3 achieves a more effective balance between the learned visual features and the textual description, leading to improved results. However, further extending the S^p_* length beyond this point induces overfitting, causing the model to overemphasize features from the reference image. This, in turn, compromises the alignment between the generated images and the textual descriptions, while also introducing visual artifacts.

On the other hand, increasing the S^n_* length generally enhances image quality, particularly in terms of texture refinement, lighting effects, and detail preservation. Nevertheless, excessively long S^n_* can also degrade image quality. Based on these empirical observations, we identify an optimal configuration of 3 for the S^p_* length and either 3 or 6 for the S^n_* length. To maintain consistency and simplicity, we adopt a length of 3 for both positive and negative embeddings in our work.

6.7 Extended Task: 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 styles, for

Fig. 18 Visualization results of DreamArtist with different positive and negative embedding length.

generating reasonable images. When combining these pseudo-words, it is necessary to add their learned embeddings of the positive and negative prompts. As illustrated in Fig. 15, combining multiple pseudo-words (highlighted in different colors) 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. 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 [27].

7 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 new characteristics. Existing methods not only require 3-5 reference images, but also suffer from over-fitting that is adverse to the image diversity and generation controllability. To mitigate this issue, we propose a simple but effective method, named DreamArtist, without bells and whistles. It employs a learning strategy of positive-negative prompt-tuning, enabling the model to learn and rectify the generation results. The learned pseudo-words can not only make the model generate high-quality and diverse

images, but also can be easily controlled by additional text descriptions. Extensive qualitative and quantitative experimental analyses have demonstrated that our method substantially outperforms existing methods. Moreover, our DreamArtist method is highly controllable and can be used in combination with complex descriptions, presenting a promising flexibility and potential for deploying other models.

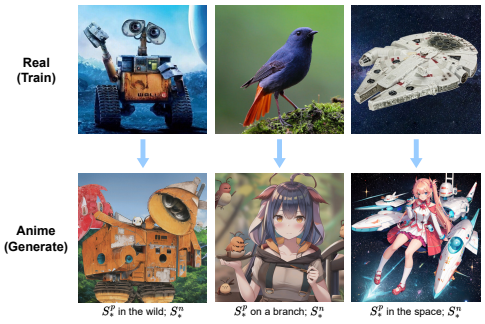

Fig. 19 Limitations of our method with domain shit.

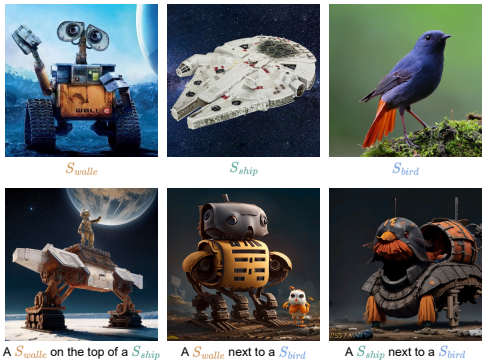

Fig. 20 Limitations of our method in entity concept composition.

8 Limitations and future work.

Domain Shift: Our DreamArtist, when trained on a base model within the real-world domain, struggles to accurately render its learned features upon application to a base model in the anime domain, as shown in Fig. 19. Only partial of the learned features can be effectively rendered, and the generated outputs are heavily influenced by the biases inherent in the anime domain's base model.

Compositions of Entity: Our DreamArtist is applicable to concept compositions by combining learned pseudo-words for promising and flexible image generation. However, it sometimes fails on compositions of entity concepts, since they are individually learned from different reference images and may interference with each other. When combining two entity concepts, the model may struggle to accurately render them as distinct entities. As shown in Fig. 20,this often leads to a fusion of their respective features, resulting in mutual interference. The model sometimes fails to distinctly map each concept to its corresponding independent entity.

Future Work: Thus, this issue can be solved by continual text-to-image generation with continual learning methods, to welcome more and more new concepts and ensure highly controllable generation from learned words and original words. Besides, domain shift between reference images and original images for pre-training would cause generation failure, which is another issue to address in future work.

References

[1] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S.K.S. Ghasemipour, B.K. Ayan, S.S. Mahdavi, R.G. Lopes, T. Salimans, J. Ho, D.J. Fleet, M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding. CoRR **abs/2205.11487** (2022). 2205.11487

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10674–10685

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with CLIP latents. CoRR **abs/2204.06125** (2022). 2204.06125

[4] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, *Alias-Free Generative Adversarial Networks*, in *Advances in Neural Information Processing Systems* (2021), pp. 852–863

[5] J. Song, C. Meng, S. Ermon, *Denoising Diffusion Implicit Models*, in *International Conference on Learning Representations* (2021)

[6] D.P. Kingma, P. Dhariwal, *Glow: Generative Flow with Invertible 1x1 Convolutions*, in *Advances in Neural Information Processing Systems* (2018), pp. 10236–10245

[7] A. Brock, J. Donahue, K. Simonyan, *Large Scale GAN Training for High Fidelity Natural Image Synthesis*, in *International Conference on Learning Representations* (2019)

[8] C. Saharia, W. Chan, H. Chang, C.A. Lee, J. Ho, T. Salimans, D.J. Fleet, M. Norouzi, *Palette: Image-to-Image Diffusion Models*, in *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference* (2022), pp. 15:1–15:10

[9] P. Dhariwal, A.Q. Nichol, *Diffusion Models Beat GANs on Image Synthesis*, in *Advances in Neural Information Processing Systems* (2021), pp. 8780–8794

[10] O. Ronneberger, P. Fischer, T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, in *Medical Image Computing and Computer-Assisted Intervention, MICCAI*, vol. 9351 (2015), pp. 234–241

[11] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole, *Score-Based Generative Modeling through Stochastic Differential Equations*, in *International Conference on Learning Representations* (2021)

[12] A.Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*, in *International Conference on Machine Learning* (2022), pp. 16784–16804

[13] J. Cheng, F. Wu, Y. Tian, L. Wang, D. Tao, *RiFeGAN: Rich Feature Generation for Text-to-Image Synthesis From Prior Knowledge*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 10908–10917

[14] A. Jain, B. Mildenhall, J.T. Barron, P. Abbeel, B. Poole, *Zero-Shot Text-Guided Object Generation with Dream Fields*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 857–866

[15] B. Li, X. Qi, T. Lukasiewicz, P.H.S. Torr, *Controllable Text-to-Image Generation*, in *Advances in Neural Information Processing Systems* (2019), pp. 2063–2073

[16] T. Qiao, J. Zhang, D. Xu, D. Tao, *MirrorGAN: Learning Text-To-Image Generation by Redescription*, in *IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 1505–1514

[17] M. Tao, H. Tang, S. Wu, N. Sebe, F. Wu, X. Jing, DF-GAN: deep fusion generative adversarial networks for text-to-image synthesis. CoRR **abs/2008.05865** (2020). 2008.05865

[18] S.E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, *Generative Adversarial Text to Image Synthesis*, in *Proceedings of the International Conference on Machine Learning*, vol. 48 (2016), pp. 1060–1069

[19] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, E. Raff, *VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance*, in *ECCV* (2022), pp. 88–105

[20] Y. Liu, J. Peng, J.J.Q. Yu, Y. Wu, *PPGAN: Privacy-Preserving Generative Adversarial Network*, in *IEEE International Conference on Parallel and Distributed Systems* (2019), pp. 985–989

[21] X. Liu, D.H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, T. Darrell, More control for free! image synthesis with semantic diffusion guidance. CoRR **abs/2112.05744** (2021). 2112.05744

[22] J. Ho, T. Salimans, *Classifier-Free Diffusion Guidance*, in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications* (2021)

[23] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, J. Tang, *CogView: Mastering Text-to-Image Generation via Transformers*, in *Advances in Neural Information Processing Systems* (2021), pp. 19822–19835

[24] J. Yu, Y. Xu, J.Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B.K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, Y. Wu, Scaling autoregressive models for content-rich text-to-image generation. CoRR **abs/2206.10789** (2022). 2206.10789

[25] H. Zhang, W. Yin, Y. Fang, L. Li, B. Duan, Z. Wu, Y. Sun, H. Tian, H. Wu, H. Wang, Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. CoRR **abs/2112.15283** (2021). 2112.15283

[26] R. Tang, A. Pandey, Z. Jiang, G. Yang, K. Kumar, J. Lin, F. Ture, What the DAAM: interpreting stable diffusion using cross attention. CoRR **abs/2210.04885** (2022). 2210.04885

[27] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A.H. Bermano, G. Chechik, D. Cohen-Or, *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion*, in *International Conference on Learning Representations* (2023)

[28] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. CoRR **abs/2208.12242** (2022). 2208.12242

[29] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *LoRA: Low-Rank Adaptation of Large Language Models*, in *ICLR The Tenth International Conference on Learning Representations* (2022)

[30] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, J. Jitsev, LAION-5B: an open large-scale dataset for training next generation image-text models. CoRR **abs/2210.08402** (2022). 2210.08402

[31] Anonymous, D. community, G. Branwen. Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset (2022). URL https://www.gwern.net/Danbooru2021

[32] L. Zhang, A. Rao, M. Agrawala, *Adding Conditional Control to Text-to-Image Diffusion Models*, in *IEEE/CVF International Conference on Computer Vision* (2023), pp. 3813–3824

[33] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, *T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models*, in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024* (2024), pp. 4296–4304

[34] H. Ye, J. Zhang, S. Liu, X. Han, W. Yang, Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. CoRR **abs/2308.06721** (2023). 2308.06721

[35] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, J.Y. Zhu, *Multi-Concept Customization of Text-to-Image Diffusion*, in *IEEE Conference on Computer Vision and Pattern Recognition* (2023)

[36] J. Ho, A. Jain, P. Abbeel, *Denoising Diffusion Probabilistic Models*, in *Advances in Neural Information Processing Systems* (2020)

[37] X. Liu, K. Ji, Y. Fu, Z. Du, Z. Yang, J. Tang, P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. CoRR **abs/2110.07602** (2021). 2110.07602

[38] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, *Language Models are Few-Shot Learners*, in *Advances in Neural Information Processing Systems* (2020)

[39] T. Schick, H. 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, EACL* (2021), pp. 255–269

[40] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models. Int. J. Comput. Vis. **130**(9), 2337–2348 (2022)

[41] K. Zhou, J. Yang, C.C. Loy, Z. Liu, *Conditional Prompt Learning for Vision-Language Models*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 16795–16804

[42] K. Ding, Y. Wang, P. Liu, Q. Yu, H. Zhang, S. Xiang, C. Pan, Prompt tuning with soft context sharing for vision-language models. CoRR **abs/2208.13474** (2022). 2208.13474

[43] B. Lester, R. Al-Rfou, N. Constant, *The Power of Scale for Parameter-Efficient Prompt Tuning*, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP* (2021), pp. 3045–3059

[44] X.L. Li, P. 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), pp. 4582–4597

[45] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*, in *IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 586–595

[46] L.A. Gatys, A.S. Ecker, M. Bethge, *Image Style Transfer Using Convolutional Neural Networks*, in *IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2414–2423