Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Correspondence:

Multistage guidance on the diffusion model inspired by human artists' creative thinking

Wang QI[§1], Huanghuang DENG[§2], Taihao LI[‡1]

[1]*AI Research Institute, Zhejiang Lab, Hangzhou 311121, China*

[2]*Department of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

E-mail: qiwang@zhejianglab.com; dhh2012@zju.edu.cn; lith@zhejianglab.com

Received Apr. 30, 2023; Revision accepted Oct. 13, 2023; Crosschecked Dec. 12, 2023; Published online Dec. 27, 2023

https://doi.org/10.1631/FITEE.2300313

Current research on text-conditional image generation shows parallel performance with ordinary painters but still has much room for improvement when compared to that of artist-ability paintings, which usually represent multilevel semantics by gathering features of multiple objects into one object. In a preliminary experiment, we confirm this and then seek the opinions of three groups of individuals with varying levels of art appreciation ability to determine the distinctions that exist between painters and artists. We then use these opinions to improve an artificial intelligence (AI) painting system from painter-level image generation toward artistic-level image generation. Specifically, we propose a multistage text-conditioned approach without any further pretraining to help the diffusion model (DM) move toward multilevel semantic representation in a generated image. Both machine and manual evaluations of the main experiment verify the effectiveness of our approach. In addition, different from previous one-stage guidance, our method is able to control the extent to which features of an object are represented in a painting by controlling guiding steps between the different stages.

1 Introduction

AI-generated content (AIGC) techniques have become popular due to their great application prospects and rapid development. As a basic research direction with multimodal attributes, text-to-image generation has attracted attention from both industry and academia in natural language and computer vision. Through large DMs trained on billion-level text–image pairs, such as Disco Diffusion and Stable Diffusion, the current deep learning technique is able to generate near-realistic images, given short and uncomplicated descriptive text.

Numerous studies have concentrated on text-to-image generation and made significant advancements in this area. Scaling up likelihood-based models, which may contain billions of parameters in autoregressive (AR) transformers, currently rules the analysis of natural landscapes (Razavi et al., 2019; Ramesh et al., 2021). In comparison, it has been found that generative adversarial networks (GANs) offer promising findings (Brock et al., 2019; Goodfellow et al., 2020; Karras et al., 2021) that are largely restricted to data with comparably low levels of variability because modeling complicated, multimodal distributions requires a more scalable adversarial learning process. The state of the art in text-conditional image generation has recently been defined by DMs (Sohl-Dickstein et al., 2015), which

§ These two authors contributed equally to this work
‡ Corresponding author
ORCID: Taihao LI, https://orcid.org/0000-0003-3279-7125

are constructed from a hierarchy of denoising autoencoders and can produce remarkable outcomes in image generation (Ho et al., 2020; Song Y et al., 2021) and beyond (Chen N et al., 2021; Kingma et al., 2021; Kong et al., 2021; Mittal et al., 2021).

Benefiting from the progress made by previous work, recent AI painting systems show performance comparable to that of human painters. However, as our preliminary experiment shows, there still exists a large gap when compared to that of human artists. We then solicit opinions from three groups of people with different levels of art appreciation ability to determine what main or important differences exist between ordinary painters and artists, so the AI painting system can be improved. We conclude two points from the perspective of artistic techniques. First, the work of artists usually contains multiple semantic levels, ranging from surface-level conceptual objects to deep-level emotion that users want to express by generated images. Second, the work of artists usually combines various objects or endows an object with the characteristics of other objects according to their own creative ideas. Both of these factors make it difficult for recent techniques to consider both levels from only one-stage text guidance.

Aiming to improve current image generation techniques toward the painting ability of human artists, we propose a multistage text-conditioned approach using a DM. Specifically, our contributions can be summarized as follows:

1. We reveal two recognized differences between current AI painting systems and human artists by soliciting the opinions of people with different levels of artistic experience.

2. We propose a multistage guidance text-conditioned approach for working on a DM without any further pretraining to help current AI painting systems approach human-artist-level painting. Both manual evaluation and machine metric evaluation verify the effectiveness of our approach.

3. Different from previous one-stage guidance, by tuning guiding steps in each stage, our method is able to control the extent to which features of an object are represented in a painting.

2 Preliminary experiment

We first constructed two datasets: AI&Painter and AI&Artist. For the former, we created 50 four-choice questions where three of these choices are images generated by three state-of-the-art (SOTA) text-to-image neural models, including DALL-E 2 (Ramesh et al., 2022), Stable Diffusion (Rombach et al., 2022), and Midjourney V3, and one of them is an image painted by ordinary painters. We obtained the latter by substituting an image drawn by artists for that drawn by human painters in each item of the AI&Painter dataset. Specifically, we first selected five different styles of artist images, including oil painting, wash painting, watercolor painting, freehand brushwork painting, and claborate-style painting, with approximately 10 images for each style. We manually annotated captions for these images, and then input them into the AI painting model to generate corresponding images. Then, we invited 20 persons without art experience and 13 art practitioners to judge which image in each question was better in their view, and then calculated the percentage distribution of approval for each model and human creator. Table 1 shows the results. We can conclude that the current AI painting system is comparable to ordinary painters in painting ability, but it is still difficult for the current AI painting system to generate images of a similar artist level.

Based on the above results, to explore the gap that exists between ordinary painters and painting artists/masters in painting artistic works, we solicited opinions of six invited artists about the question. After that, we invited the aforementioned 13 practitioners and 20 persons without art experience to vote for those six artists' opinions. Table 2 shows their corresponding answers and voting results.

A common argument that can be concluded from Table 2 is that all invited artists stressed the expression of deep thoughts in painting, not only painting skills. In both the voting results of art practitioners and the public, the deep meaning (A2)/leading thinking (A3) of painting artists was most favored when distinguished from ordinary painters, while the integration of deep level emotion/thinking/culture elements and visible entities in art works (A1, A4, A5, and A6) was also supported.

The above evidence motivated us to design a method of multistage guidance for a DM that focuses on artistic image generation. First, the proposed method accepts multiple stages of text guidance, which is consistent with the multilevel semantic representation of artistic work to enable users

Table 1 Vote ratio for the two datasets

Dataset	Evaluator	Vote ratio (%)			
		Stable Diffusion	DALL-E 2	Midjourney V3	Human
AI&Painter	Public	23	24	26	27
	Art practitioners	21	23	27	29
AI&Artist	Public	15	17	16	52
	Art practitioners	14	12	12	62

Table 2 Survey results for three groups of people with different artistic appreciation abilities for the question "What is the difference between painting artists/masters and ordinary painters?"

No.	Answer by painting artists/masters	Vote ratio (%)	
		Practitioners of painting art	Public
A1	Painting artists/masters can integrate their own ideas and personality into painting works to innovate	6.67	15
A2	The works of painting artists/masters are infectious, can reflect deep meaning, and cause spiritual and emotional resonance	13.33	45
A3	Painting artists/masters have their own mental outlook and can lead the times in thinking and creative style	53.33	10
A4	Painting artists/masters integrate the work technique and spirit and emotion properly and naturally	6.67	5
A5	Painting artists/masters combine painting with cultural elements and express them in artistic form	6.67	10
A6	Painting artists/masters combine the deep feelings they want to express with the images they paint to show the characteristics of the times	6.67	15
A7	Others	6.67	0

to realize their own artistic ideas. Second, because the principle of the DM is denoising from a pure Gaussian noise image under external guidance, current stage generation is based on previous denoising results such that the model naturally and harmoniously merges multiple segments of guidance into an image.

3 Diffusion model

In this section, we briefly introduce the preliminaries of the DM for AI painting. We consider the original denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) and illustrate its diffusion and reverse diffusion processes. The diffusion process of DDPM aims to iteratively add diagonal Gaussian noise to the initial data sample x and to turn it into an isotropic Gaussian distribution after T steps:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1}, t \in \{1, 2, ..., T\}, \quad (1)$$

where the sequence x_t starts with $x_0 = x$ and ends with $x_T \sim \mathcal{N}(0, I)$, the added noise at each step is $t \sim \mathcal{N}(0, I)$, and $\{\alpha_t\}_{1,2,...,T}$ is a pre-defined schedule (Ho et al., 2020). The denoising process is the reverse of diffusion, in which the Gaussian noise $x_T \sim \mathcal{N}(0, I)$ is converted back into the data distribution x_0 through iterative denoising steps $t = T, T-1, ..., 1$. During training, for a given image x, the model calculates x_t by sampling a Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad (2)$$

where $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. Given x_t, the target of the denoising network $\epsilon_\theta(\cdot)$ is to restore x_0 by predicting the noise ϵ. It is learned via the loss function

$$\mathcal{L} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]. \quad (3)$$

With the predicted $\epsilon_\theta(x_t, t)$, the prediction of x_0 at step t by converting Eq. (2) can be obtained:

$$\hat{x}_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}[x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)]. \quad (4)$$

The reverse diffusion process of inference aims to obtain a target feature representation x_0 from an initial Gaussian noise x_T iteratively with a timestep sequence $\{T, T-1, ..., 1\}$ by using Eq. (2). Usually,

U-Net is used as the backbone network to predict $\epsilon_\theta(x_t, t)$. Moreover, classifier-free guidance is a technique to guide the iterative sampling process of a DDPM toward a conditioning signal c by mixing the predictions of a conditional model and an unconditional model:

$$\bar{\epsilon}_\theta(x_t, t, c) = (w+1)\epsilon_\theta(x_t, t, c) - w\epsilon_\theta(x_t, t), \quad (5)$$

where $w \geq 0$ is the guidance strength. In AI painting, the conditioning signal c refers to a text or an image input.

4 Method

In artistic image generation, as previously analyzed, current text-to-image methods fall short of generating images with multilevel semantics or integrating the features of multiple objects into one expressive object. To address this, we introduce our proposed multistage guidance on the text-conditioned DM. Fig. 1 shows the process.

Given a well-pretrained latent DM $\mathcal{M}$ with t steps of sampling and a sequence of k text prompts $\{P_1, P_2, ..., P_k\}$ as semantic guidance, text-conditional image synthesis generates images under the guidance of these text prompts by $\mathcal{M}$. The inference, i.e., the reverse diffusion process of $\mathcal{M}$, first encodes an image of purely Gaussian noise to the initial latent Z_t, and then the text encoder, which is usually a pretrained language model such as BERT or a pretrained vision-language model such as CLIP, encodes k text prompts $\{P_1, P_2, ..., P_k\}$ into k embeddings $\{e_1, e_2, ..., e_k\}$. We assume that the sampling steps of $\mathcal{M}$ guided by k text prompts are $\{s_1, s_2, ..., s_k\}$, where $\sum_{i=0}^{k} s_i = t$. The initial latent Z_t is guided to denoise by $\{e_1, e_2, ..., e_k\}$ with $\{s_1, s_2, ..., s_k\}$ steps and converted to the final latent Z_0. In this process, Attention U-Net (Rombach et al., 2022) or Transformer (Vaswani et al., 2017) is usually used as a backbone to predict the noise n_T given the current timestep T, the input latent Z_t, and the text embedding guidance e_T. Finally, Z_0 is inputted into the decoder to generate an obtained image.

Specifically, in Fig. 1, because the Chinese dragon is the totem of the Chinese nation, the generated fish can be given the connotation of Chinese traditional culture in artwork, so that a user wants to generate a fish with the characteristics of a Chinese dragon. By conducting two-stage guidance with "A Chinese dragon" of stage 1 with 8 steps and "A fish" of stage 2 with 12 steps, the image with Gaussian noise is denoised into an image where the characteristics, movements, and momentum of the Chinese dragon are well integrated into the fish, showing strong artistic creativity. In the following practical experiment, we focus on two-stage guidance.

5 Experiments

5.1 Main experiment

5.1.1 Setup

As shown in Fig. 2, we used ChatGPT (https://chat.openai.com/) to generate 100 image captions by imitating the original human annotation where each caption mentions two similar objects. Later, we used the extracted object prompt of each caption to retrieve the top-40 related images from the Baidu image search engine (https://image.baidu.com/). In this way, we collected a small dataset as a test set. We then forced each AI painting model to generate 40 images for evaluation. As illustrated in

$$\alpha\text{-FID} = \sqrt{[x^2 + y^2 + (x-y)^2]/2}, \quad (6)$$

where

$$x = \text{FID}(G, I^{(1)}), y = \text{FID}(G, I^{(2)}),$$

we proposed a new machine metric α-FID to measure the quality of the generated images in this test set. The design of α-FID for the evaluation objective can be illustrated as follows: On one hand, we hope that the generated images are similar to the images of both objects mentioned in a label caption, constrained by minimizing $x^2 + y^2$, where $x = \text{FID}(G, I^{(1)})$, $y = \text{FID}(G, I^{(2)})$, G represents the generated images, and $I^{(1)}$ and $I^{(2)}$ are images containing objects 1 and 2, respectively. On the other hand, we hope to balance the similarity to these two objects, that is, to minimize $x - y$. Based on the above two considerations, we proposed α-FID described in Eq. (6) as a reasonable machine metric to evaluate the quality of the generated images for artistic creativity, and the smaller, the better.

Our method was proposed to enable users to generate images of artistic creativity with their own ideas by applying current diffusion-based text-to-image models. We evaluated our method by applying it to Stable Diffusion (Since it is the only

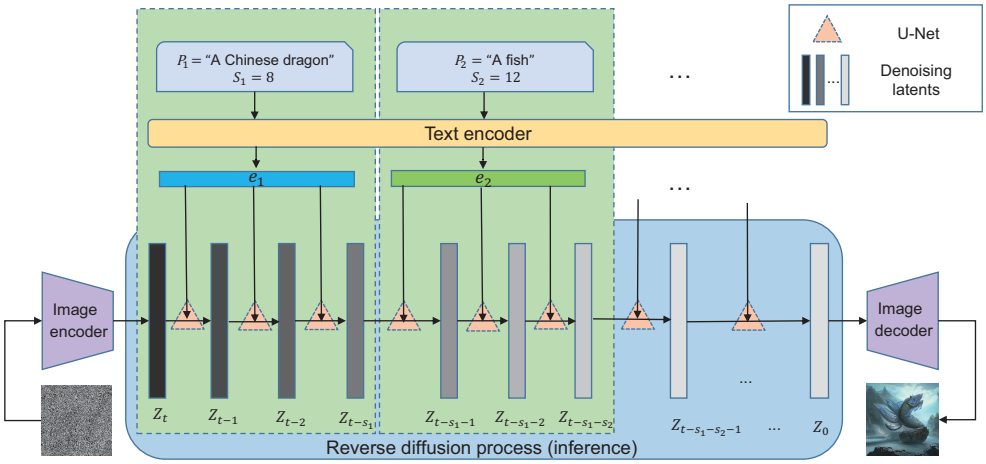

Fig. 1 Illustration of our method. In this example, there are only two-stage guidance (k=2) and the number of total guidance steps is t=20 with s_1=8 and s_2=12. Dark gray to light gray embeddings represent the denoising process in the latent space; the lighter the color, the less noise the image latent embedding contains

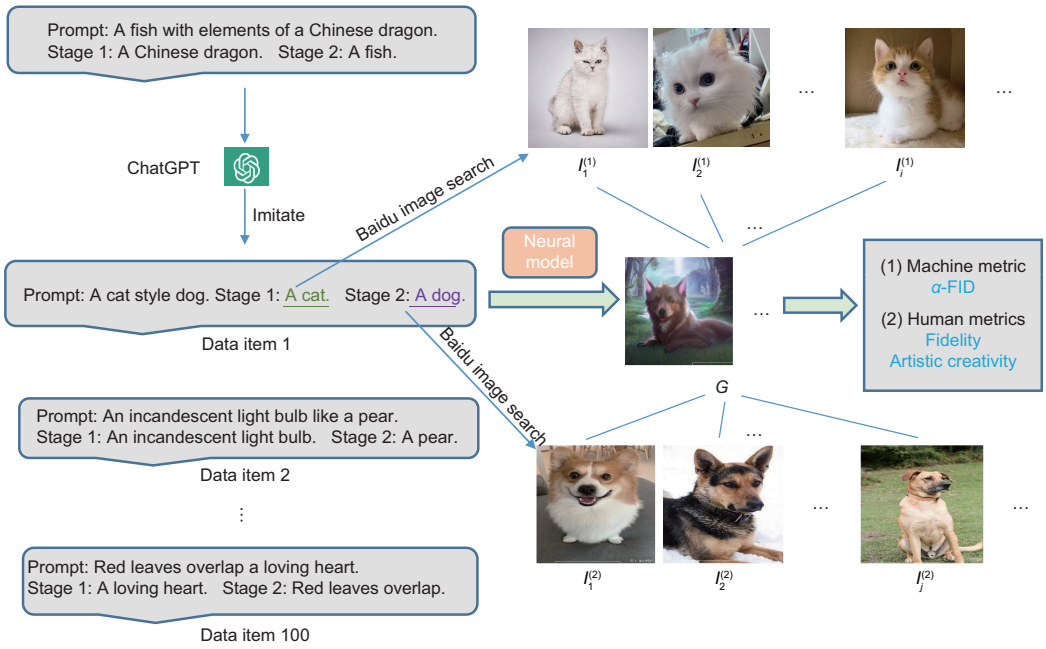

Fig. 2 Construction process of our test dataset and image data generated by AI models for evaluation

method for which the code is open over all baselines) on the aforementioned collected dataset and made a comparison with three SOTA baselines: Midjourney V3, DALL-E 2, and Stable Diffusion (without our method). We used both the machine evaluation α-FID and human evaluation metrics. For the latter, to compare the performance of different methods under the aesthetic standards of the public, 20 persons were invited to score the generated images on fidelity and artistic creativity from 1 to 5.

5.1.2 Results

As shown in Table 3, our method applying Stable Diffusion outperformed SOTA diffusion methods with one-stage guidance by a large margin in both machine evaluation and human evaluation. This verified the advantages of our proposed multistage guidance on artistic image generation, where multilevel semantic information is required for representation or multiple objects are required to be aggregated

together. Moreover, previous methods all showed unsatisfactory performance on fidelity and artistic creativity of human evaluation, which validates our motivation that artistic image generation is different from conventional photorealistic image generation, so it is difficult for previous methods to be applied directly in this field. In addition, the latent-based diffusion method (i.e., Stable Diffusion) was slightly inferior to the pixel-based diffusion methods (i.e., Midjourney and DALL-E 2), even though the former had a higher inference speed.

Table 3 Evaluation results on our collected artistic dataset

Method	Machine evaluation	Human evaluation	
	α-FID	Fidelity	Artistic creativity
Stable Diffusion	35.17	3.50	3.43
DALL-E 2	33.57	3.64	3.65
Midjourney V3	32.94	3.58	3.69
Ours	**24.51**	**3.92**	**3.95**

The bold number indicates the best performance

5.2 Case study

We randomly sampled four examples from the generated images for evaluation and made a visual comparison among DALL-E 2, Stable Diffusion, and our method. As shown in Fig. 3a, the input text required the model to generate a fish with features of Chinese dragons. DALL-E 2 generated pure dragons or a fish with fewer dragon features, while Stable Diffusion misunderstood the instruction to generate fish and generated only dragons. Benefitting from two-stage guidance on diffusion latents, our method can better fuse a fish with the characteristics of Chinese dragons by first generating a vague framework of a Chinese dragon and then detailing it with a fish. Through this two-stage guidance, our method can endow a fish with the artistic connotation of the soul of the Chinese nation, enabling users to realize their own artistic ideas. More examples are shown in Figs. 3b–3d. Moreover, even with long and complex prompt inputs, as shown in Fig. 4, our method was still able to generate satisfactory results.

5.3 Controllable guidance steps

Our method was able to control the extent to which features of an object are represented in a painting because the steps guided in each stage are tunable. We showed two examples to verify this. As shown in Fig. 5, in each example, we set seven different compositions of guiding steps ranging from "0→20" to "20→0," whose total number of guiding steps equaled 20.

In the example of Fig. 5, with the increasing number of guiding steps by "dragon" in the first stage, the generated object gradually evolved from a pure fish to a fish with the characteristics of a Chinese dragon showing magnificent momentum and serious demeanor, and then was completely transformed into a dragon. Similarly, in the second one, benefiting from our method, the dog was gradually endowed with more cuteness and the clever characteristics of a cat. Among them, guiding compositions from "$x = 6$, $y = 14$" to "$x = 10$, $y = 10$" was suitable for an artwork in which features were mixed and users can set the detailed compositions according to their needs.

6 Related works

6.1 Text-conditional image synthesis

Text-conditional image synthesis generates photorealistic images under the guidance of input text. Generative modeling faces unique difficulties due to the high dimensionality of pictures. GANs (Goodfellow et al., 2020) enable effective sampling of high-resolution pictures with excellent perceptual quality (Brock et al., 2019; Karras et al., 2020), but it is challenging to improve them (Arjovsky et al., 2017; Gulrajani et al., 2017; Mescheder, 2018) and there are problems in capturing the complete data distribution (Metz et al., 2017). While likelihood-based techniques place more emphasis on accurate density prediction, optimization behaves better as a result. High-resolution pictures can be synthesized effectively by using variational autoencoders (VAEs) (Kingma and Welling, 2014) and flow-based models (Dinh et al., 2015, 2017), but sample quality is not on pace with that of GANs. A sequential sampling procedure and computationally intensive architectures (Vaswani et al., 2017) restrict the sharpness of the images that autoregressive models (ARMs) (van den Oord et al., 2016a, 2016b; Child et al., 2019; Chen M et al., 2020) can produce, despite their good performance in density estimation. Each generative model has its own distinct shortcomings; this makes it difficult to reach the level of human artist painting.

Fig. 3 Comparison of images generated using DALL-E 2 (top), Stable Diffusion (middle), and our method (bottom). We show four examples, which are dragon & fish (a), cat & dog (b), light bulb & pear (c), and red leaves & heart (d), where each method represents three generated images

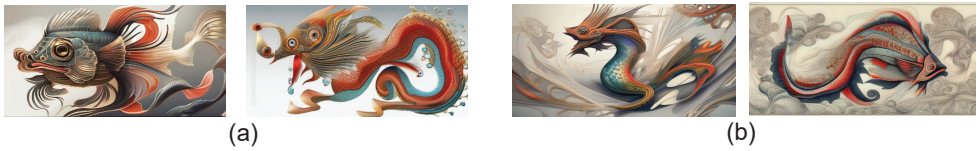

Fig. 4 Two examples of the generated result with long and complex prompt inputs. (a) Stage 1: A Chinese dragon looking at the front horizontally is flying on the sky; stage 2: A fish with big eyes is looking at the surroundings, gradient colors background. (b) Stage 1: A colorful Chinese dragon with dark red and dark blue scales is flying on the sky; stage 2: A fish with long scales is dropping out of water stirring up waves

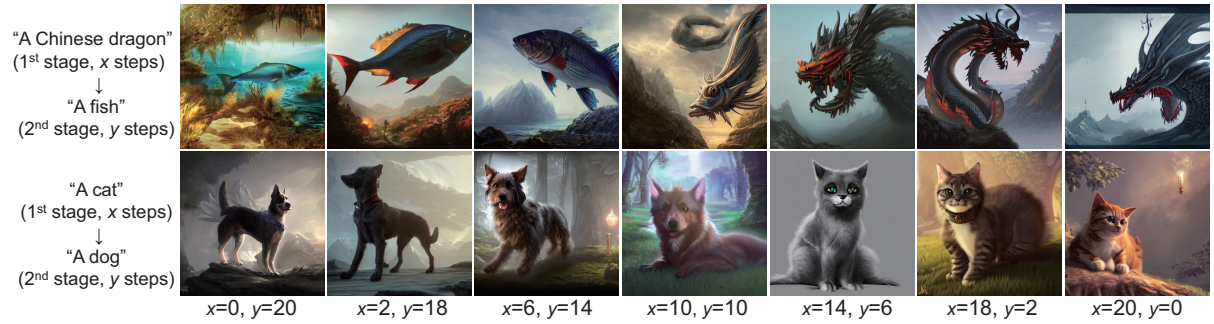

Fig. 5 Two examples of the study of different compositions for the guiding steps

6.2 Diffusion model

As a generative model with multistep inference, DM has achieved success in a wide range of generation tasks, such as Stable Diffusion (Rombach et al., 2022) in image synthesis, DiffWave in audio generation, and LatentOps in natural language generation. Specifically, in computer vision, Ho et al. (2020) proposed DDPM, and Nichol and Dhariwal (2021) improved it with several simple modifications. Song J et al. (2021) proposed denoising diffusion implicit models (DDIMs) to accelerate the sampling of DDPMs with the same training procedure. Ho and Salimans (2021) proposed classifier-free diffusion guidance, which enables DM to work on text-conditional image synthesis.

There are also several deep neural models that apply DM to image synthesis, such as GLIDE (Nichol et al., 2022), Imagen (Saharia et al., 2022), DALL-E 2 (Ramesh et al., 2022), and Stable Diffusion (Rombach et al., 2022). However, the performance of these methods is still inferior to that of the artwork of human artists. Our proposed multistage text-conditioned approach ingeniously uses the multistep inference characteristic of DM and enables it to generate creative images according to users' ideas, which helps current diffusion-based methods go a step further toward human-artist-level painting.

7 Conclusions and future work

In this paper, by soliciting opinions from three groups of people with different levels of art appreciation ability, we reveal that a recognized gap exists between recent SOTA text-to-image methods and human artist painters. Based on this, we propose a multistage text-conditioned approach to help current diffusion-based methods move toward human-artist-level painting. Both manual evaluation and machine metric evaluation verify the effectiveness of our approach. Finally, by tuning the number of guiding steps in each stage, our method is able to control the extent to which features of an object are represented in a painting.

Theoretically, because our proposed multistage guidance approach can be applied not only to text-conditioned image generation, but also to other generation tasks using DMs, such as conditional audio generation and conditional text generation, it is worth verifying its effectiveness in those areas in future work.

Contributors

Taihao LI designed the research. Wang QI and Huanghuang DENG developed the methodology, collected the data, and worked on the software. Wang QI drafted the paper. Huanghuang DENG helped organize the paper. All the authors revised and finalized the paper.

Compliance with ethics guidelines

Wang QI, Huanghuang DENG, and Taihao LI declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Arjovsky M, Chintala S, Bottou L, 2017. Wasserstein GAN. https://arxiv.org/abs/1701.07875

Brock A, Donahue J, Simonyan K, 2019. Large scale GAN training for high fidelity natural image synthesis. Proc 7th Int Conf on Learning Representations.

Chen M, Radford A, Child R, et al., 2020. Generative pretraining from pixels. Proc 37th Int Conf on Machine Learning, p.1691-1703.

Chen N, Zhang Y, Zen H, et al., 2021. WaveGrad: estimating gradients for waveform generation. Proc 9th Int Conf on Learning Representations.

Child R, Gray S, Radford A, et al., 2019. Generating long sequences with sparse transformers. https://arxiv.org/abs/1904.10509

Dinh L, Krueger D, Bengio Y, 2015. NICE: non-linear independent components estimation. Proc 3rd Int Conf on Learning Representations.

Dinh L, Sohl-Dickstein J, Bengio S, 2017. Density estimation using real NVP. Proc 5th Int Conf on Learning Representations.

Goodfellow I, Pouget-Abadie J, Mirza M, et al., 2020. Generative adversarial networks. *Commun ACM*, 63(11):139-144. https://doi.org/10.1145/3422622

Gulrajani I, Ahmed F, Arjovsky M, et al., 2017. Improved training of wasserstein GANs. Proc 31st Int Conf on Neural Information Processing Systems, p.5767-5777.

Ho J, Salimans T, 2021. Classifier-free diffusion guidance. Proc Workshop on Deep Generative Models and Downstream Applications.

Ho J, Jain A, Abbeel P, 2020. Denoising diffusion probabilistic models. Proc 34th Int Conf on Neural Information Processing Systems, Article 574.

Karras T, Laine S, Aittala M, et al., 2020. Analyzing and improving the image quality of StyleGAN. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8107-8116. https://doi.org/10.1109/CVPR42600.2020.00813

Karras T, Laine S, Aila T, 2021. A style-based generator architecture for generative adversarial networks. *IEEE Trans Patt Anal Mach Intell*, 43(12):4217-4228. https://doi.org/10.1109/TPAMI.2020.2970919

Kingma DP, Welling M, 2014. Auto-encoding variational Bayes. Proc 2nd Int Conf on Learning Representations.

Kingma DP, Salimans T, Poole B, et al., 2021. Variational diffusion models. https://arxiv.org/abs/2107.00630

Kong ZF, Ping W, Huang JJ, et al., 2021. DiffWave: a versatile diffusion model for audio synthesis. Proc 9th Int Conf on Learning Representations.

Mescheder L, 2018. On the convergence properties of GAN training. https://arxiv.org/abs/1801.04406v1

Metz L, Poole B, Pfau D, et al., 2017. Unrolled generative adversarial networks. Proc 5th Int Conf on Learning Representations.

Mittal G, Engel JH, Hawthorne C, et al., 2021. Symbolic music generation with diffusion models. Proc 22nd Int Society for Music Information Retrieval Conf, p.468-475.

Nichol AQ, Dhariwal P, 2021. Improved denoising diffusion probabilistic models. Proc 38th Int Conf on Machine Learning, p.8162-8171.

Nichol AQ, Dhariwal P, Ramesh A, et al., 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. Proc 39th Int Conf on Machine Learning, p.16784-16804.

Ramesh A, Pavlov M, Goh G, et al., 2021. Zero-shot text-to-image generation. Proc 38th Int Conf on Machine Learning, p.8821-8831.

Ramesh A, Dhariwal P, Nichol A, et al., 2022. Hierarchical text-conditional image generation with clip latents. https://arxiv.org/abs/2204.06125

Razavi A, van den Oord A, Vinyals O, 2019. Generating diverse high-fidelity images with VQ-VAE-2. Proc 33rd Int Conf on Neural Information Processing Systems, Article 1331.

Rombach R, Blattmann A, Lorenz D, et al., 2022. High-resolution image synthesis with latent diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10684-10695. https://doi.org/10.1109/CVPR52688.2022.01042

Saharia C, Chan W, Saxena S, et al., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Proc 36th Int Conf on Neural Information Processing Systems, p.36479-36494.

Sohl-Dickstein J, Weiss EA, Maheswaranathan N, et al., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. Proc 32nd Int Conf on Machine Learning, p.2256-2265.

Song J, Meng C, Ermon S, 2021. Denoising diffusion implicit models. Proc 9th Int Conf on Learning Representations.

Song Y, Sohl-Dickstein J, Kingma DP, et al., 2021. Score-based generative modeling through stochastic differential equations. Proc 9th Int Conf on Learning Representations.

van den Oord A, Kalchbrenner N, Espeholt L, et al., 2016a. Conditional image generation with pixelcnn decoders. Proc 30th Int Conf on Neural Information Processing Systems, p.4797-4805.

van den Oord A, Kalchbrenner N, Kavukcuoglu K, 2016b. Pixel recurrent neural networks. Proc 33rd Int Conf on Machine Learning, p.1747-1756.

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010.