Wang QI, Huanghuang DENG, Taihao LI, 2023. Multistage guidance on the diffusion model inspired by human artists' creative thinking. *Frontiers of Information Technology & Electronic Engineering*, early access. https://doi.org/10.1631/FITEE.2300313

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

Key words: Text-conditional image generation; Diffusion model; Multilevel semantics; AI painting system

Taihao Li
E-mail: lith@zhejianglab.com
ORCID: https://orcid.org/0000-0003-3279-7125

Motivation

- Benefiting from large diffusion models (DMs), such as Disco Diffusion and Stable Diffusion, recent artificial intelligence (AI) painting systems show performance comparable to that of human painters. However, as our preliminary experiment in Table 1 shows, there still exists a large gap when compared to that of human artists. It is still difficult for the current AI painting system to generate images of a similar artist level.

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

The evaluations of different levels of evaluators on the works of AI-generated models and human painters

Main idea

- We first reveal two recognized differences between current AI painting systems and human artists by soliciting the opinions of people with different levels of artistic experience. The differences lie in the ability to construct works that contain multiple semantic levels and combine various objects together, or endow one object with features of other objects.
- Aiming to improve current image generation techniques toward the painting ability of human artists, we propose a multistage text-conditioned approach using a DM. The proposed approach is for working on a DM without any further pretraining to help current AI painting systems approach human-artist-level painting. Different from previous one-stage guidance, by tuning guiding steps in each stage, the proposed method is able to control the extent to which features of an object are represented in a painting.

Diffusion model

- Diffusion models (DMs) are the fundamental component of artistic image generation. The diffusion process of DMs is to iteratively add diagonal Gaussian noise to the initial data sample:

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

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 $\{\alpha_t\}_{1,2,\cdots,T}$ is a predefined schedule.

- The reverse diffusion process of inference aims to obtain a target feature representation x_0 from an initial Gaussian noise x_T iteratively:

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

Usually, U-Net is used as the backbone network to predict $\epsilon_\theta(x_t, t)$. Moreover, classifier-free guidance is used to guide the process: $\bar{\epsilon}_\theta(x_t, t, c) = (w+1)\epsilon_\theta(x_t, t, c) - w\epsilon_\theta(x_t, t)$

Framework and method

- The proposed multistage guidance on the text conditioned focuses on the inference of a well-pretrained latent DM. Given a sequence of text prompts as semantic guidance, text-conditional image synthesis generates images guided by different step sizes of these text prompts.

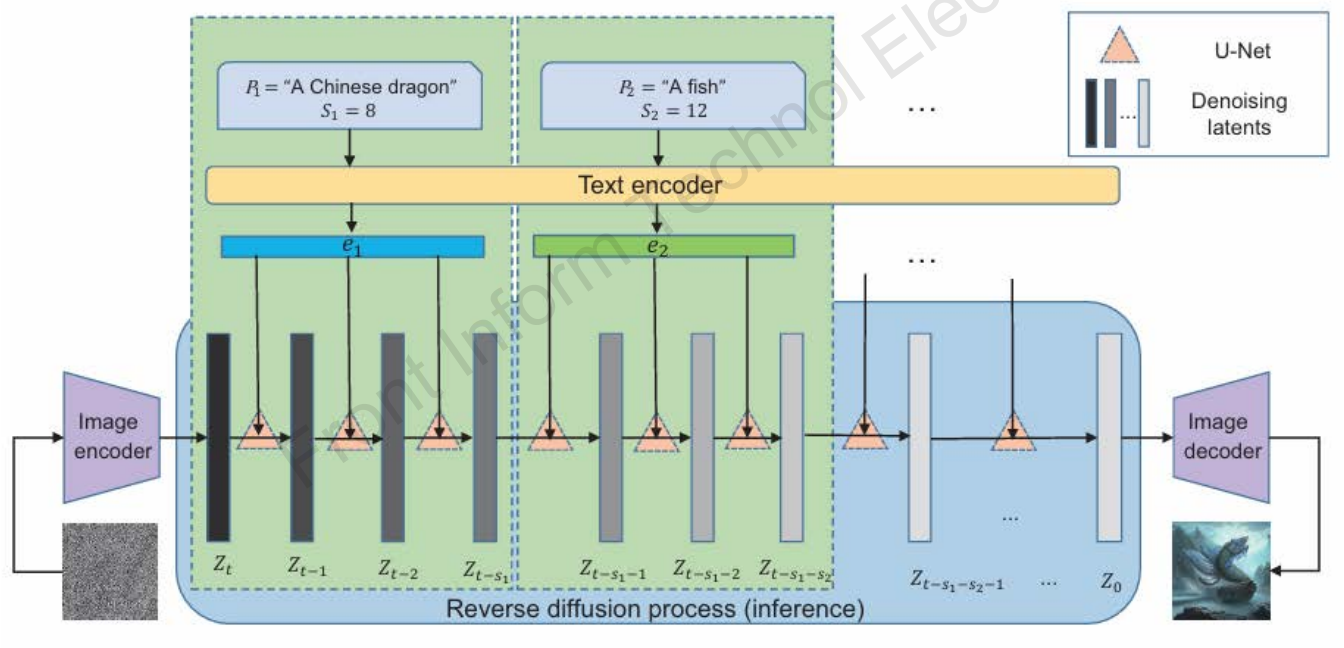
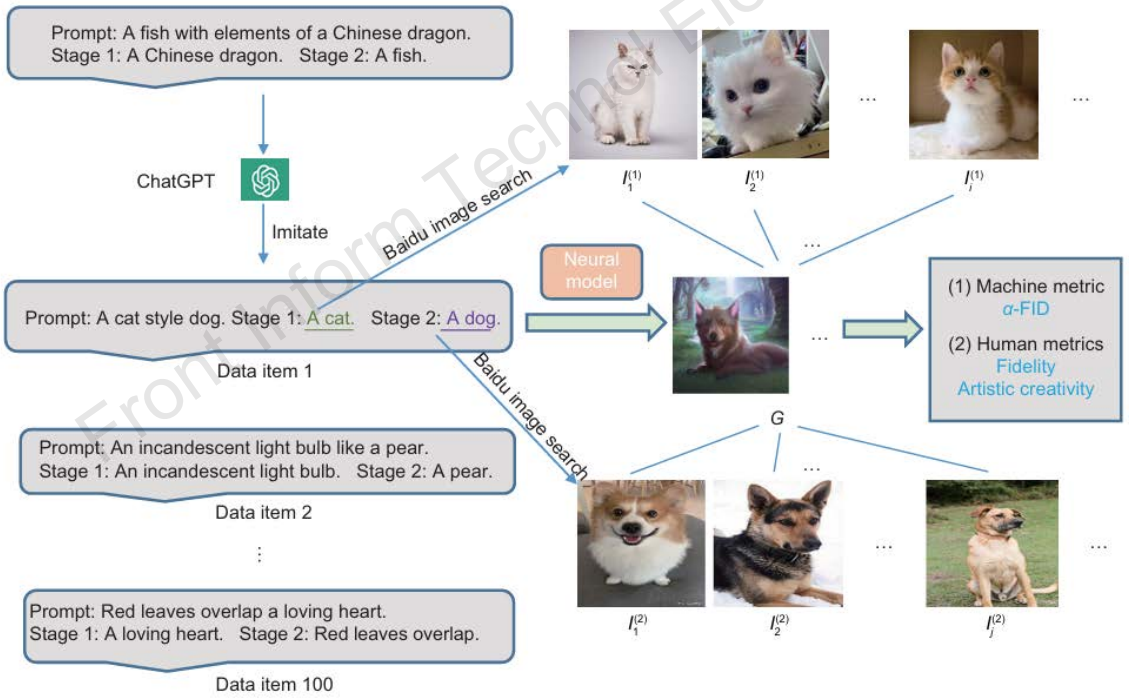


Illustration of the proposed method

Framework and method

- We used ChatGPT to generate 100 image captions. Later, we used the extracted object prompt of each caption to retrieve the top-40 related images. 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.



Process of constructing the test dataset

Major results

- We used both the machine evaluation α-FID and human evaluation metrics. We randomly sampled four examples from the generated images for evaluation and made a visual comparison.

Table 3 Evaluation results on our collected artistic dataset

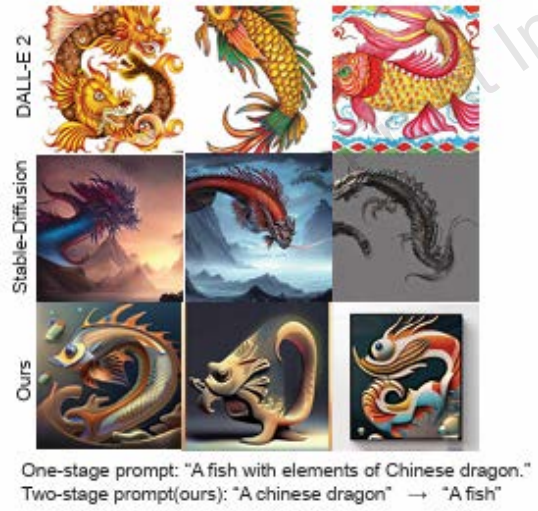
Method	Machine evaluation	Human evaluation	
	α-FID	Fidelity	Artistic creativity
Stable Diffusion	35.17	3.50	3.43
DALL-E 2	33.57	3.64	3.65
Midjourney V3	32.94	3.58	3.69
Ours	**24.51**	**3.92**	**3.95**

The bold number indicates the best performance

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

where

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



One-stage prompt: "A fish with elements of Chinese dragon."
Two-stage prompt(ours): "A chinese dragon" → "A fish"

One-stage prompt: "A cat style dog."
Two-stage prompt(ours): "A cat" → "A dog"

One-stage prompt: "An incandescent light bulb like a pear."
Two-stage prompt (ours): "An incandescent light bulb" → "A pear"

Ablation study

- The long text testing of the proposed method


(a)

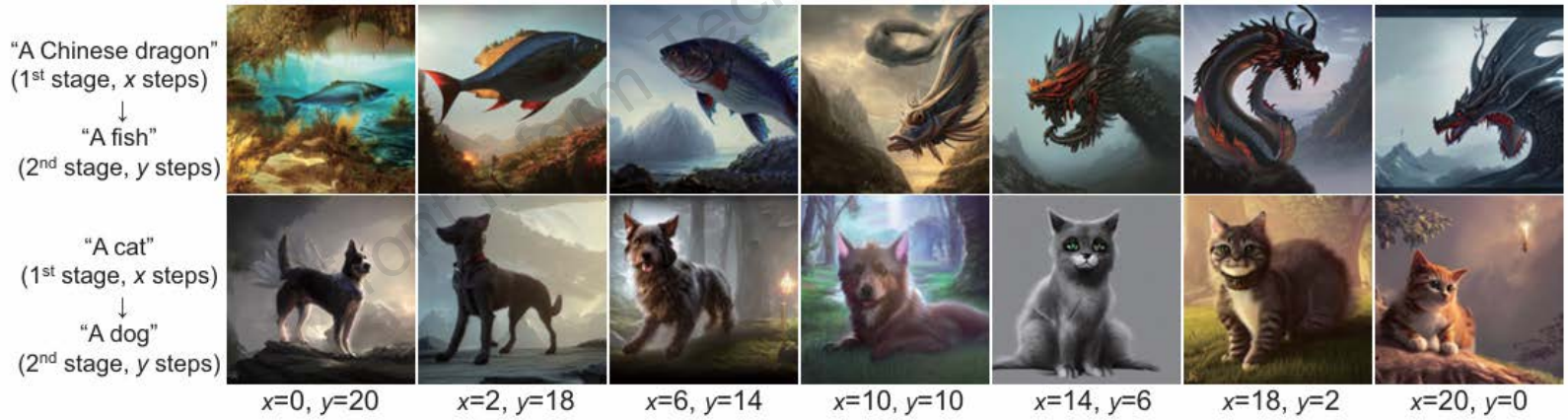
1st stage: A Chinese dragon looking at the front horizontally is flying on the sky.
2nd stage: A fish with big eyes is looking at the surroundings, gradient colors background.


(b)

1st stage: A colorful Chinese dragon with dark red and dark blue scales is flying on the sky.
2nd stage: A fish with long scales is dropping out of water stirring up waves.

Ablation study

- The proposed 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. 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.



Two examples of the study of different compositions for the guiding steps

Conclusions

- 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 state-of-the-art text-to-image methods and human artist painters.
- Based on the observations, we propose a multistage text-conditioned approach to help current diffusion-based methods move toward human-artist level painting. 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. Both manual evaluation and machine metric evaluation verify the effectiveness of our approach.