Supplementary Materials: InstantAS: Minimum Coverage Sampling for Arbitrary-Size Image Generation

1 DENOISING DIFFUSION MODEL

Given a dataset $D = \{x_i\}_{i=1}^{N}$, the diffusion model learns a mapping from the standard normal distribution $\mathcal{N}(0, I)$ to the distribution $p(x)$ to which D belongs. As a result, we can generate a random vector $\epsilon \sim \mathcal{N}(0, I)$ and feed it into the diffusion model to obtain a corresponding x, which will be a sample from the distribution $p(x)$.

In the training phase, the diffusion model is divided into a forward process and a reverse process. The forward process is a Markov process that adds noise to samples x_0 in the dataset:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon \tag{1}$$

where β_t represents a sequence of hyperparameters that is manually specified, and ϵ is a random noise vector sampled from the standard normal distribution $\mathcal{N}(0, I)$. We can write Equation 1 in the form of a probability distribution:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

After performing the above forward process T times, we obtain x_T, and by setting β_t appropriately, we can make x_T approximately follow the distribution $\mathcal{N}(0, I)$. In addition, due to the characteristics of the Markov process, we can simplify the forward process into one step:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{3}$$

where $\bar{\alpha}_t = \prod_{s=0}^{t}\alpha_s$ and $\alpha_t = 1-\beta_t$. Due to the reparameterization trick, ϵ still obeys $\mathcal{N}(0, I)$. Therefore, we can write Equation 3 in the form of a probability distribution:

$$q(x_t|x_0) = \mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{4}$$

According to Bayes rule, we can obtain the probability distribution expression for the reverse process:

$$q(x_{t-1}|x_t) = \frac{q(x_t|x_{t-1})q(x_{t-1})}{q(x_t)} \tag{5}$$

Since we do not know the exact form of $q(x_0)$, we cannot solve the integral for computing $q(x_t)$: $q(x_t) = \int q(x_t|x_0)q(x_0)dx_0$. Instead, we use the empirical distribution to approximate $q(x_0)$, by adding x_0 as a condition to the probability distribution. Therefore, Equation 5 can be written as:

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{6}$$

Obviously Equation 6 can be calculated from Equation 2 and Equation 4:

$$\begin{aligned} &q(x_{t-1}|x_t, x_0) \\ &= \mathcal{N}\left(x_{t-1}|\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{(1-\bar{\alpha}_{t-1})\sqrt{1-\beta_t}}{1-\bar{\alpha}_t}x_t, \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t I\right) \end{aligned} \tag{7}$$

We cannot get x_0 directly during generation, so we use a neural network $s_\theta(x_t, t)$ to fit x_0:

$$\mathcal{L} = \|s_\theta(x_t, t) - x_0\|_2^2 \tag{8}$$

In the sampling phase, we use the trained neural network $s_\theta(x_t, t)$ to replace x_0 in Equation 6, thus obtaining the reverse process in the sampling step:

$$x_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}s_\theta(x_t, t)\right) + \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\epsilon \tag{9}$$

The reverse process is iterated T times, and the final x_0 is the generated output of the model. For a detailed understanding of diffusion models, please refer to reference [1, 3–11].

2 EXPERIMENT IMPLEMENTATION DETAILS

In all experiments conducted in this paper, we use the unified pretrained text-to-image model StableDiffusion 2.0[1] for InstantAS and all compared sampling methods. In the comparison experiment of sampling speed, we use the DDIM method with a fixed step size of 50 for all models. For the sake of fairness, we removed the acceleration packages in the code of some models, including xformers and accelerate, etc., so that all models can be compared under the same conditions.

3 REGIONAL CONTROL GENERATION COMPARISON

In this section, we demonstrate the region control generation capabilities of InstantAS and MultiDiffusion[2], as shown in Figure 1. The mask-based generation region limitation method used by MultiDiffusion employs masks to segment the latent representation of an image into different semantic regions, and applies different prompts to these regions separately. In contrast, InstantAS aligns semantic information to different regions in cross-attention through masks. Compared to MultiDiffusion, the target region shape control of MultiDiffusion is more refined. However, since InstantAS performs alignment after modification in the deeper feature cross-attention, the boundaries it affects cannot be represented very precisely. Nevertheless, InstantAS demonstrates significantly better fusion effects between regions than MultiDiffusion, resulting in more consistent and natural-looking generated images overall.

4 FAILURE CASES AND DISCUSSION

Experimental results show that compared to previous work, our method has significant advantages in terms of generation quality and sampling speed. However, our method still has some shortcomings. First, because larger-sized images exceed the training resolution of the diffusion model, it is difficult to find their latent representations in the latent space of the diffusion model. As a result, the generated images struggle to account for information that is far apart in the image, losing some global structure. This is specifically manifested in prompts with a single target or larger entities, where the method cannot accurately restore the content. Secondly, when the content described by the text is difficult to be compatible with the specified target image size, sometimes the content described by

[1] https://huggingface.co/stabilityai/stable-diffusion-2-base

the text will be stacked in the image. Some failure cases are shown in Figure 2.

According to the manifold distribution hypothesis, for a fixed prompt, the latent feature distribution of its corresponding image data is located near a low-dimensional manifold within the high-dimensional space represented by a fixed-size image training set. However, for images of any size larger than this fixed size, they exist in a higher dimensional space. Therefore, how to utilize this low-dimensional manifold to search for embedding representations in the higher dimensional space, thereby obtaining images with consistency in the latent space, is the direction of our future research.

REFERENCES

[1] Brian D.O. Anderson. 1982. Reverse-time diffusion equation models. *Stochastic Processes and their Applications* 12, 3 (1982), 313–326. https://doi.org/10.1016/0304-4149(82)90051-5

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation. (2023).

[3] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Conference and Workshop on Neural Information Processing Systems* 34 (2021).

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[5] Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. (2021). https://openreview.net/forum?id=qw8AKxfYbI

[6] Cheng Lu et al. 2022. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *arXiv preprint arXiv:2211.01095* (2022).

[7] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan LI, and Jun Zhu. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. 35 (2022), 5775–5787. https://proceedings.neurips.cc/paper_files/paper/2022/file/260a14acce2a89dad36adc8eefe7c59e-Paper-Conference.pdf

[8] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[9] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. (2021). https://openreview.net/forum?id=St1giarCHLP

[11] Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

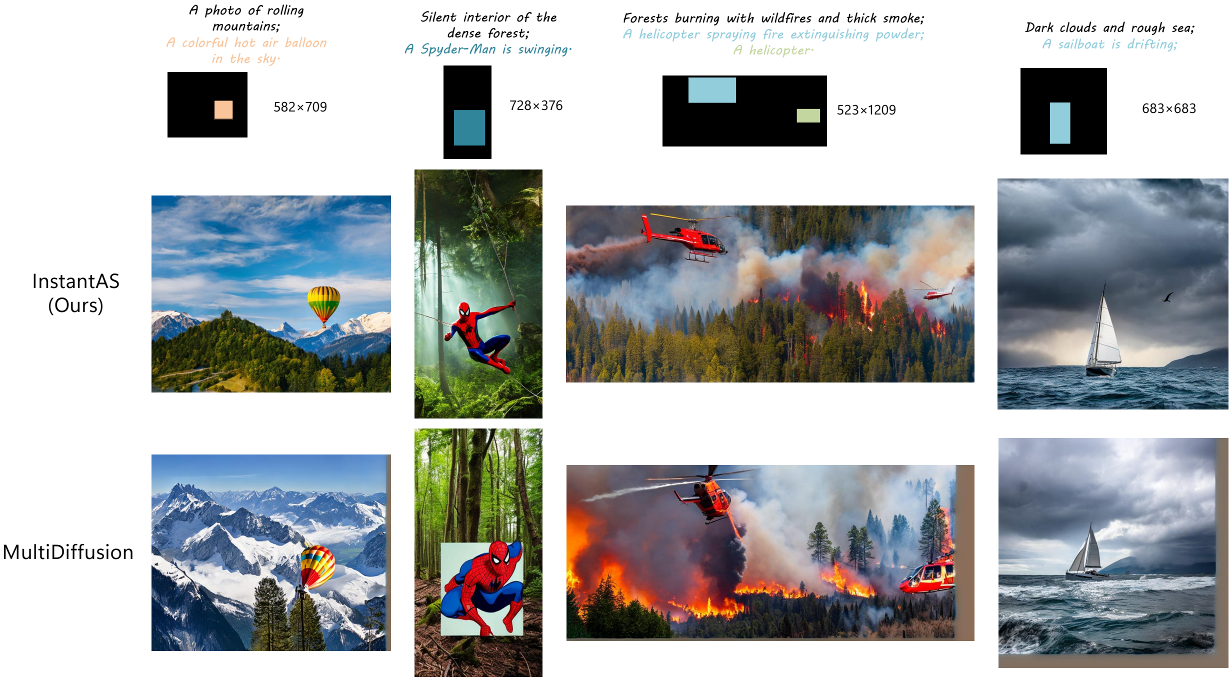
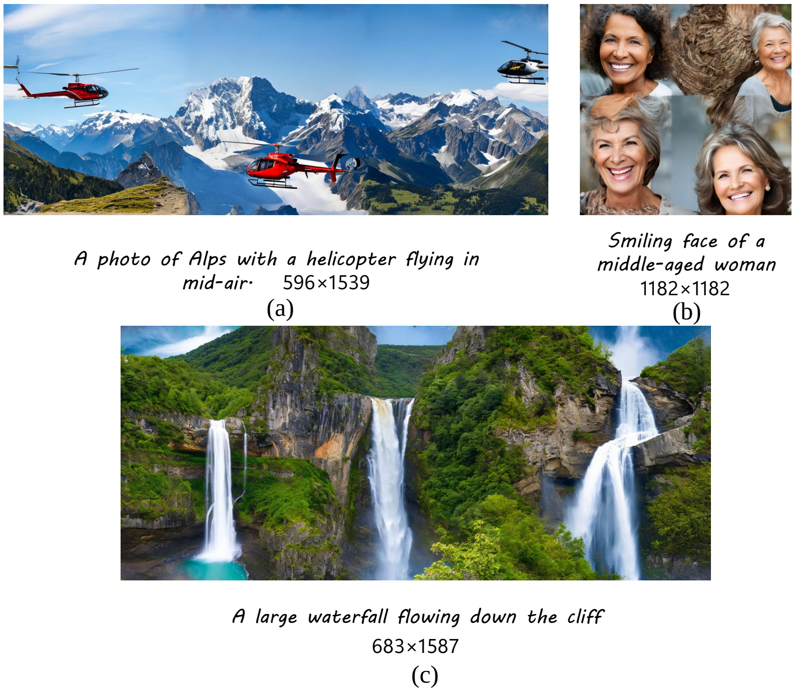

A photo of rolling mountains;
A colorful hot air balloon in the sky.
582×709
Silent interior of the dense forest;
A Spyder-Man is swinging.
728×376
Forests burning with wildfires and thick smoke;
A helicopter spraying fire extinguishing powder;
A helicopter.
523×1209
Dark clouds and rough sea;
A sailboat is drifting;
683×683
InstantAS (Ours)
MultiDiffusion

Figure 1: Comparison results of regional information control generation between InstantAS and MultiDiffusion. Our proposed InstantAS method has a more natural fusion of foreground and background, and does not appear contentless boundaries.



A photo of Alps with a helicopter flying in mid-air. 596×1539
(a)

Smiling face of a middle-aged woman
1182×1182
(b)

A large waterfall flowing down the cliff
683×1587
(c)

Figure 2: Some failure cases. (a) (b) When the described content includes specific quantities or entities with complete structures that occupy most of the image area, repetition of elements often occurs. (c) When the description content is incompatible with the image size, for example, when it is required to generate a waterfall from top to bottom in a horizontally long image, the generated results will often have content stacking.