Not All Parameters Matter: Masking Diffusion Models for Enhancing Generation Ability

Lei Wang [1] Senmao Li [1] Fei Yang [1] (✉) Jianye Wang [1] Ziheng Zhang [1] Yuhan Liu [1] Yaxing Wang [1, 2] Jian Yang [1] (✉)

[1]PCA Lab, VCIP, College of Computer Science, Nankai University [2]Shenzhen Futian, NKIARI

Motivation

The diffusion models, in early stages focus on constructing basic image structures, while the refined details, including local features and textures, are generated in later stages. Thus the same network layers are forced to learn both structural and textural information simultaneously, significantly differing from the traditional deep learning architectures (e.g., ResNet or GANs) which captures or generates the image semantic information at different layers. This difference inspires us to explore the time-wise diffusion models. We initially investigate the key contributions of the U-Net parameters to the denoising process and identify that properly zeroing out certain parameters (including large parameters) contributes to denoising, substantially improving the generation quality on the fly (see Fig. 1). Capitalizing on this discovery, we propose a simple yet effective method—termed "MaskUNet"— that enhances generation quality with negligible parameter numbers (see Fig. 2).

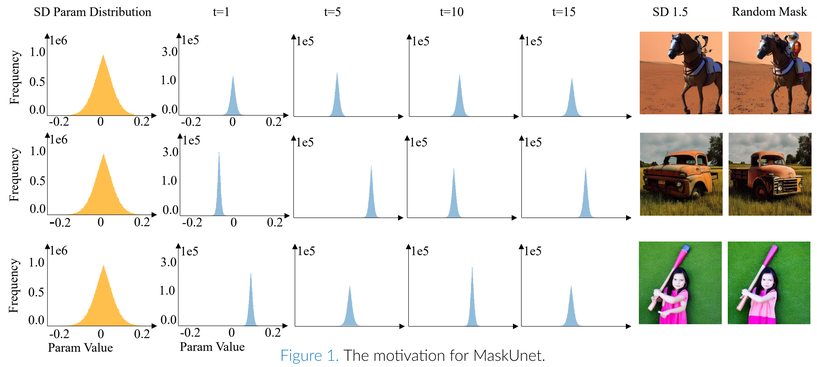

Figure 1. The motivation for MaskUnet.

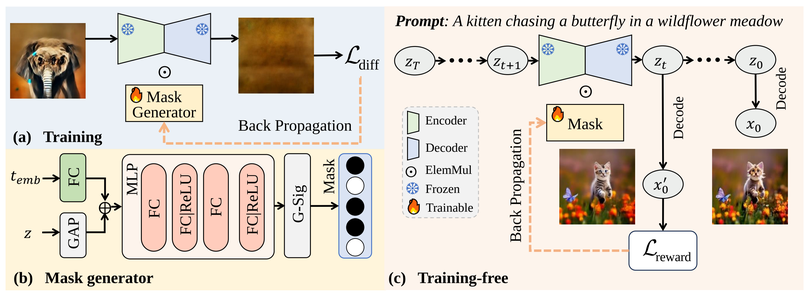

Figure 2. The pipeline of the MaskUnet.

Highlights

- We conduct an in-depth study of the relationship between parameters in the pre-trained U-Net, samples, and timesteps, revealing the effectiveness of parameter independence, which provides a new perspective for efficient utilization of U-Net parameters.
- We propose a novel fine-tuning framework for text-to-image pre-trained diffusion models, called MaskUNet. In this framework, the training-based method optimizes masks through diffusion loss, while the training-free method uses a reward model to optimize masks. The learnable masks enhance U-Net's capabilities while preserving model generalization.
- We evaluate MaskUNet on the COCO dataset and various downstream tasks. Experimental results demonstrate significant improvements in sample quality and substantial performance gains in key metrics.

Zero-shot Text-to-image Generation

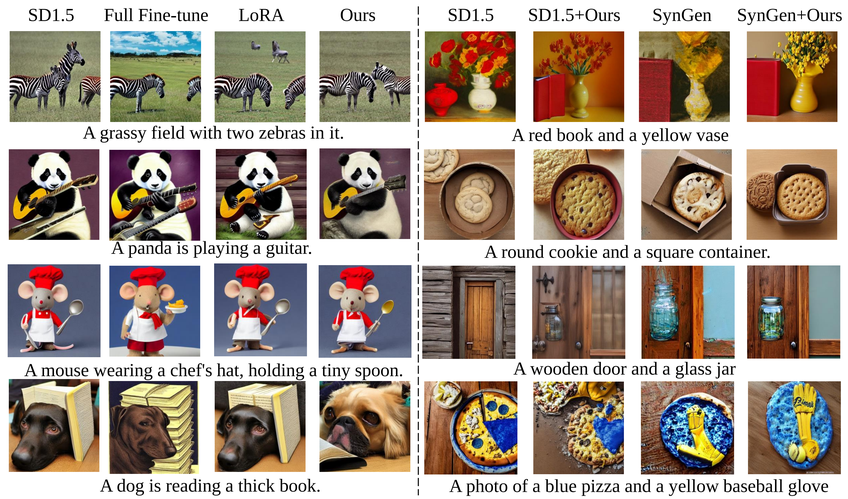

Figure 3. Quality results compared to other methods.

Table 1. Quantitative results of zero-shot generation on the COCO 2014 and COCO 2017 datasets, with the best results in **bold**.

Method	COCO 2014		COCO 2017	
	FID-30k (↓)	CLIP (↑)	FID-5k (↓)	CLIP (↑)
SD 1.5	12.85	0.32	23.39	0.33
Full Fine-tune	14.06	0.32	24.45	0.33
LoRA	12.82	0.32	23.18	0.33
MaskUnet	**11.72**	0.32	**21.88**	0.33

Text-to-video Generation

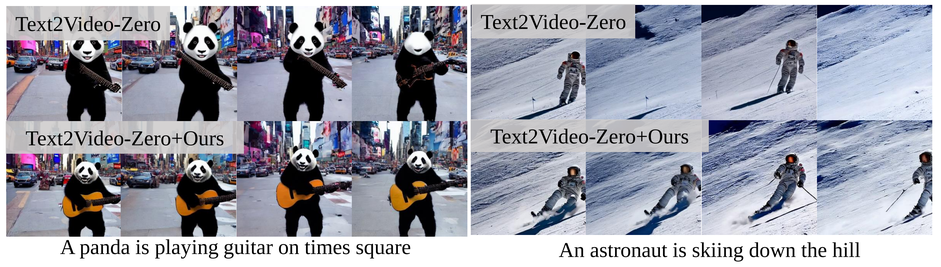

Figure 4. Quality results by Text2Video-Zero with or without mask.

Image Customization

Figure 5. Quality results compared to other methods.

Figure 6. Quality results by ReVersion w or w/o mask.

Acknowledgements

This work was supported by the National Science Fund of China under Grant Nos. 62361166670 and U24A20330, the "Science and Technology Yongjiang 20" key technology breakthrough plan project (2024Z120), the Shenzhen Science and Technology Program (JCYJ20240813114237048), and the Supercomputing Center of Nankai University (NKSC).