Not All Parameters Matter: Masking Diffusion Models for Enhancing Generation Ability (*Supplementary Material*)

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
{scitop1998, senmaonk, feiyangflyhigher}@gmail.com, {yaxing,csjyang}@nankai.edu.cn

A. Additional Ablation Studies

We present the results of different reward models in Table 2. Overall, the impact of excluding a specific reward model (e.g., HPSv2 [11] or ImageReward [12]) demonstrates that while these individual models positively contribute to specific tasks, they are not the sole determining factors, and their combination maximizes performance improvement. This further validates the importance of leveraging multiple reward signals comprehensively, enabling the capture of more holistic semantic features for semantic binding tasks.

We observe that after 7k iterations, the mask ratio stabilizes, while FID briefly fluctuates before settling. This temporary fluctuation is likely due to the model adapting to the finalized mask pattern (Figure 2 (a)).

Inspired by Faster Diffusion [6], PnP [4], and DIFT [10], which emphasize the decoder's role in generation, we applied masking to the decoder. To reduce inference latency (Table 1), we limited masking to linear layers, excluding convolutional layers.

Table 1. Ablation study on the impact of different layers, with the best results in **bold**.

Setting	FID	CLIP	Add Param.	Latency (s)
All	22.10	0.32	160.15M	6.27
Decoder (q, k, v, o, conv)	**21.79**	0.33	81.69M	4.02
Decoder (q, k, v, o)	21.88	0.33	**55.06M**	**2.99**

B. Additional Results

B.1. Visualization of Mask Position

We visualized the mask of the v matrix in the third attention layer of the second decoder block in the U-Net, as shown in Figure 4. Although the mask ratio consistently remains at 8.24%, its positions vary across different timesteps. This indicates that the mask introduces timestep dependency, allowing each timestep to have its unique U-Net weight distribution.

[†]Corresponding authors.

Table 2. Ablation study on the impact of different reward models on T2I-CompBench, with the best results in **bold**.

Method	BLIP-VQA		
	Color (↑)	Texture (↑)	Shape (↑)
SD 1.5 [8]	0.3750	0.4159	0.3742
w/o HPSv2 [11]	0.4530	0.4871	0.4202
w/o ImageReward [12]	0.4502	**0.4949**	0.4254
MaskUNet	**0.4958**	0.4938	**0.4529**

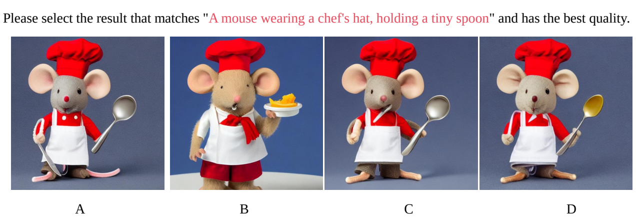

Figure 1. An example of the user study.

B.2. Mask in different layers

A high mask ratio in layers 2-5 indicates that blocks 2 and 3 in the decoder are the key decision-making layers in UNet computation (Figure 2 (b)).

B.3. More Mask Visualizations

We add more mask visualizations for image customization task in Figure 2 (c) and will include them in the final version. MaskUnet uses varying masking parameters for different samples.

B.4. Mask analysis

We visualized the masks of q, k, and o (Figure 2 (d)), revealing significant timestep inconsistencies due to their dynamic adaptation to timestep-specific features (q/k for attention

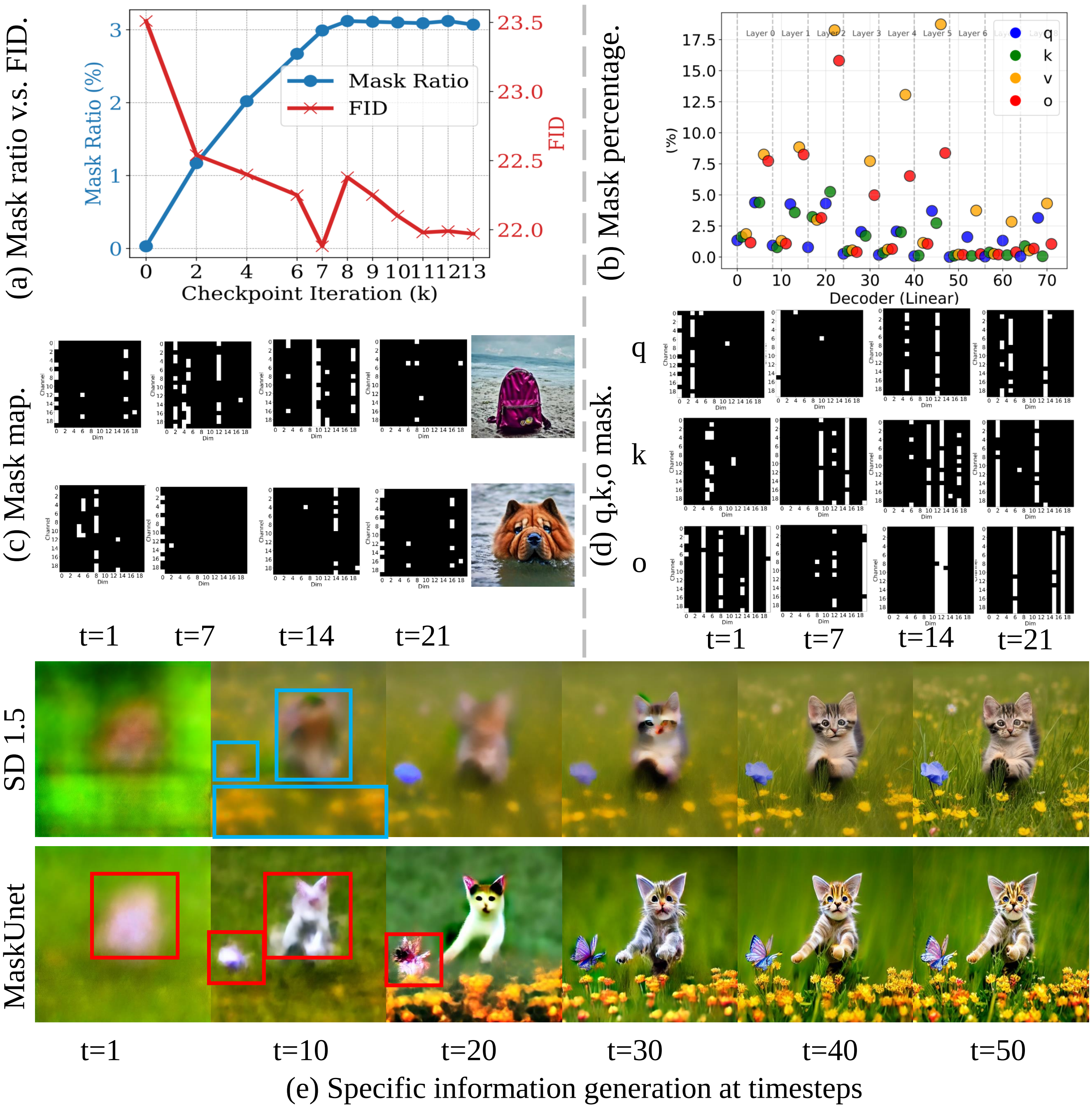

Figure 2. More results.

weights, o for attention outputs). In contrast, v masks remain consistent, reflecting its role in preserving key content features.

B.5. Timestep result difference analysis

MaskUNet selectively generates key structures earlier (red boxes), while SD 1.5 uniformly refines the entire image (blue boxes), see Figure 2 (e).

B.6. Efficiency analysis

The Table 3 shows that while training-free incurs higher latency (66.82s), its computational overhead remains minimal (+0.77 GFLOPS). Training achieves a trade-off, with a moderate latency increase (2.99s vs. 2.14s for SD 1.5) while keeping efficiency practical.

Figure 3. Quality results compared to other methods.

Table 3. Efficiency analysis.

Method	Add Param	GFLOPS	Latency (s)
SD 1.5	0	156.00	2.14
Training	55.06M	156.03	2.99
Training-free	49.03M	156.77	66.82

B.7. More comparative results

The random mask can improve results in certain cases but has worse overall performance (FID 270.22), which motivated us to propose two more stable and controllable masking strategies (Table 4, third-to-last row). We did not apply dropout during FT. Thanks for the suggestion. We conduct Full FT with dropout, it degraded performance (Table 4, second-to-last row).

Table 4. More comparative results.

Method	FID	CLIP
Full Fine-tune	24.45	0.33
Random Mask	270.22	0.06
Dropout (ratio=0.1)	26.56	0.32
MaskUNet	**21.88**	0.33

B.8. Zero-out strategy in DiT-based models

Table 5. More comparative results.

ImageNet 256 5k		COCO 2017 5k			Geneval
Model	FID	Model	FID	CLIP	Overall
DiT	17.07	PixArt-alpha	39.40	0.33	0.48
MaskUnet	**16.15**	MaskUnet	**37.83**	**0.33**	**0.53**

The Table 5 demonstrates that our method is also applicable to DiT-based models.

B.9. Text-to-video for More Visualization Results

Additional visualization results of zero-shot generation are shown in the Figure 3. Figure 5 is the complete visualization result of Text2Video-Zero [5]. It clearly demonstrates that MaskUNet generates videos with greater temporal continuity and semantic consistency, validating its effectiveness in video generation.

B.10. User study details

The study participants consisted of 26 volunteers from our university. The questionnaire comprised 46 questions, each presenting several images: one generated by our method, MaskUNet, and the others generated by alternative methods (Dreambooth [9], Textual Inversion [1], Reversion [3], Text2Video-zero [5], SynGen [7], LoRA [2], SD [8], etc.). An example of the questionnaire is shown in the Figure 1.

References

[1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[2] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

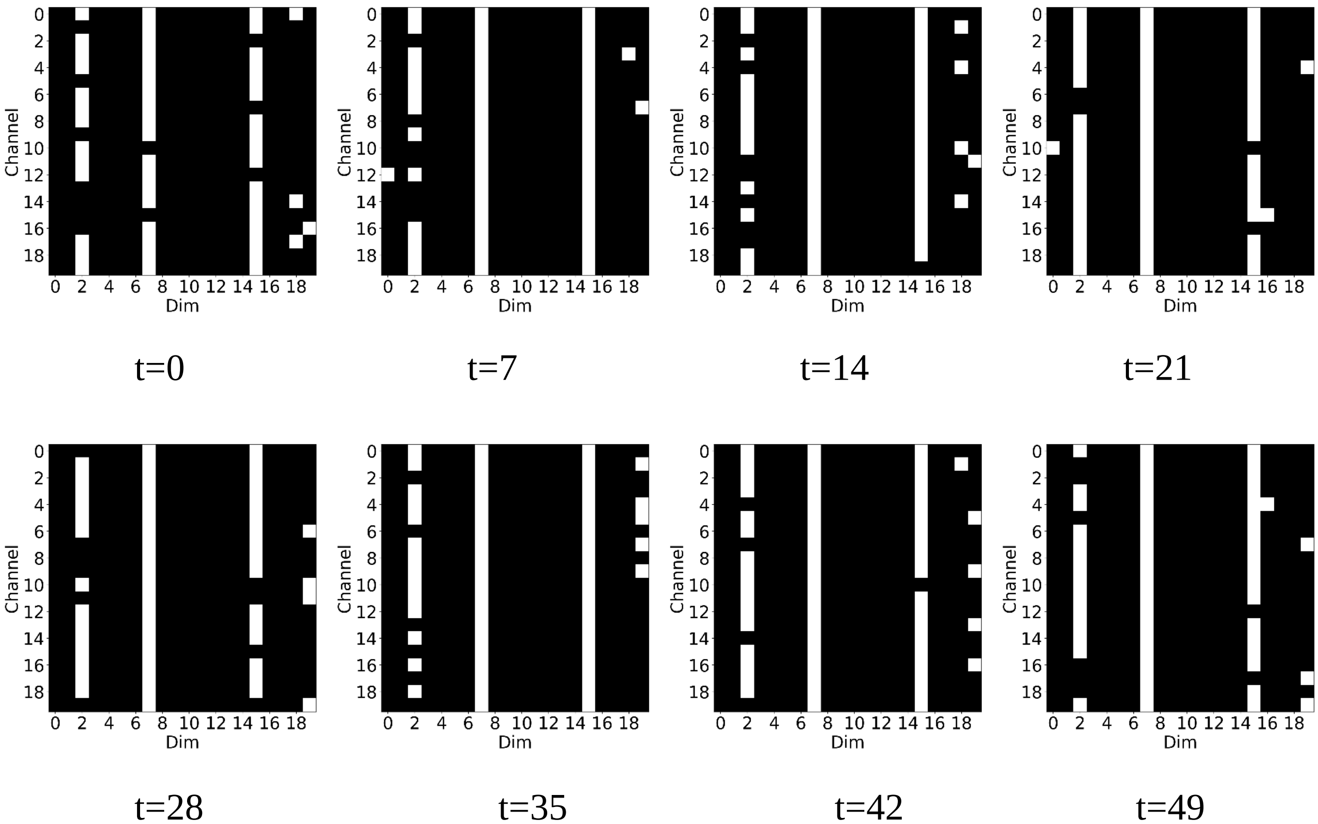

Figure 4. Visualization of mask position at different time steps.

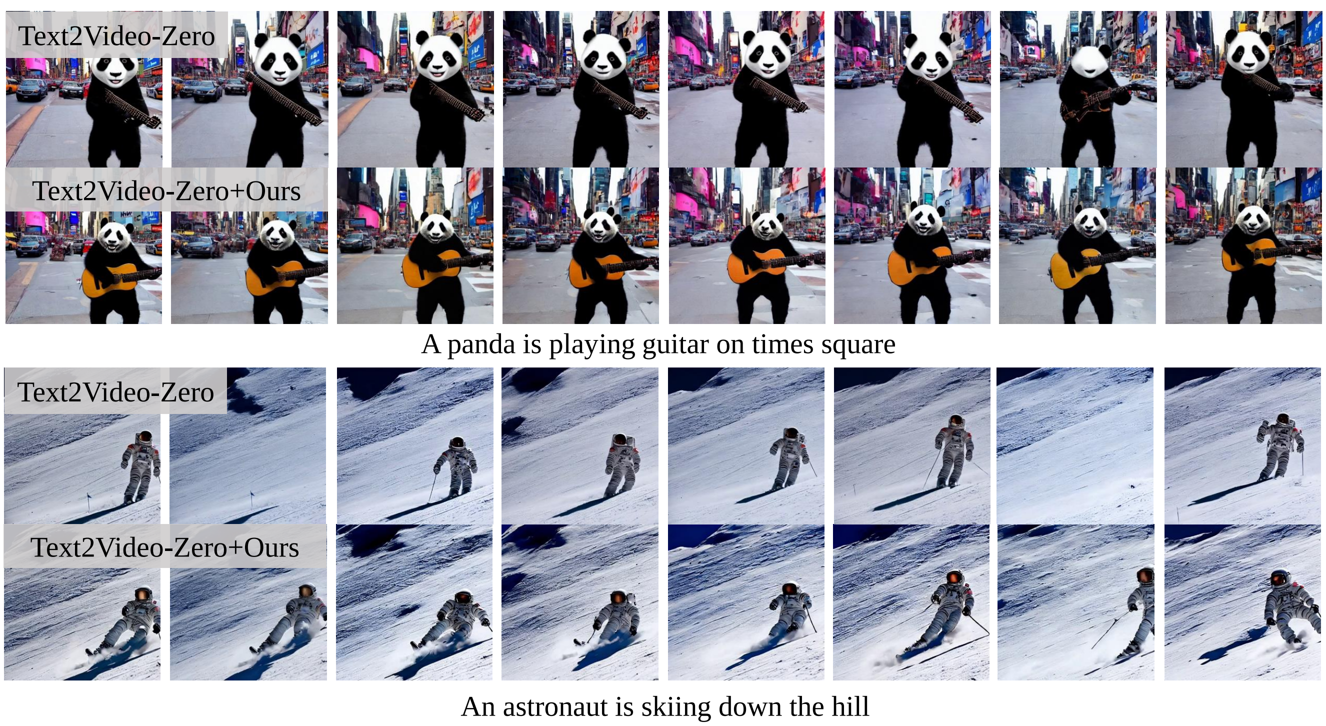

Figure 5. Quality results by Text2Video-Zero [5] with or without mask.

[3] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 3

[4] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 1

[5] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 3, 4

[6] Senmao Li, Taihang Hu, Fahad Shahbaz Khan, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of unet encoder in diffusion models. *arXiv e-prints*, pages arXiv–2312, 2023. 1

[7] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[10] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 1

[11] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 1

[12] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1