MaskFocus: Focusing Policy Optimization on Critical Steps for Masked Image Generation

Supplementary Material

1. Detailed Experiment Setup

1.1. Benchmark

HPS benchmark [7]. This is a large-scale image-text dataset constructed based on human preferences. We use the prompts from the training set to train models to generate high-quality images that align with human preferences. These prompts possess sufficient complexity to effectively capture differences in model performance on preference-oriented tasks.

GenEval benchmark [1]. This benchmark evaluates the performance of T2I models on compositional tasks, like object counting, spatial relations, and attribute binding—across six increasingly difficult compositional image generation tasks. We use this benchmark to evaluate the compositional image generation capabilities of our models. In addition, we use this evaluation pipeline as our reward model, following Flow-GRPO [3].

DrawBench [6]. This is a comprehensive and challenging text-to-image model evaluation benchmark that allows for a deeper assessment and comparison of different models. We use it to test prompts, including metrics such as PickScore and DEQA.

T2I-CompBench [2]. This is a comprehensive benchmark designed to evaluate model performance in compositional text-to-image generation within open-world scenarios. It includes several specially designed evaluation metrics, covering attribute binding, spatial relations, and complex compositions.

1.2. Detailed implement

1.2.1. Critical Step Selection

(1) During each image generation sampling process, the intermediate image embeddings at each sampling step are recorded. (2) The cosine similarity between each intermediate embedding and the final generated image embedding is calculated, yielding a similarity score for each step. (3) The change in similarity between adjacent sampling steps (the absolute value of the difference) is calculated as the information gain metric for that step. (4) All sampling steps are ranked according to their information gain scores, and the Top-K steps with the highest scores are selected as critical steps. (5) In the RL optimization process, for these critical steps, firstly, some mask locations are randomly generated according to the mask schedule of these steps. Then, the images generated during the sampling process are masked and then fed into the model and the reference model to predict the token log-probability at these mask locations. Finally, the RL training loss can be calculated, and policy optimization can be performed.

1.2.2. Dynamic Routing Sampling

(1) Before each round of sampling, obtain the token distribution probability of all samples in the same group and calculate the entropy value of each sample. (2) Samples are sorted from the highest to the lowest entropy value. (3) Half of the samples with high entropy values (high uncertainty, the model is not confident) are assigned to the "Exploitation" branch. In this branch, token sampling is performed according to the highest confidence strategy to maintain model stability. (4) Half of the samples with low entropy values (the model is highly certain) are assigned to the "Exploration" branch. In this branch, the sampling temperature at each token is dynamically adjusted according to its entropy value, appropriately increasing the sampling perturbation to encourage the model to explore. (5) Each round of sampling automatically completes routing based on the current group entropy value, without the need for manual group specification. (6) The sampling results from both branches are concatenated and then used for subsequent RL policy learning and optimization.

1.3. Training Hyperparameters

Table 1. MaskFocus training hyperparameters.

Name	MaskFocus
Learning rate	2e-6
Beta β	0.01
Group Size G	8
Classifier-Free Guidance Scale	5
Max Gradient Norm	1.0
Batchsize	1
Training Steps	1200
Gradient Accumulation Steps	1
Image Resolution $h \times w$	1024×1024

2. More Discussion

2.1. Comparison of different selection strategies

In the main text, we conduct a comprehensive analysis of the performance of various selection strategies. To fur-

ther demonstrate the advantages of our critical-step selection strategy, we provide visual validation as shown in the figure. Specifically, when 40% of the first tokens are randomly selected during the later stages, we observe evident reward hacking phenomena and a pronounced decline in image quality. Alternatively, selecting the earlier 20% of steps helps alleviate this issue to some extent; however, the resulting visual outcomes remain inferior compared to those produced by our proposed approach. These results clearly highlight the effectiveness and superiority of our critical-step selection strategy in producing high-quality images while avoiding unintended reward hacking behaviors.

2.2. Ablation on critical steps K

We set the number of training steps to 6 to achieve a trade-off between efficiency and model fidelity, as shown in Table 2. Our method outperforms random selection from the entire sequence. Although increasing K enhances performance, it incurs higher computational overhead. However, further increasing K to 32 or 64 results in negative gains. We attribute this to the over-optimization of fine-grained details and approximation errors for importance sampling.

Table 2. Ablation study on the number of critical steps.

Critical Steps	Geneval↑	DEQA↑	PickScore↑	Time↓
K=3	0.71	4.29	22.22	**2.5h**
K=6 (Random)	0.72	4.30	22.24	2.8h
K=6 (Our)	**0.76**	4.39	22.39	2.8h
K=10	**0.76**	**4.42**	**22.42**	3.4h
K=32	0.70	4.25	22.19	6.7h
K=64 (Full-trajectory)	0.66	4.20	22.08	9.5h

2.3. Difference from previous methods.

The core distinction between our approach and previous methods lies in our ability to dynamically identify critical steps during the sampling process and optimize policies specifically for these steps—an aspect that has not been effectively addressed in prior research. Unlike the strategy of selecting all steps, which maximizes information utilization but imposes considerable memory and inference burdens, our method selectively focuses on the most critical steps, thereby maintaining high efficiency and practical feasibility. On the other hand, approaches that simply pre-select a fixed portion of steps fail to account for the varying contributions of different steps, often resulting in suboptimal results. By targeting the selection of critical steps, our method strikes a better balance between performance and effectiveness, achieving improved sample quality with manageable resource consumption.

Mask-GRPO [4] reformulates the sampling trajectory as a multi-step decision-making problem, sharing certain similarities withRL-based strategies in diffusion models. Furthermore, Mask-GRPO emphasizes per-step log-probability analysis and posits that predicting these masked tokens in subsequent steps is more valuable. Our approach, in contrast, is simpler and aligns the loss computation objective with that used during pre-training by optimizing the log-probabilities for all masked tokens. Importantly, our automatic selection of critical steps also reduces unnecessary computation and enhances inference efficiency.

MaskGRPO [5] is specifically designed for multimodal discrete diffusion models. While our masking strategy is consistent with theirs, our primary difference lies in the critical step selection and sampling policy. MaskGRPO points out that MaskGiT's sampling strategy tends to generate overly smooth images with missing details. Therefore, it proposes a token emerge sampling strategy. Our method uses entropy-based dynamic sampling to balance exploration and exploitation during reinforcement learning training. Additionally, MaskGRPO does not differentiate between the impact of individual steps on output quality, which limits its ability to further improve performance. In contrast, our method leverages critical-step selection to deliver higher image quality and greater inference efficiency.

3. More Visual Comparison Results

We present additional visual examples in Figure 1, which clearly showcase the outstanding performance of our method. Our technique leads to substantial improvements not only in the visual quality and diversity of the generated images, but also in their structural coherence and richness of details. The images produced by our method exhibit vibrant colors, sharper features, and more realistic textures, setting a new benchmark for fidelity and expressiveness in Meissonic generation. These results convincingly demonstrate the superiority and effectiveness of our approach.

References

[1] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 1

[2] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 1

[3] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 1

[4] Yifu Luo, Xinhao Hu, Keyu Fan, Haoyuan Sun, Zeyu Chen, Bo Xia, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement learning meets masked generative models: Mask-grpo for text-to-image generation. *arXiv preprint arXiv:2510.13418*, 2025. 2

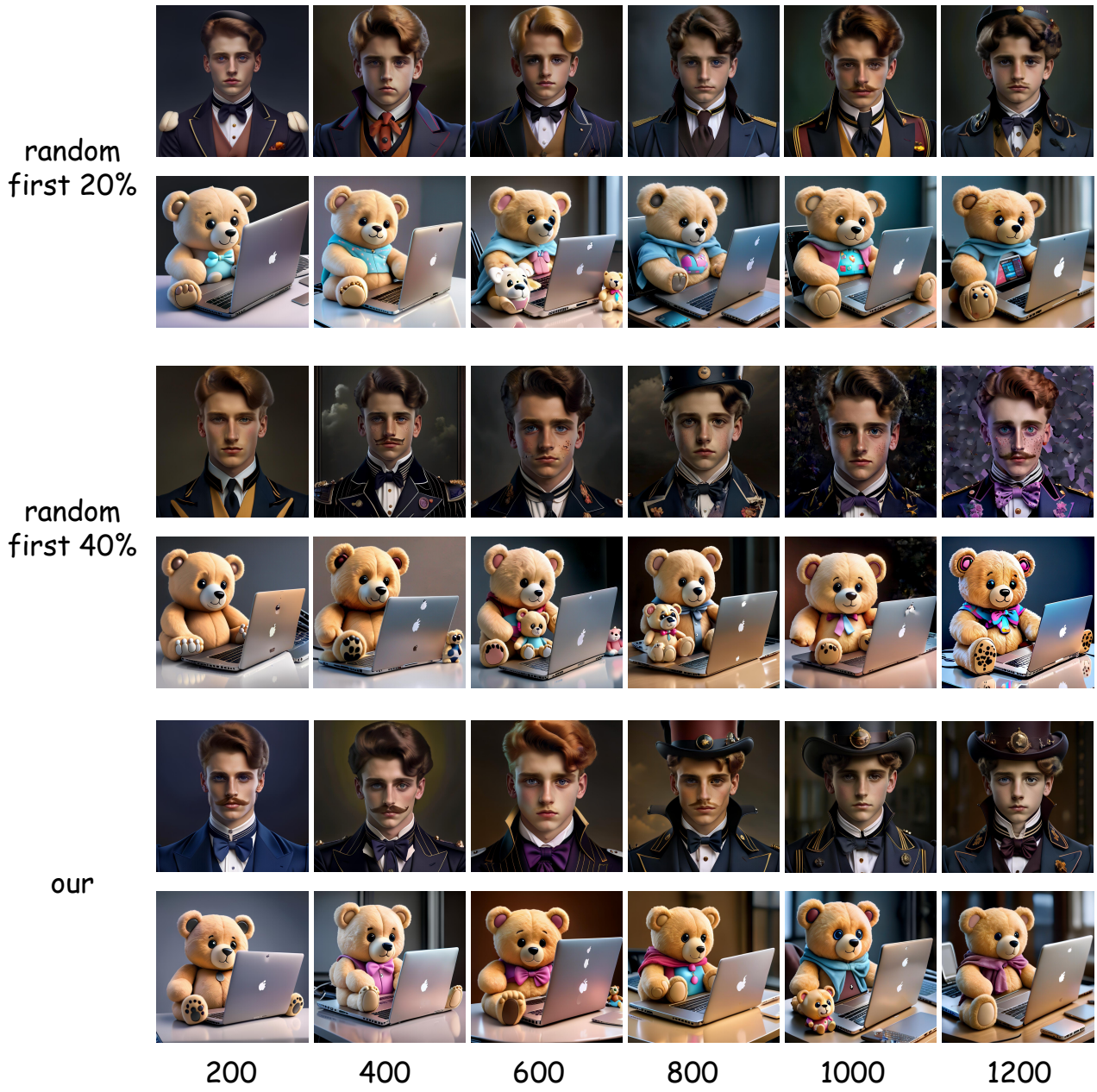

Figure 1. **Visual Comparison.** These prompts are sampled from the HPS evaluation dataset. Our method does not exhibit reward hacking and achieves the best image quality and aesthetic quality.

[5] Tianren Ma, Mu Zhang, Yibing Wang, and Qixiang Ye. Consolidating reinforcement learning for multimodal discrete diffusion models. *arXiv preprint arXiv:2510.02880*, 2025. 2

[6] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[7] 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

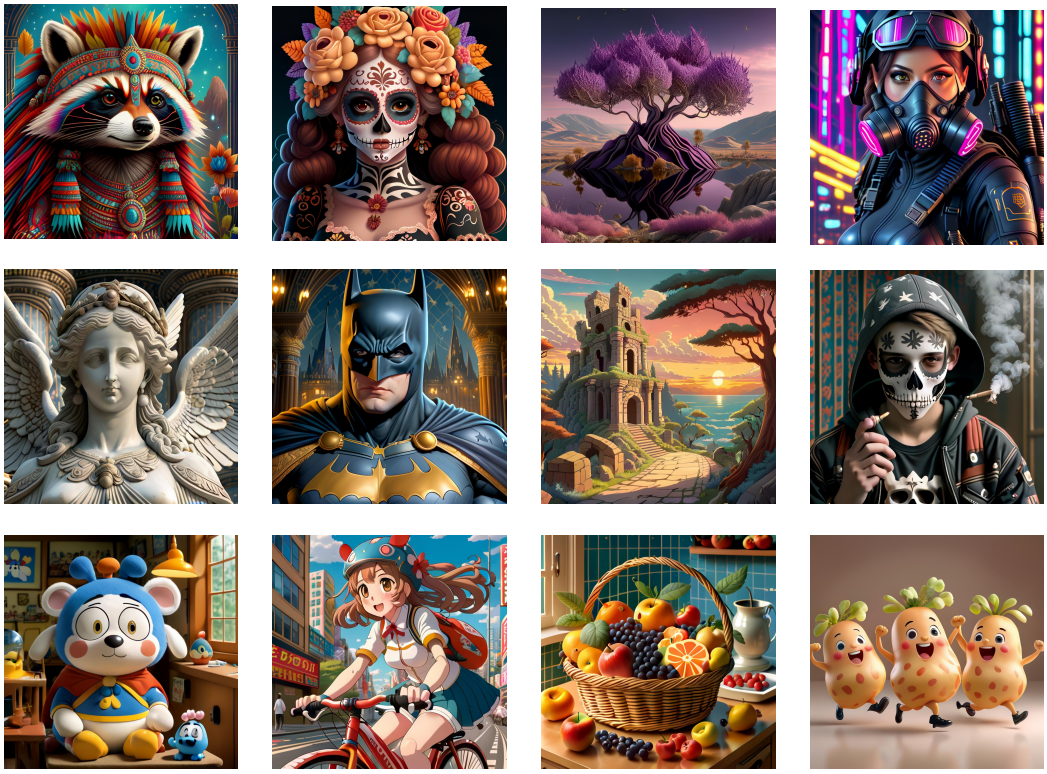

Figure 2. **Visual Comparison on HPS benchmark.**

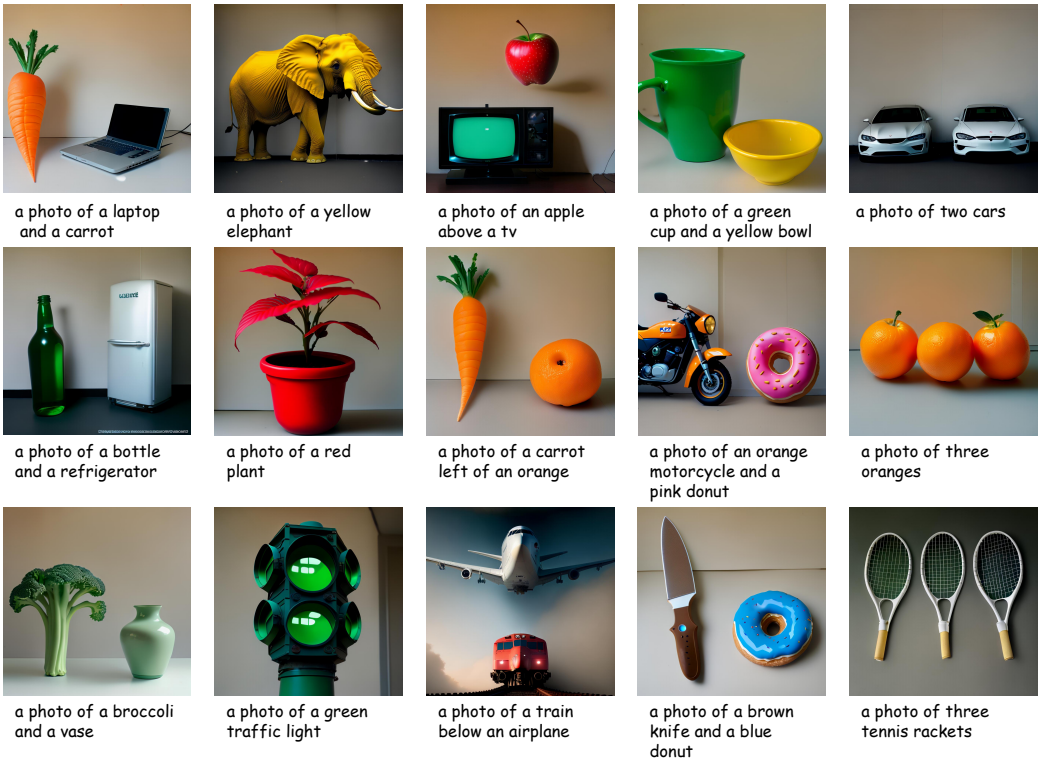

Figure 3. **Visual Comparison on GenEval benchmark.**