Performance Plateaus in Inference-Time Scaling for Text-to-Image Diffusion Without External Models

Changhyun Choi [1] **Sungha Kim** [1] **H. Jin Kim** [1 2 3]

Abstract

Recently, it has been shown that investing computing resources in searching for good initial noise for a text-to-image diffusion model helps improve performance. However, previous studies required external models to evaluate the resulting images, which is impossible on GPUs with small VRAM. For these reasons, we apply Best-of-N inference-time scaling to algorithms that optimize the initial noise of a diffusion model without external models across multiple datasets and backbones. We demonstrate that inference-time scaling for text-to-image diffusion models in this setting quickly reaches a performance plateau, and a relatively small number of optimization steps suffices to achieve the maximum achievable performance with each algorithm.

1. Introduction

In large language models, an inference-time scaling method has attracted a lot of attention because it has been discovered that this can improve performance without increasing the model size (OpenAI, 2024; Guo et al., 2025; Zhang et al., 2025). Therefore, studies have recently been conducted on whether inference-time scaling can also be applied to diffusion models (Ma et al., 2025; Li et al., 2025; Zhuo et al., 2025). An influential study (Ma et al., 2025) in this area demonstrated that allocating more compute to search for a better initial noise helps to improve performance on text-to-image (T2I) tasks. Despite these advances, these studies all rely on additional vision-language models (VLMs) or other external models to evaluate the quality of the generated images. This makes it difficult to implement on consumer-grade GPUs typically found in personal desktops rather than in well-funded research laboratories or enterprises.

Independently, several approaches (Meral et al., 2024; Guo et al., 2024; Qiu et al., 2024) have been proposed to optimize the initial noise and intermediate outputs of the denoising process at inference time so that the generated image aligns better with the input prompt. These methods do not require additional models; instead, they optimize the noise using only a pre-trained T2I diffusion model. In particular, the authors of InitNO (Guo et al., 2024) argue that not every noise sample drawn from a standard normal distribution precisely adheres to a given text prompt, implying the existence of both valid and invalid noise. This raises an important question:

When selecting good initial noise solely with the T2I diffusion model,
does applying inference-time scaling help improve performance?

We perform extensive experiments across various combinations of algorithms and models at multiple scales and reveal a clear performance plateau as we invest more computational resources, which contradicts both the default hyperparameter settings and common expectations. Based on these findings, we show how much computational effort should ideally be invested without relying on additional models when performing T2I tasks on GPUs with small VRAM. Furthermore, we show that the state-of-the-art (SOTA) algorithm among these initial noise optimization algorithms changes depending on the underlying T2I diffusion model, thereby indicating opportunities for future research.

[1]Interdisciplinary Program in Artificial Intelligence, Seoul National University [2]Aerospace Engineering, Seoul National University [3]ASRI, AIIS, Seoul National University. Correspondence to: H. Jin Kim <hjinkim@snu.ac.kr>.

Accepted at Methods and Opportunities at Small Scale (MOSS), ICML 2025, Vancouver, Canada.

2. Preliminaries

Stable Diffusion Model. In a T2I task, a generative model receives a text prompt as input and generates an image that precisely matches the prompt. One of the most widely used models in this task is the Stable Diffusion model (SD) (Rombach et al., 2022). SD belongs to the family of latent diffusion models and therefore operates within the latent space of an autoencoder. Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) runs with a UNet (Ronneberger et al., 2015) backbone (for both SD1 and SD2) in this latent space. In the denoising process, SD gradually denoises the initial noise sampled from a standard normal distribution. If the initial noise differs, the generated image is also different.

For T2I tasks, SD incorporates cross-attention layers. The text encoder of SD transforms the input text prompt into a sequence of embeddings, which are then projected to serve as keys, while the UNet's intermediate features undergo a similar projection to serve as queries. In addition, SD includes self-attention layers, in which the UNet's intermediate features serve as both keys and queries. The resulting cross-attention maps and self-attention maps derived from these key-query pairs are used by all the algorithms examined in this paper.

3. Best-of-N Inference-Time Scaling

3.1. Metrics for the given initial noise

In this study, we apply Best-of-N inference-time scaling to the T2I diffusion model by sampling multiple initial noises and selecting the best one. This is similar to one of the methods of the previous study (Ma et al., 2025) which applied inference-time scaling to the T2I diffusion model with an additional verifier model. However, we evaluate the quality of a given initial noise with loss functions of various algorithms that optimize the initial noise of diffusion models. CONFORM (Meral et al., 2024) utilizes the cross-attention maps of T2I diffusion models for contrastive learning. It employs InfoNCE (Oord et al., 2018) as its loss, grouping the objects and attributes within the input text prompt into positive and negative pairs. InitNO (Guo et al., 2024) uses the sum of two metrics as its loss. First, it considers $1 - \textbf{\textit{minmax\_cross}}$, where ***minmax_cross*** is the smallest value among the maximum cross-attention weight values for each object that must appear in the resulting image. Second, InitNO measures the degree of overlap between self-attention maps corresponding to the spatial patch with the largest cross-attention weight of each object. Although the paper formally includes a regularization term to prevent deviations from the standard normal distribution for every optimization step, the official implementation[1] uses this term in optimization process only when it exceeds a certain threshold. Since we sample noises from a standard normal distribution, we do not consider this regularization term. The loss of Self-Cross guidance (Qiu et al., 2024) is largely similar to that of InitNO. The key difference is that it does not focus solely on the self-attention map of the patch with the largest cross-attention weight; instead, it uses the entire patches' self-attention maps with each of their cross-attention weights when computing overlaps between objects. The authors propose to use InitNO before Self-Cross guidance.

3.2. Best-of-N

To examine whether allocating more compute is beneficial for initial noise optimization algorithms that rely solely on a T2I diffusion model, we adopt a Best-of-N approach to the number of candidate initial noises in each algorithm. We allocate a total of N loss calculations, assess each candidate noise using these loss values, and select the noise achieving the best (lowest) loss. Note that unlike the other two algorithms, InitNO (Guo et al., 2024) optimizes a single initial noise multiple times. Concretely, InitNO first samples an initial noise and runs up to 10 optimization steps, checking at each step whether the resulting noise is valid, i.e., the loss value falls under the predefined threshold. If the noise becomes valid at any step, the optimization is halted immediately and that noise is used as the initial noise for the diffusion model. If it remains invalid after all 10 steps, a new noise is sampled, and the process is repeated up to 5 times. If all 5 trials are invalid, InitNO picks the noise that has the smallest loss among them, optimizes it for more 40 iterations, and ultimately uses that result as the initial noise for the diffusion model. As a result, in the case of InitNO, each noise candidate is optimized over 10 loss calculations (without early stopping for fairness), so the effective number of candidates is $\frac{N}{10}$. In contrast, both CONFORM (Meral et al., 2024) and Self-Cross guidance (Qiu et al., 2024) retain N candidates. As N increases, the number of candidate initial noises will also increase. If the loss function of each algorithm accurately reflects the quality of the resulting image, the performance will always increase monotonically as the candidate initial noise increases. And if this is true, we can conclude that applying inference-time scaling to algorithms that only use the soley T2I diffusion model helps improve performance. We conduct extensive experiments to confirm this.

[1] https://github.com/xiefan-guo/initno/tree/main

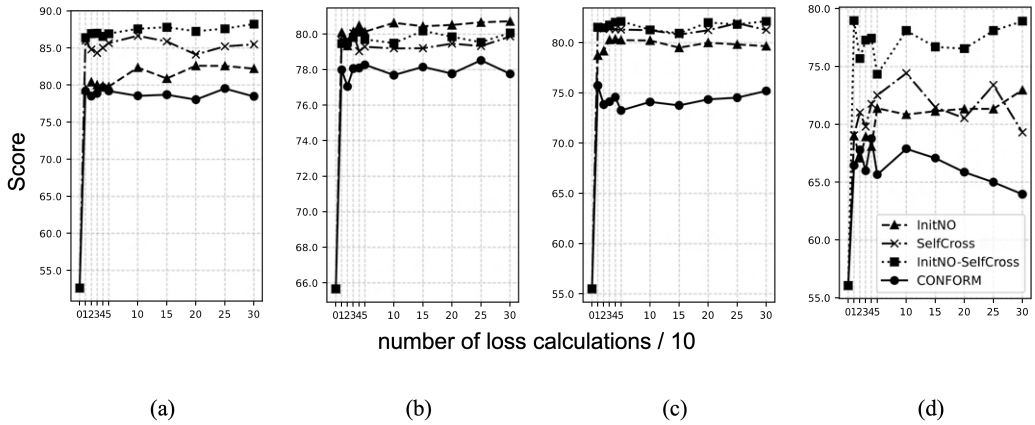


Figure 1. **Experimental results in SD1.5.** We conducted experiments on four datasets in SD1.5: (a) *animal_animal*, (b) *animal_object*, (c) *object_object*, (d) *similar_subjects*. In each graph, the horizontal axis (x-axis) represents the number of loss calculations N divided by 10 ($N/10$), while the vertical axis (y-axis) shows the resulting scores. The legend in (d) lists the names of each algorithm evaluated.

4. Experiments

4.1. Experimental settings

We followed the experimental settings described in the Self-Cross guidance paper (Qiu et al., 2024). The input text prompts are categorized into four datasets: (1) *animal_animal*, consisting of 66 prompts featuring various pairs of animals; (2) *animal_object*, consisting of 144 prompts describing diverse animal–object pairs; (3) *object_object*, consisting of 66 prompts containing two distinct subjects; and (4) *similar_subjects*, consisting of 31 prompts illustrating two similar subjects. For the score calculation, we use GPT4o (Achiam et al., 2023) and predefined questions introduced in the Self-Cross guidance paper to determine whether each resulting image (i) includes both subjects (*Existence*), (ii) presents both subjects in a recognizable form (*Recognizability*), and (iii) does not exhibit any undesired mixture between the two subjects (*Not a Mixture*). The percentage of the positive responses from GPT4o was calculated as a score. All experiments were repeated with 10 different random seeds to ensure reliability and statistical robustness of the reported results. In our experiments, we used the official implementation of Self-Cross guidance[2], which also includes implementations of InitNO (Guo et al., 2024) and CONFORM (Meral et al., 2024). Since both the InitNO's and Self-Cross guidance's official implementations note that InitNO does not work well on SD2.1 and we also confirmed this in our own tests, we did not use InitNO for SD2.1. Additionally, to investigate how much InitNO contributes to the performance of Self-Cross guidance, we distinguish between the version that uses InitNO first (InitNO-SelfCross) and the one that does not (SelfCross).

4.2. Results

Figure 1 and 2 present the main experimental results, while more detailed quantitative and qualitative findings are provided in the Appendix. When the number of loss calculations N is set to 0, the result corresponds to the default SD's output without any additional algorithms. For SD1.5, InitNO-SelfCross (Qiu et al., 2024) performs best overall, as this is the last proposed algorithm. For SD2.1, CONFORM (Meral et al., 2024) is generally the top-performing method in most cases, even though this is the earliest algorithm among the algorithms used in this study.

As we increase the number of candidate initial noises, we can see the result that contradicts the common expectation. **In multiple backbones, various algorithms do not guarantee an increase in performance as candidate initial noise increases**. Although the default setting for InitNO (Guo et al., 2024) is $N = 50$ and the general expectation is that larger N should offer better results, all initial noise optimization algorithms do not show a consistent performance increase beyond $N = 10$. In other words, for InitNO, using just a single noise sample and optimizing it suffices to achieve the maximum possible performance with the algorithm and the other two algorithms also show similar patterns.

[2] https://github.com/mengtang-lab/selfcross-guidance/tree/main

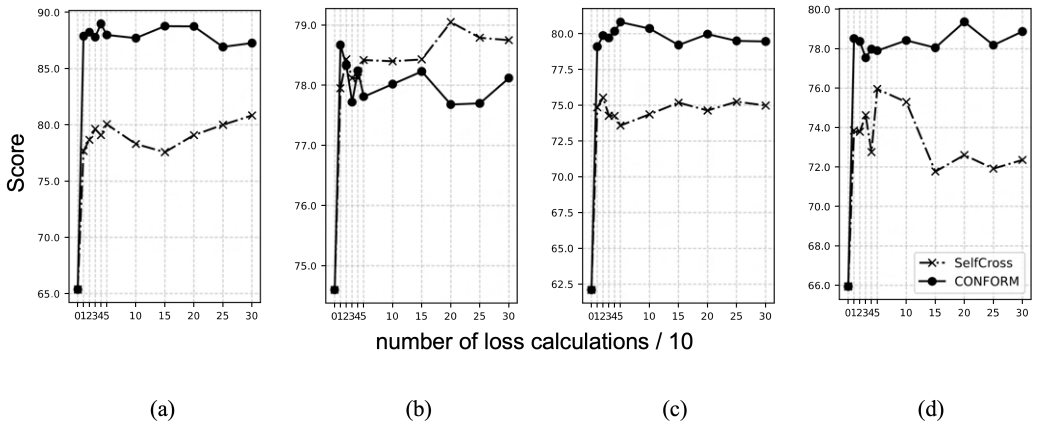


Figure 2. **Experimental results in SD2.1.** We conducted experiments on four datasets in SD2.1: (a) *animal_animal*, (b) *animal_object*, (c) *object_object*, (d) *similar_subjects*. In each graph, the horizontal axis (x-axis) represents the number of loss calculations N divided by 10 ($N/10$), while the vertical axis (y-axis) shows the resulting scores. The legend in (d) lists the names of each algorithm evaluated.

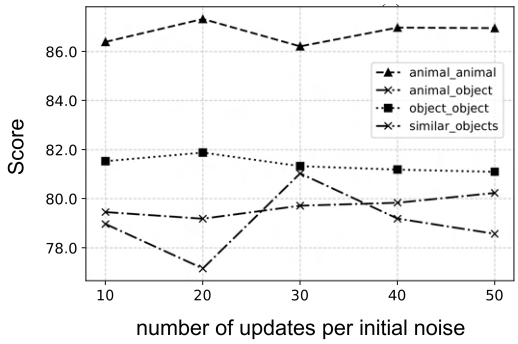


Figure 3. **Experimental results in InitNO.** We conducted experiments on four datasets in SD2.1. In each graph, the horizontal axis (x-axis) represents the number of updates per initial noise, while the vertical axis (y-axis) shows the resulting scores. The legend lists the names of the prompt datasets.

These findings suggest that **the losses employed by current initial noise optimization algorithms do not perfectly capture the alignment between the input text prompt and the resulting image**. To examine this point more closely, we varied the number of optimization iterations in the InitNO phase that is conducted first in the Self-Cross guidance. We retained the original setting of 5 candidate initial noise samples but increased the optimization steps for each sample from 10 to 50 gradually. As Figure 3 shows, more optimization steps for the initial noise is not helpful for the performance. This can be interpreted, like over-optimization (Zhang et al., 2024; Kim et al., 2025), as the result of excessively optimizing a loss function that fails to accurately capture the true objective.

We confirmed that the performance does not increase even if we sample more random noises and select the noise with the smallest loss among them, or further optimize using the loss of the sampled noise. Nevertheless, the performance clearly increased compared to not using any of these algorithms. Therefore, we find that it is better to optimize the initial noise with the least computational resource (but not zero) for GPUs with limited VRAM.

5. Conclusion

In this study, we found that when we only have VRAM-limited GPUs, the most efficient way to improve the performance of the T2I diffusion model is to optimize the initial noise with the least computational resources. In addition, we can see that existing training-free approaches for initial noise optimization have considerable potential for further improvement.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Guo, X., Liu, J., Cui, M., Li, J., Yang, H., and Huang, D. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9380–9389, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kim, S., Kim, M., and Park, D. Test-time alignment of diffusion models without reward over-optimization. In *Proceedings of the International Conference on Learning Representations*, 2025.

Li, S., Kallidromitis, K., Gokul, A., Koneru, A., Kato, Y., Kozuka, K., and Grover, A. Reflect-dit: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. *arXiv preprint arXiv:2503.12271*, 2025.

Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.-C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Meral, T. H. S., Simsar, E., Tombari, F., and Yanardag, P. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9005–9014, 2024.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Learning to reason with llms, September 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

Qiu, W., Wang, J., and Tang, M. Self-cross diffusion guidance for text-to-image synthesis of similar subjects. *arXiv preprint arXiv:2411.18936*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Zhang, Q., Lyu, F., Sun, Z., Wang, L., Zhang, W., Guo, Z., Wang, Y., King, I., Liu, X., and Ma, C. What, how, where, and how well? a survey on test-time scaling in large language models. *arXiv preprint arXiv:2503.24235*, 2025.

Zhang, Z., Zhang, S., Zhan, Y., Luo, Y., Wen, Y., and Tao, D. Confronting reward overoptimization for diffusion models: A perspective of inductive and primacy biases. In *Proceedings of the International Conference on Machine Learning*, pp. 60396–60413. pmlr, 2024.

Zhuo, L., Zhao, L., Paul, S., Liao, Y., Zhang, R., Xin, Y., Gao, P., Elhoseiny, M., and Li, H. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. *arXiv preprint arXiv:2504.16080*, 2025.

Appendix

Table 1, 2, 3 and 4 present the quantitative results corresponding to Figure 1 and 2. Table 5 shows the qualitative outcomes for the input text prompt "*a bear and a red balloon*".

Table 1. **Quantitative results for each initial noise optimization algorithm for *animal_animal* dataset.** N represents the number of loss calculations. The numbers in the parentheses indicate the versions of Stable Diffusion models.

N	CONFORM(1.5)	InitNO(1.5)	SelfCross(1.5)	InitNO-SelfCross(1.5)	CONFORM(2.1)	SelfCross(2.1)
0	52.64	52.64	52.64	52.64	65.38	65.38
10	79.22	79.41	86.00	86.39	87.88	77.68
20	78.53	80.41	84.81	86.95	88.23	78.67
30	78.89	80.02	84.35	86.99	87.79	79.64
40	79.69	79.88	85.04	86.58	88.98	79.09
50	79.22	79.79	85.67	86.93	87.99	80.06
100	78.56	82.36	86.65	87.58	87.71	78.31
150	78.72	80.89	85.89	87.80	88.77	77.58
200	78.05	82.60	84.13	87.25	88.75	79.09
250	79.55	82.58	85.22	87.60	86.93	80.01
300	78.48	82.21	85.50	88.23	87.27	80.84

Table 2. **Quantitative results for each initial noise optimization algorithm for *animal_object* dataset.** N represents the number of loss calculations. The numbers in the parentheses indicate the versions of Stable Diffusion models.

N	CONFORM(1.5)	InitNO(1.5)	SelfCross(1.5)	InitNO-SelfCross(1.5)	CONFORM(2.1)	SelfCross(2.1)
0	65.66	65.66	65.66	65.66	74.60	74.60
10	77.99	80.10	79.65	79.46	78.67	77.95
20	77.05	79.34	79.71	79.54	78.33	78.43
30	78.06	80.14	80.05	79.81	77.72	78.12
40	78.09	80.48	79.04	80.11	78.24	78.12
50	78.27	80.13	79.29	79.70	77.81	78.42
100	77.69	80.63	79.20	79.49	78.02	78.40
150	78.15	80.44	79.21	80.18	78.23	78.43
200	77.77	80.51	79.47	79.85	77.68	79.05
250	78.52	80.66	79.33	79.54	77.70	78.79
300	77.76	80.72	79.87	80.07	78.12	78.75

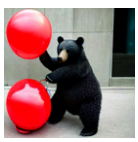
Table 3. **Quantitative results for each initial noise optimization algorithm for *object_object* dataset.** N represents the number of loss calculations. The numbers in the parentheses indicate the versions of Stable Diffusion models.

N	CONFORM(1.5)	InitNO(1.5)	SelfCross(1.5)	InitNO-SelfCross(1.5)	CONFORM(2.1)	SelfCross(2.1)
0	55.50	55.50	55.50	55.50	62.10	62.10
10	75.72	78.72	81.58	81.53	79.09	74.85
20	73.84	79.17	81.40	81.49	79.87	75.53
30	74.13	80.28	81.37	81.74	79.70	74.26
40	74.57	80.35	81.32	82.04	80.17	74.25
50	73.25	80.25	81.31	82.12	80.81	73.58
100	74.10	80.21	81.25	81.27	80.36	74.36
150	73.75	79.50	80.81	80.92	79.21	75.18
200	74.35	79.99	81.24	81.99	79.97	74.63
250	74.52	79.82	81.95	81.81	79.50	75.24
300	75.19	79.65	81.27	82.12	79.46	74.98

Table 4. **Quantitative results for each initial noise optimization algorithm for *similar_subjects* dataset.** N represents the number of loss calculations. The numbers in the parentheses indicate the versions of Stable Diffusion models.

N	CONFORM(1.5)	InitNO(1.5)	SelfCross(1.5)	InitNO-SelfCross(1.5)	CONFORM(2.1)	SelfCross(2.1)
0	56.06	56.06	56.06	56.06	65.94	65.94
10	66.44	69.00	69.06	78.97	78.52	73.85
20	67.78	67.09	71.03	75.69	78.37	73.79
30	65.99	68.92	69.80	77.28	77.54	74.63
40	68.77	68.08	71.77	77.44	77.98	72.76
50	65.65	71.38	72.52	74.33	77.90	75.96
100	67.89	70.84	74.43	78.10	78.42	75.30
150	67.08	71.13	71.48	76.70	78.05	71.77
200	65.87	71.33	70.54	76.55	79.36	72.61
250	65.00	71.33	73.40	78.10	78.18	71.92
300	63.96	72.96	69.31	78.92	78.87	72.36

Table 5. **Resulting images for each initial noise optimization algorithm.** Input text prompt is "*a bear and a red balloon*". N represents the number of loss calculations. The numbers in the parentheses indicate the versions of Stable Diffusion models.

N	CONFORM(1.5)	InitNO(1.5)	SelfCross(1.5)	InitNO-SelfCross(1.5)	CONFORM(2.1)	SelfCross(2.1)
0						
10						
20						
30						
40						
50						
100						
150						
200						
250						
300						