Manipulation of noise in generative modeling

Artem Serebriakov
Technical University of Munich

Abstract

Although the field of Text-to-Image synthesis has made significant progress, challenges related to adhering to prompts and maintaining visual quality persist. Reward-based Noise Optimization (ReNO) offers a novel approach to address these challenges by iteratively optimizing the initial noise vector at inference time using pre-trained human-preference reward models. This method avoids model fine-tuning, thereby maintaining computational efficiency and mitigating the risk of "reward hacking". The review focuses on the ReNO framework, discussing the distillation techniques and reward models employed within it, and also highlighting other noise optimization methods. Furthermore, this paper extends its original findings through additional experiments. Metrics such as L2 distances, coherence analysis, and visualizations of noise differences were used to analyze the noise optimization in ReNO. Experimental results demonstrate that ReNO improves prompt fidelity, compositional complexity and visual aesthetics across multiple one-step diffusion models, reaching the performance of the state-of-the-art proprietary models. The research on the noise optimization process indicates ReNO achieves these results with minimal modifications to the initial noise while preserving high-frequency details and refining broader structural patterns. The findings demonstrate the significant impact of noise optimization techniques on generative modeling and point to exciting opportunities for future research to further advance text-to-image synthesis.

1 Introduction

Generative modeling, particularly in the domain of Text-to-Image (T2I) synthesis, has made remarkable advancements in recent years. Powered by the availability of large-scale image-text datasets [1, 2] and the evolution of deep generative models, mainly diffusion models [3, 4], these systems are now capable of producing highly detailed and diverse images from textual descriptions. However, significant challenges persist, particularly in ensuring the generated images accurately reflect the complexity of the input prompts. Problems such as incorrect attribute binding, compositional errors, and visual artifacts continue to hinder their performance, especially when dealing with intricate or abstract prompts [5].

One of the key factors influencing the quality and prompt adherence of T2I models is the initialization of the noise vector, which serves as the starting point for the generation process. Noise plays a critical role in diffusion models, as it determines the trajectory of the denoising process and, eventually, the characteristics of the generated image. Although the significance of noise manipulation for enhancing generative modeling is clear, this approach has been underexplored until recently.

Reward-based Noise Optimization (ReNO) [6] introduces a novel approach to address this research gap. Unlike traditional methods that fine-tune model parameters, ReNO optimizes the initial noise vector at inference using signals from human-preference reward models. This iterative optimization process enables T2I models to produce images that are not only aesthetically superior but also more aligned with the input prompts. By leveraging pre-trained reward models such as ImageReward [7], PickScore [8], HPSv2 [9] and CLIPScore [10], ReNO avoids issue of "reward hacking", a

phenomenon where models optimize for high reward scores at the expense of generating undesirable artifacts. Moreover, by utilizing the distilled one-step T2I model [11, 12, 13, 14], ReNO circumvents the issue of exploding and vanishing gradients. This design significantly reduces inference time, making the proposed method practical for real-world applications.

This review paper aims to provide a comprehensive analysis of the ReNO framework, extending its original findings with additional experimental insights. Specifically, it explores the impact of noise optimization through metrics such as L2 distances, coherence analysis, and visual differences between noise vectors and generated images. Using the Parti-Prompts dataset, which includes diverse and challenging textual descriptions, this study evaluates how ReNO modifies noise to enhance the generative process.

Several key observations arise from this exploration. Firstly, the distance analysis reveals that the optimized noise remains close to the initial noise, suggesting a controlled and targeted refinement process. Surprisingly, the optimized noise often leads to a larger L2 distance from the generated image compared to the initial noise. This finding indicates that ReNO facilitates more "far-reaching" predictions, potentially enhancing the model's ability to capture intricate details in complex prompts. Secondly, coherence analysis shows that ReNO primarily adjusts low-frequency components of the noise, which correspond to broader patterns and structural features, while preserving high-frequency details. This method's behavior means that the optimization process focuses on refining the underlying structure of the noise, which may contribute to the improved prompt faithfulness and aesthetic quality observed in the results. Finally, visualizations of the noise differences confirm that no significant artifacts or distortions are introduced into the initial noise during optimization.

The insights gained from this research have profound implications because they emphasize the fundamental importance of noise manipulation techniques in advancing the capabilities of generative modeling. By shifting the focus from model parameters to noise optimization, ReNO opens up new ways of improving T2I models in a computationally efficient manner. This approach not only aligns with the growing demand for scalable and adaptable generative systems but also sets the stage for future research into more advanced noise optimization techniques.

2 Background and Related Work

2.1 T2I Diffusion Models

Text-to-Image diffusion models [15] work by gradually learning to reverse a process that adds noise to data. Firstly, they progressively introduce noise to an image until it becomes completely random and then learn to reverse this process to generate images from a Gaussian noise conditioned on a text prompt. The forward diffusion process can be described as follows:

$$x_{t+\Delta t} = x_t + \sigma_q \sqrt{\Delta t}\, \zeta, \tag{1}$$

where $x_t \sim p_t(x)$ is data at time step t, σ_q^2 represents the variance of x_T at final time step T, $\Delta t = 1/T$ and $\zeta \sim \mathcal{N}(0, 1)$. In the continuous limit, as $\Delta t \to 0$, the discrete diffusion process turns into a stochastic differential equation (SDE). The general form of the SDE is often represented by this expression:

$$dx = f(x, t)\, dt + g(t)\, dw, \tag{2}$$

where $f(x, t)$ denotes the drift, w is a Brownian motion and g(t) represents the diffusion schedule. Furthermore, the SDE approach is compelling since existing theory provides a general analytical solution for the time-reversed SDE which is given by:

$$dx = \big(f(x, t) - g(t)^2 \nabla_x \log p_t(x)\big)\, dt + g(t)\, d\bar{w}. \tag{3}$$

Here $\nabla_x \log p_t(x)$ is a score function which guides the reverse process by indicating the direction of decreasing noise.

As a result, reverse SDE tells us how to run any forward SDE backward in time. This means that we do not have to re-derive the reversal in each case, and we can choose any SDE solver to yield a sample from the image distribution.

2.2 Distillation Techniques

One of the key challenges in deploying the models in real-world applications is their computational inefficiency due to the high number of denoising steps required for the image generation. Distillation techniques aim to address this issue by reducing the number of inference steps while maintaining image quality. Their solution is to transfer knowledge from a complex, high-capacity teacher model to a smaller, more efficient student model.

Several approaches have been developed to distill diffusion models into single-step generators, which learn to approximate the full stochastic differential equation in one step. We will focus on Adversarial Diffusion Distillation (ADD) [13], Diffusion Matching Distillation (DMD) [14], and Trajectory Segmented Consistency Distillation (TSCD) [12] techniques since they will be further used in the ReNO method.

ADD [13] is a novel approach that reduces the number of inference steps of a pre-trained diffusion model to 1–4 sampling steps while preserving high image quality. The goal of this approach is to combine the superior sample quality of diffusion models with the inherent speed of Generative Adversarial Networks. For this purpose, ADD combines two training objectives: distillation loss and adversarial loss. Score distillation transfers the knowledge from a high-quality diffusion teacher model to the student model by minimizing the difference in their outputs. Adversarial loss ensures that generated images resemble real images more closely by using a discriminator which is trained to distinguish between them. Combining losses, ADD achieves a balance between speed and quality, enabling near real-time image synthesis with performance comparable to state-of-the-art diffusion models.

DMD [14] is another distillation technique for creating efficient versions of diffusion models. It enforces the one-step image generator match the diffusion model at distribution level. To reach this goal, DMD minimizes the Kullback–Leibler (KL) divergence between the output distributions of the pre-trained diffusion model (teacher) and the one-step generator (student) by computing the gradient update based on the difference between their score functions. These score functions are parameterized as two diffusion models trained separately on each distribution. Furthermore, an additional regression loss is used to measure the pointwise distance between the generator and base diffusion model outputs, given the same input noise. It serves as a regularization mechanism to stabilize training and ensure that the one-step generator is aligned with the initial multi-step diffusion model. This combination of KL divergence and regression loss allows DMD to produce high-quality images in a single step, significantly speeding up image generation.

Lastly, TSCD [12] is a combination of trajectory-preserving and trajectory-reformulating distillation techniques. It divides the timesteps into segments and enforces consistency within each segment while gradually reducing the number of segments to achieve all-time consistency. This approach addresses the issue of suboptimal consistency model performance caused by insufficient model fitting capability and accumulated errors in inference.

2.3 Human Preference Reward Models

One of the key challenges in text-to-image synthesis is ensuring the outputs generated by diffusion models align with human preferences. Traditional generative models optimize for likelihood-based objectives, which do not always capture the user intent. Human Preference Reward Models (HPRMs) address this issue. These models quantify and encode human subjective preferences into a reward function, enabling generative models to learn what users find visually or contextually desirable. They take an image and a corresponding prompt as input and output a numerical reward score, indicating how well the image adheres to the given prompt according to human preferences. Usually, HPRMs are trained on the datasets of ranked human preferences, where users rate or compare generations based on various qualitative characteristics (e.g., attribute binding, aesthetics). Then they can be leveraged to fine-tune T2I models resulting in the improved quality of the generated outputs. Next, we will discuss such notable reward models like ImageReward [7], PickScore [8], HPSv2 [9] and CLIPScore [10] that are also used in ReNO framework.

ImageReward [7] is a reward model designed specifically for evaluating and enhancing text-to-image generation. Its annotation pipeline involves a prompt annotation stage, which includes categorizing prompts and identifying problematic ones, and a text-image rating stage, where images are rated based on alignment, fidelity, and harmlessness by professional annotators. Subsequently, the images

are ranked in order of preference. As human evaluation is limited by labor costs and scalability, ImageReward aims to model human preference based on annotations, which can lead to a virtual evaluator that is independent of direct human involvement. For this purpose, it utilizes BLIP [16] as the backbone architecture. The reward model extracts image and text features, combines them with cross attention, and uses an MLP to generate a scalar for preference comparison. Then, it compares pairs of images and learns to rank the more preferred image higher. After pre-training on ranking data, ImageReward can be integrated into the training process of text-to-image generation models.

Another reward model PickScore [8] is trained on the Pick-a-Pic dataset, a large-scale open dataset containing over half-a-million examples of real user preferences on text-to-image generations. It follows the architecture of CLIP [17], utilizing separate text and image encoders to generate embeddings for prompts and images, respectively. The similarity between these embeddings, measured by their inner product, weighted by a learned temperature parameter, serves as the image's quality score. Unlike traditional metrics that focus on image fidelity or aesthetics, PickScore is trained to directly predict human preferences. Its training objective is based on maximizing the likelihood of correctly predicting which image a human would choose from a pair, given a text prompt. It has been shown that using this approach, PickScore correlates better with human rankings than other automatic evaluation metrics.

Human Preference Score v2 (HPSv2) [9] is an advanced reward model trained on Human Preference Dataset v2 (HPDv2), the largest dataset of human-annotated image comparisons, for evaluating human preferences in text-to-image synthesis. It uses CLIP ViT-H/14 model fine-tuned on the HPDv2 dataset to predict a quality score for the given image. Its training objective is to minimize the difference between the predicted preference and the human-assigned rating for that image. Experiments show that HPSv2 generalizes better than previous models, including ImageReward and PickScore.

Finally, CLIPScore [10] is a reference-free metric for evaluating image captions. It measures an image-text alignment by leveraging cosine similarity between visual and textual CLIP embeddings, making it a human-oriented metric for evaluating generative models.

2.4 Initial Noise Optimization

Initial noise optimization is a powerful mechanism for improving the performance of T2I models. Different approaches using this mechanism for different purposes were introduced recently. For instance, D-Flow framework [18] allows for controlled generation by refining the initial noise to achieve desired properties in the final image. Another example is SeedSelect technique [19]. It focuses on finding optimal starting points in the latent space of the diffusion model (seeds) that are more likely to generate images of the desired rare concept.

Most related to the explored ReNO approach is Direct Optimization of Diffusion Latents (DOODL) method [20]. It aims to improve classifier guidance in diffusion models. Other classifier guidance approaches require either training new noise-aware models to obtain accurate gradients or using a one-step denoising approximation of the final generation, which leads to misaligned gradients and sub-optimal control. Instead of training noise-aware classifiers, DOODL utilizes pre-trained classifiers (CLIP models for reinforcing text guidance and visual personalization guidance, Fine-grained visual classification (FGVC) models for vocabulary expansion, "Aesthetics Predictor" model for aesthetic improvement) to optimize the initial latent vector. DOODL employs discretely invertible diffusion algorithm EDICT [21], which admits backpropagation with constant memory cost w.r.t. the number of diffusion steps, to compute classifier gradients on the pixels of the final generation w.r.t. the original noise vectors. This enables efficient iterative optimization of diffusion latents w.r.t. any differentiable loss on the image pixels and accurate calculation of gradients for classifier guidance.

However, aforementioned methods take 10 (DOODL) to 40 (D-Flow) minutes to generate a single image due to their application on time-dependent generative models with a large number of denoising steps. To mitigate this, SeedSelect method utilizes a bootstrapping technique to accelerate image generation, but it is limited to settings where the goal is to generate samples including a concept jointly represented by a set of input images. ReNO framework is proposed to address these issues.

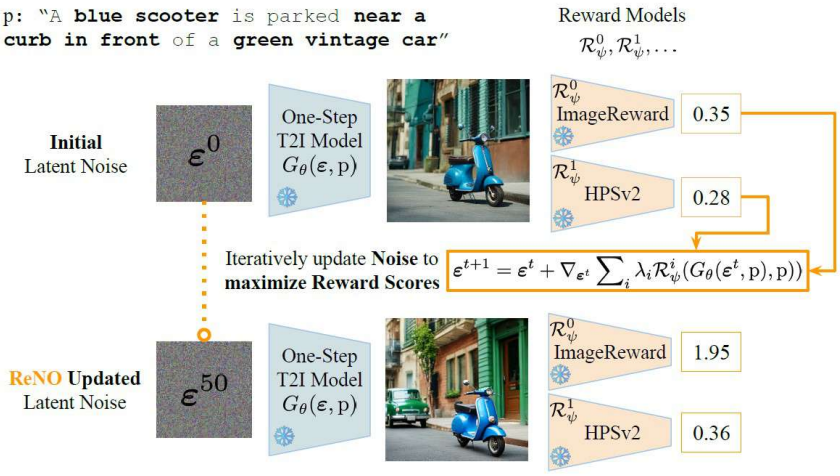

Figure 1: Overview of ReNO framework

3 Reward-based Noise Optimization (ReNO)

3.1 Method

Reward-based Noise Optimization (ReNO) [6] method optimizes the initial random noise during inference without adapting any of the model's parameters. Therefore, to optimize the initial noise ε we should solve the following problem:

$$\varepsilon^* = \arg\max_{\varepsilon} C(G_\theta(\varepsilon, p)), \tag{4}$$

where diffusion model G_θ generates images based on a noise ε and a prompt p, C is a criterion function evaluated on the generated image.

The solution can be obtained through iterative optimization via gradient ascent techniques. However, backpropagating through $C(G_\theta(\varepsilon, p))$ is non-trivial as T2I models are based on the simulation of SDEs as described in Section 2.1. Selecting a one-step model as G_θ enables efficient backpropagation through (4). ReNO builds upon several methods that distill models into one-step generators, which learn to approximate the full SDE in one step (ADD [13] is employed to train SD-Turbo based on SD 2.1 [22] as a teacher and SDXL-Turbo [13] based on SDXL [23], DMD [14] is applied for PixArt-α DMD [11, 24], TSCD is used to distill HyperSDXL [12]). Moreover, this approach achieves image generation, including noise optimization, in 20-50 seconds, making it suitable for practical applications.

One important consideration is that it is desirable for noise ε to stay within the proximity of the initial noise distribution as otherwise the model G_θ might provide unwanted generations. This can be realized by including a regularization term $K(\varepsilon)$ inside of C: $C(\mathbf{x}_0, \varepsilon) = \tilde{C}(\mathbf{x}_0) + K(\varepsilon)$.

In ReNO case, it is suggested to use a weighted combination of pre-trained reward models $R^0_\psi, \ldots, R^n_\psi$ as the criterion function:

$$\tilde{C}(\mathbf{x}_0, p) = \sum_{i=1}^{n} \lambda_i R^i_\psi(\mathbf{x}_0, p), \tag{5}$$

where λ_i denotes the weighting for reward model R^i_ψ.

This can help prevent "reward-hacking" and allow capturing various aspects of image quality and prompt adherence, as different reward models are trained on different prompt and preference sets. This not only effectively combines the strengths of multiple reward models, but also helps mitigate their weaknesses. ReNO leverages four reward models for this purpose (HPSv2 [9], PickScore [8], ImageReward [7], and CLIPScore [10]) that were introduced in Section 2.3.

- Mean L2 distance between initial and optimized noises: 27.0122;
- Mean L2 distance between two random noises: 180.7667;
- Mean L2 distance between initial noise and generated image: 123.3377;
- Mean L2 distance between optimized noise and generated image: 142.3693.

Figure 2: Mean distances between noises and generated images

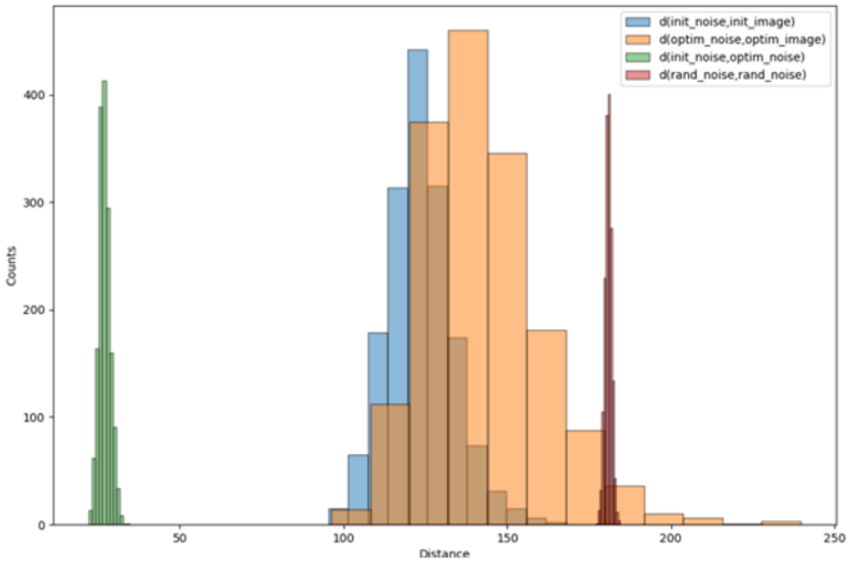

Figure 3: Histogram of distances between noises and generated images

ReNO then boils down to iteratively solving (4) with gradient ascent:

$$\varepsilon^{t+1} = \varepsilon^t + \eta \nabla_{\varepsilon^t} \left[K(\varepsilon^t) + \sum_{i=1}^{n} \lambda_i R^i_\psi(G_\theta(\varepsilon^t, p), p) \right], \tag{6}$$

where η is the learning rate. It is worth mentioning that it is actually not desirable to find the optimal ε^* as we want to prevent adversarial samples that exploit the reward models.

It has been proven that already a few optimization steps (<50) of ReNO lead to significant improvements in both prompt following and visual aesthetics, striking a good balance between reward optimization and the prevention of "reward hacking". A sketch of ReNO is provided in Figure 1.

3.2 Experiments

In experimental part the effectiveness of the ReNO method was evaluated on three benchmarks (T2I-CompBench [5], GenEval [25], Parti-Prompts [26]). The latent noise vector ε was optimized for 50 steps using gradient ascent with Nesterov momentum and gradient norm clipping for stability.

Firstly, it was shown that combining all four reward models leads to having strong improvements in faithfulness, while also increasing the image quality. Then, according to quantitative results on T2I-CompBench, ReNO improved color, shape, and texture binding by 20-25% across multiple models. Moreover, there were significant boosts in the spatial, non-spatial, and complex categories. Similar trends were also noticed for GenEval where applying the considered noise optimization framework helps improve the performance of various one-step diffusion models. The strongest model, HyperSDXL + ReNO, even beat DALL-E 3 [27] on 4/6 categories in GenEval. In both of these benchmarks, ReNO improves results for all the models in all the categories. Further, a user study on the Parti-Prompts dataset was held to validate ReNO. It showed that ReNO-enhanced images were preferred in more than 60% of cases and the performance of HyperSDXL + ReNO model was competitive with proprietary SD3 model [28]. Finally, attribute binding results on T2ICompBench demonstrated that even when restricted to the same compute budget as SDXL (50 steps, 7sec),

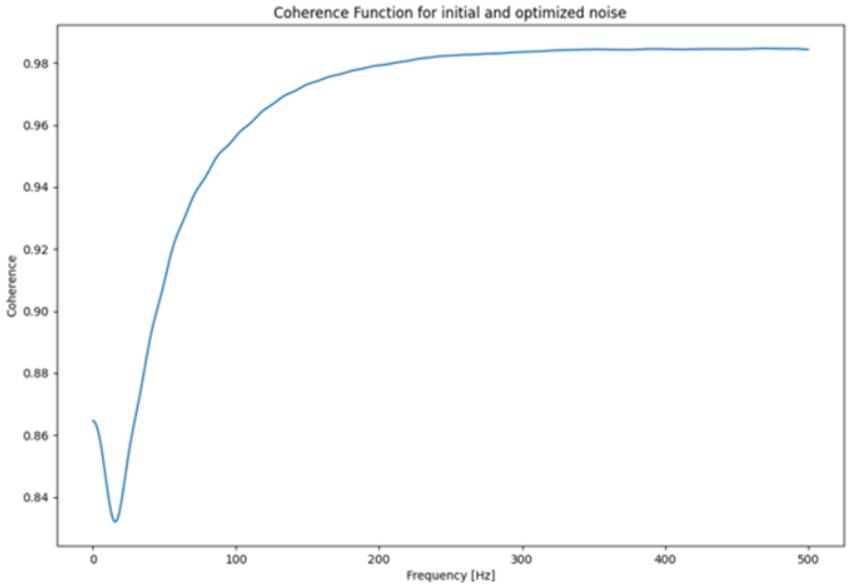

Figure 4: Coherence function for the initial and optimized noises

SD-Turbo + ReNO significantly outperforms it while in this comparison PixArt-α (20 steps, 7sec) lies shortly below the Pareto-frontier of ReNO.

In addition to the results given by the authors of the ReNO, further exploration of the method was performed. The goal of this research is to understand how the optimization process changes the initial noise. Diverse Parti-Prompts dataset (1.6k prompts) was used for this purpose.

Firstly, the distances between the noises and generated images were measured. According to the mean distances (see Figure 2) and the histogram in Figure 3, we see that the distance between the initial and optimized noises is much less than distance between random noises. This observation suggests that the optimization process does not drastically change the noise, and the optimized noise stays close to the initial noise sample. It means that the optimized image is an improved version of the initial one which is a desired outcome.

Surprisingly, the mean L2 distance between the initial noise and image is smaller than the distance between the optimized ones (see Figure 2 and Figure 3). It means that the optimized noise makes the model make a "further" prediction w.r.t. the L2-distance. This finding is probably the most interesting and requires further research.

Next, a plot of the coherence function for the initial and optimized noises was created (see Figure 4). It illustrates that the coherence is lower at low frequencies (which correspond to the broad patterns) and it is increasing at higher frequencies (which capture the finer features). The coherence analysis shows that ReNO primarily modifies these slower-varying components while preserving the rapid fluctuations, suggesting the optimization process focuses on adjusting the underlying structure of the noise rather than its fine-scale variations. Lastly, the difference between the initial and the optimized noise was also visualized in Figure 5. The absence of noticeable artifacts or distortions in the noise indicates that ReNO likely preserves high-frequency components, which correspond to fine details in the final image.

3.3 Limitations

No doubt, ReNO has some limitations. Firstly, despite using different image generation models of varying architectures and sizes, they converge to similar performance on both T2I-Compbench [5] and GenEval [25]. Secondly, the amount of needed GPU VRAM is significantly higher when using ReNO. Thirdly, ReNO is only designed for distilled diffusion models.

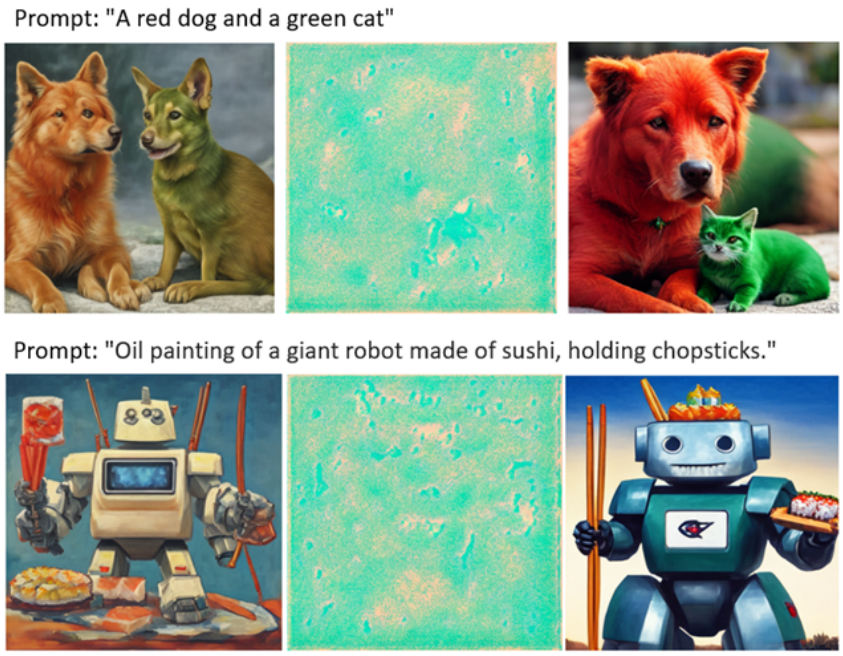

Figure 5: Visualization of the difference between the initial and optimized noises

4 Conclusion

In this work, we review current approaches for manipulating noise in generative modeling, with a primary focus on the ReNO method [6]. We also discuss the mechanism of diffusion models, some popular distillation techniques and notable human preference reward models. Furthermore, we evaluate the performance of ReNO on multiple benchmark datasets across different tasks and analyze how the optimization process modifies the initial noise in this method. In conclusion, ReNO demonstrates that techniques for optimizing initial noise deserve significantly more attention, and further research is necessary to gain a deeper understanding of this essential component of generative models.

References

[1] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. L. Komatsuzaki, "400m: Open dataset of clip-filtered 400 million image-text pairs. arxiv," *arXiv preprint arXiv:2111.02114*, 2021.

[2] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[3] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[5] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 723–78 747, 2023.

[6] L. Eyring, S. Karthik, K. Roth, A. Dosovitskiy, and Z. Akata, "Reno: Enhancing one-step text-to-image models through reward-based noise optimization," *arXiv preprint arXiv:2406.04312*, 2024.

[7] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36 652–36 663, 2023.

[9] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," *arXiv preprint arXiv:2306.09341*, 2023.

[10] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[11] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," in *European Conference on Computer Vision*. Springer, 2024, pp. 74–91.

[12] Y. Ren, X. Xia, Y. Lu, J. Zhang, J. Wu, P. Xie, X. Wang, and X. Xiao, "Hyper-sd: Trajectory segmented consistency model for efficient image synthesis," *arXiv preprint arXiv:2404.13686*, 2024.

[13] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–103.

[14] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, "One-step diffusion with distribution matching distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6613–6623.

[15] P. Nakkiran, A. Bradley, H. Zhou, and M. Advani, "Step-by-step diffusion: An elementary tutorial," *arXiv preprint arXiv:2406.08929*, 2024.

[16] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[18] H. Ben-Hamu, O. Puny, I. Gat, B. Karrer, U. Singer, and Y. Lipman, "D-flow: Differentiating through flows for controlled generation," *arXiv preprint arXiv:2402.14017*, 2024.

[19] D. Samuel, R. Ben-Ari, S. Raviv, N. Darshan, and G. Chechik, "Generating images of rare concepts using pre-trained diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4695–4703.

[20] B. Wallace, A. Gokul, S. Ermon, and N. Naik, "End-to-end diffusion latent optimization improves classifier guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7280–7290.

[21] B. Wallace, A. Gokul, and N. Naik, "Edict: Exact diffusion inversion via coupled transformations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 532–22 541.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[23] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[24] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu *et al.*, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," *arXiv preprint arXiv:2310.00426*, 2023.

[25] D. Ghosh, H. Hajishirzi, and L. Schmidt, "Geneval: An object-focused framework for evaluating text-to-image alignment," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[26] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

[27] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.

[28] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first International Conference on Machine Learning*, 2024.