Dual-Seed Evolutionary Algorithm for Noise Optimization in Diffusion Models

Yuzheng Tan[1], Yuan He[2*], Yao Zhu[3*], Tianlin Huo[2], Huanqian Yan[4], Hang Su[2], Shuxin Zhang[1], Guangneng Hu[5]

[1]State Key Laboratory of Integrated Electromechanical Manufacturing of High-performance Electronic Equipments, Xidian University, Xi'an, China
[2]Tsinghua University, Beijing, China
[3]Zhejiang University, Hangzhou, China
[4]Beihang University, Beijing, China
[5]School of Computer Science and Technology, Xidian University, Xi'an, China
tan0225@stu.xidian.edu.cn, heyuan.1997@tsinghua.org.cn, ee_zhuy@zju.edu.cn

Abstract

Diffusion models have emerged as state-of-the-art generative methods, particularly excelling in conditional tasks such as prompt-driven image synthesis. While recent research emphasizes the pivotal role of noise seeds in enhancing text-image alignment and generating human-preferred outputs, these works predominantly rely on random Gaussian noise or heuristic local adjustments, , overlooking the potential of global optimization strategies to systematically improve generation quality. To bridge this gap, we propose Seed Optimization based on Evolution (SOE), a hybrid framework that integrates global evolutionary search with local semantic refinement. The global evolutionary stage conducts seed selection by jointly optimizing text-image alignment (via CLIP-Score) and human preference estimation (via ImageReward), while the local stage employs diffusion inversion to inject conditional semantics into the noise seed. Together, these components constitute a model-agnostic, training-free optimization framework for conditional diffusion models. Extensive experiments across various diffusion models demonstrate that SOE consistently improves semantic fidelity and visual quality, highlighting its generalizability and potential as a plug-and-play enhancement for generative diffusion pipelines.

Code — github.com/T899work/Dual-seed-evolutionanry-algortithm-for-noise-optimization-in-diffusion-model

Introduction

In recent years, diffusion models have achieved remarkable breakthroughs in multiple fields, including computer vision[(Xing et al. 2024; Nguyen et al. 2023; Garibi et al. 2024)], natural language processing[(Yi et al. 2024)], robotics[(Chi et al. 2024)] and bioinformatics[(Guo et al. 2024b)], significantly propelling the innovation of artificial intelligence technology. As a pivotal application of diffusion models in the visual domain, Text-To-Image (T2I), despite its notable achievements, still encounters substantial technical challenges when handling complex semantic combination prompts. T2I models struggle to accurately parse and integrate diverse semantic information within prompts, leading to artifacts such as missing target entities, mismatched attribute features, and logically inconsistent object compositions(Chefer et al. 2023; Karthik et al. 2023).

Recent studies (Samuel et al. 2024; Karthik et al. 2023) have revealed that diffusion models are highly sensitive to input noise, such that even minor perturbations can lead to significant changes in the output images. This characteristic indicates that the initial noise not only profoundly impacts the visual aesthetics of synthetic images but also largely determines the semantic fidelity between the images and given text prompts (Miao et al. 2025). Currently, one category of methods enhances missing concepts by introducing attention modules (Chefer et al. 2023; Guo et al. 2024a; Agarwal et al. 2023), yet these methods are only applicable to generative models with U-net architectures. Some other approaches focus on optimizing noise using gradients or inversion processes (Miao et al. 2025; Eyring et al. 2024; Li et al. 2024), while additional methods employ neural networks to map the initial noise into "golden noise" (Zhou et al. 2024), which requires training a dedicated neural network for noise mapping based on the generative model. However, all these methods fall under local optimization strategies, suffering from high implementation costs and strong dependence on the initial noise regions.

This paper explores a distinct approach rooted in the hypothesis that diffusion models possess the intrinsic capability to generate semantically faithful images, yet their performance is inherently sensitive to the semantic content of initial random noise—images aligned with text prompts are more likely to emerge when noise is sampled from latent regions rich in prompt-relevant semantics. Inspired by evolutionary algorithms, we introduce a global optimization framework that mimics the evolutionary process of population selection and variation to explore the noise space efficiently. The proposed Seed Optimization based on Evolution (SOE) employs a dual-metric evaluation system: one measuring semantic fidelity and the other assessing visual quality. Leveraging inversion diffusion local optimization, SOE enhances semantic consistency of candidate noise points while introducing an annealed seed search strategy for dynamic interpolation between dual extremal

*Corresponding author.

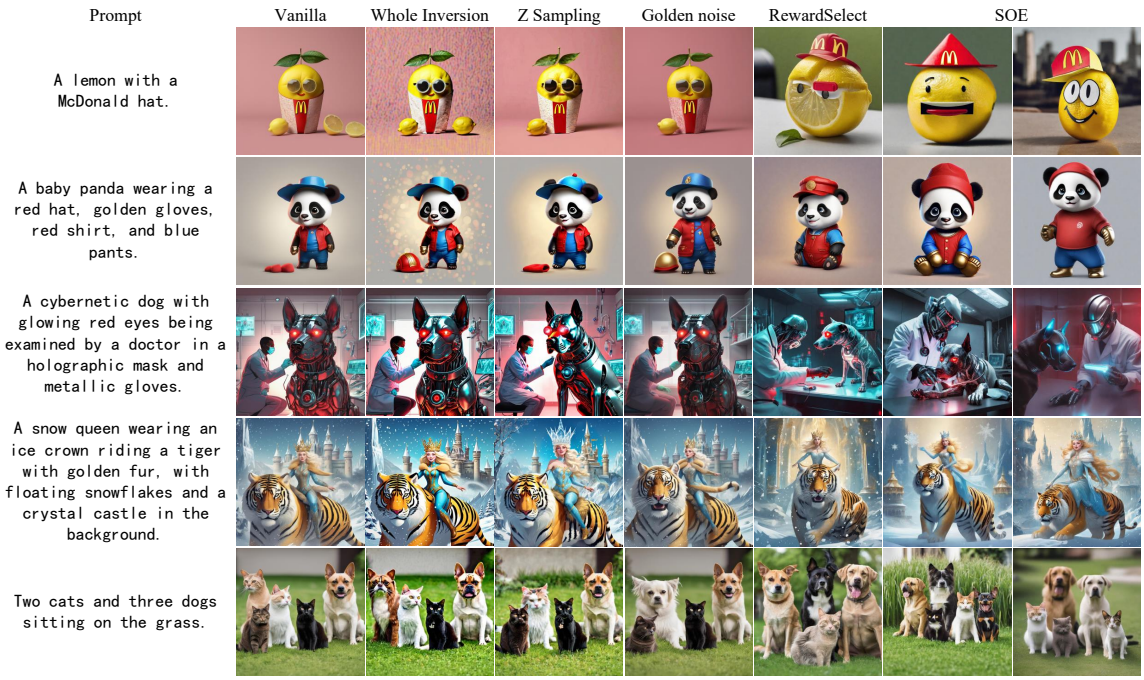

Figure 1: Comparison of SDXL model generation results across 5 prompts: vanilla SDXL, SDXL with Whole Inversion, Z-sampling, Golden noise, RewardSelect, and our SOE.

points. Specifically, inspired by the norm-constrained interpolation theory(Samuel et al. 2023), we employ annealed norm-constrained interpolation between two optimal points. By controlling the path length between these two optimal points, we conduct random wide-range searches in the early stage of optimization to cover diverse semantic interpretations within the initial noise space, and shorten the path length in the later stage for fine-grained optimization—thus enabling the gradual evolution of initial noise.

We conducted comprehensive experiments on three benchmark datasets across mainstream diffusion models, including SDv2.1(Rombach et al. 2022), SDXL(Podell et al. 2023), and SDv3.5(Esser et al. 2024). Model performance was rigorously evaluated using six core metrics: ClipScore(Hessel et al. 2021), ImageReward(Xu et al. 2023), Aesthetic Evaluation Score (AES)(Schuhmann 2022), PickScore(Kirstain et al. 2023), Human Preference Score (HPSv2)(Wu et al. 2023), and the composite metric Average Gain derived from the former five. As illustrated in Fig. 1, the initial noise optimized by SOE leads to substantial improvements in both semantic fidelity and visual aesthetics of the generated images. Across all datasets, consistent and significant enhancements were observed in all evaluation metrics. Crucially, SOE relies solely on generic evaluation models rather than being tied to specific diffusion model architectures. This not only endows it with excellent cross-model generalization and practical applicability but also makes it a plug-and-play solution that requires no pre-training, additional modules, or reliance on specific architectural constraints.

Our main contributions are summarized as follows:

- We have developed the dual-seed optimization framework based on evolution (SOE) method, which leverages the collaboration between dual-evaluation models and few step inversion to optimize the noise in diffusion models.
- Our study reveals that global noise optimization plays a more critical role than local refinement in improving both semantic alignment and perceptual quality. Additionally, SOE exhibits strong compatibility with prior methods (e.g., Golden Noise), functioning as an orthogonal enhancement that can be seamlessly integrated into existing pipelines.
- The effectiveness of SOE has been validated through extensive experiments. Across three datasets, it outperforms existing methods by 40% in SDv2.1, more than 30% in SDXL, and over 5% in SDv3.5 medium.

Related Work

Text-to-Image Generation The goal of text-to-image synthesis is to generate images consistent with given text, with early research focusing on GANs (Goodfellow et al. 2014), autoregressive models(Xiong et al. 2024), flow-matching model(Ben-Hamu et al. 2024), though limited by issues like mode collapse. In recent years, diffusion models have become dominant, where denoising diffusion probabilistic models (Ho, Jain, and Abbeel 2020) transform noise into images via progressive denoising, and conditional diffusion models (Dhariwal and Nichol 2021) achieve semantic control by integrating textual embeddings with guided sampling—early works assumed initial noise follows a stan-

dard normal distribution $\mathcal{N}(0, I)$. However, precise alignment with text prompts remains challenging, with key factors being prompt encoding accuracy and initial noise distribution(Feng et al. 2023a; Karthik et al. 2023). To address the issue of diffusion models generating results inconsistent with given conditions, numerous scholars have proposed various optimization strategies(Samuel et al. 2023; Wallace et al. 2024; Li et al. 2023; Agarwal et al. 2023; Feng et al. 2023b; Zhang et al. 2024). Current studies address this in two main ways: some optimize prompt representations(Yuksekgonul et al. 2025; Xue et al. 2025) or embedding encodings (Feng et al. 2023a), while another category focuses on optimizing initial noise (Qi et al. 2024; Zhou et al. 2024; Karthik et al. 2023; Chen et al. 2024), which aligns with our research.

Initial Noise Optimization Recent advances in noise optimization for text-to-image synthesis include: (Chefer et al. 2023) introduced Generative Semantic Nursing, enhancing text-image alignment by augmenting cross-attention to missing targets and slightly adjusting noisy images at each denoising timestep; (Guo et al. 2024a) designed cross-attention response scores and self-attention conflict scores to partition the latent space and optimize noise toward effective regions; (Meng et al. 2023) observed that denoising with an inversion step improves image quality, which inspiring Golden Noise to generate noise via one-step reverse, construct a dataset, and train a transformation network; (Xu et al. 2025) address style transfer by introducing negative guidance in the reverse stage to shift sampling origins away from original style content; (Bai et al. 2025; Bai, Sugiyama, and Xie 2025) accumulates semantic information by alternating denoising and inversion to leverage guidance gaps; (Lu et al. 2025) aligns diffusion models with human preferences via Inversion Preference optimization; in gradient-based approaches, (Eyring et al. 2024) optimizes noise using multi-evaluator scores to boost prompt faithfulness and aesthetic quality, while (Miao et al. 2025) employs a visual question answering model for scoring and simplified gradient-based noise optimization. Notably, these methods primarily rely on local optimization and some depend on specific models. Several studies have also explored global optimization strategies (Karthik et al. 2023), which generate high-quality images through batch generation and evaluator-based screening. However, their reliance on a single evaluation criterion results in less significant improvements in semantic alignment compared to enhancements in visual quality.

To address the above challenges, we propose a novel training-free optimization pipeline that identifies high-quality initial noise with both strong visual fidelity and semantic alignment. The method selects two initial noise with optimal image quality and highest semantic matching within a batch, and combines them through a reverse semantic injection mechanism. Benefiting from its independence from specific diffusion model versions or architectures, our approach offers broad applicability and ease of integration.

Preliminaries

We first present preliminaries about DDIM(Song, Meng, and Ermon 2023) and DDIM Inversion and the conditional sampling. Given the noise schedule α_t and σ_t and time step t. we denote the forward noising process of diffusion models as $x_t = \alpha_t x_0 + \sigma_t \epsilon_t$. In the sampling process, we first sample the initial noise $x_T \sim \mathcal{N}(0, I)$. Then, the forward noising process proceeds iteratively from $t = T$ down to $t = 1$ according to the following update rule:

$$\begin{aligned} x_{t-1} =& \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} x_t + \sqrt{\alpha_{t-1}} \\ & \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \varepsilon_\theta \left(x_t, t, \omega \right) \end{aligned} \tag{1}$$

DDIM Inversion Denoising Diffusion Implicit Models(Song, Meng, and Ermon 2023) (DDIM) represents a seminal deterministic sampling technique that constructs implicit non-Markovian diffusion paths, achieving significant improvements in generation efficiency while maintaining sample fidelity. The sampling process can be formulated as:

$$\begin{aligned} x_t =& \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} x_{t-1} + \sqrt{\alpha_t} \\ & \left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1} \right) \bar{\varepsilon}_\theta \left(x_{t-1}, t-1, \gamma \right) \end{aligned} \tag{2}$$

DDIM Inversion (Lugmayr et al. 2022; Mokady et al. 2023; Pan et al. 2023; Zhang, Lewis, and Kleijn 2024; Hong et al. 2024) is the inverse process of DDIM sampling. Its core lies in leveraging a deterministic path to reverse the diffusion process from a low noise level image back to images with higher noise levels, ultimately recovering the initial noise x_T that aligns with the given text prompt.

Conditional Guidance Mechanism Classifier-Free Guidance (CFG)(Ho and Salimans 2020) enhances generative controllability by interpolating conditional and unconditional diffusion model outputs. At time step t, the guided noise prediction is given by:

$$\bar{\varepsilon}_\theta \left(x_t, t, \omega \right) = \omega \varepsilon_\theta \left(x_t, t | \, c \right) - (\omega - 1) \, \varepsilon_\theta \left(x_t, t | \, \Omega \right) \tag{3}$$

where ω denotes the guidance scale and $\emptyset$ represents the unconditional input. Notably, the initial noise x_T in forward diffusion should satisfy the conditional distribution $x_T \sim N(0, I | \, condition)$(Dhariwal and Nichol 2021). Existing methods often overlook this prior assumption, leading to potential mismatches between the reverse process optimization target and the true data distribution.

Seed Optimization Based on Evaluation

The initial noise X_T significantly influences both the quality and semantic consistency of generated images (Karthik et al. 2023; Eyring et al. 2024). As shown in Fig. 2, we innovatively propose a dual-noise collaborative search mechanism

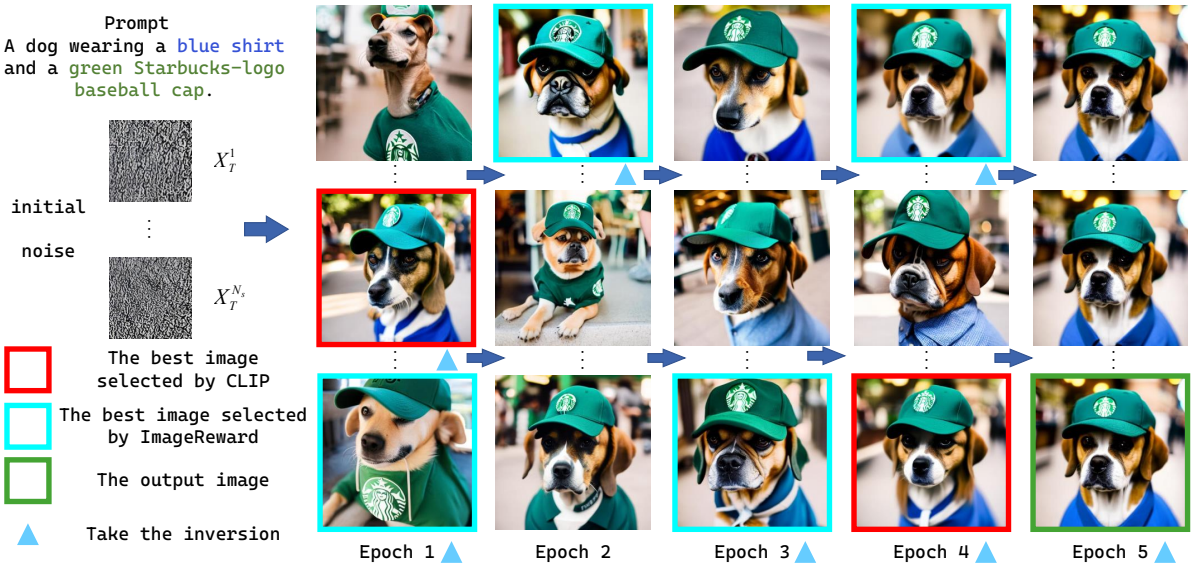

Figure 2: (Left) Schematic diagram of the Prompt and initial noise. (Right) Optimized iterative generation based on SDv2.1: evolution from initial noise X_T to selection by ClipScore, ImageReward, and final result output. In each epoch of SOE, a batch of images is first generated, and the current optimal samples are selected via ImageReward and ClipScore. After information is injected through Inversion, the optimal sample of the current epoch is obtained. For the initial noise corresponding to each batch, random sampling is used in the first round; for other rounds, interpolated noise generated based on the two optimal samples is adopted. The final output is the best image selected by ImageReward.

combined with local mutation strategy, which jointly optimizes semantic and visual quality by maximizing in parallel reward signals from diverse preference dimensions. Since inversion with guided difference can inject conditional information into the initial noise, local mutation is implemented via inversion.

Dual-noise Collaborative Search Mechanism Prior work (Karthik et al. 2023; Eyring et al. 2024) has shown that seed selection guided by learned reward models can significantly enhance performance. However, most existing approaches focus on optimizing a single or composite metric globally, which often introduces bias and fails to balance visual quality and semantic consistency. To this end, we propose a **dual-noise collaborative optimization strategy** under a multi-preference reward formulation, inspired by principle of linear interpolation between noise seed, which can integrate the content of two images.

Concretely, during each optimization round, we first sample a population of noise vectors $\{x_j\}_{j=1}^{N_s}$ from the standard Gaussian prior $\mathcal{N}(0, I)$. Each sample is evaluated using two distinct reward functions representing orthogonal human preferences: $\mathcal{R}_1(\cdot)$ for semantic relevance (ClipScore) and $\mathcal{R}_2(\cdot)$ for human preference reward (ImageReward). This dual-metric evaluation yields two optimal seeds:

$$x_{\text{sem}} = \arg\max_{x_j} \mathcal{R}_1(x_j), \quad x_{\text{hum}} = \arg\max_{x_j} \mathcal{R}_2(x_j) \quad (4)$$

These seeds capture the most promising directions in latent space under their respective objectives. To facilitate exploration between these competing optima, we construct a new noise candidate via linear interpolation with stochastic perturbation:

$$x_i = \frac{(1-e_i)x_{\text{sem}} + e_i x_{\text{hum}}}{\sqrt[n_i]{2}} + \frac{z}{\sqrt{2^{n_i}}}, \quad z \sim \mathcal{N}(0, I) \quad (5)$$

where $e_i \in [0, 1]$ controls the interpolation bias, and n_i denotes the iteration index. The design of scaling term gradually diminishes the influence of random perturbations by reducing the weight of the stochastic term as the iteration count n_i increases. The initial noise also satisfies the χ^d distribution in terms of norm:

$$\|x_i\| \sim \chi^d = \frac{\|x_i\|^{d-1} e^{-\|x_i\|^2/2}}{2^{d/2-1}\Gamma(d/2)} \quad (6)$$

where $\|\cdot\|$ is the stand Euclidean norm and $\Gamma(\cdot)$ is the Gamma function.

The prior over the seed space can be defined as $P(z_T) := \chi_d(\|z_T\|)$, where this probability density function represents the likelihood that a seed with norm $\|z_T\|$ is drawn from a Gaussian distribution. To guide sampling toward high-quality regions, we redefine an objective function based on the probability density over N_e interpolated samples:

$$\mathcal{L}_{\text{SOE}} = -\sum_{j=1}^{N_e} \log P(x_j) \cdot \|x_j - x_{j-1}\|^{\exp(n_i - N_i/2)} \quad (7)$$

which allows the interpolation to respect both the statistical properties of the diffusion prior and the evolutionary principle of intermediate recombination. The exponential weight

term progressively suppresses large transitions as optimization proceeds, transitioning from global search to local refinement.

This search-evaluate-update scheme integrates semantic and visual preferences in a principled way, analogous to crossover and selection in evolutionary algorithms. Unlike purely random or single-objective selection schemes, our method encourages diverse and high-reward seeds to collaboratively guide the noise trajectory. Furthermore, it is compatible with any reward model and agnostic to downstream diffusion architectures, making it a universal plug-in for seed selection.

Local Mutation via Guided Inversion While the dual-seed crossover mechanism enables global exploration in the noise space, evolutionary algorithms also rely on *mutation* to introduce diversity and perform local refinement. To this end, we incorporate a lightweight, model-agnostic mutation strategy via **guided DDIM inversion**.

In particular, we leverage the Classifier-Free Guidance (CFG) formulation (Ho and Salimans 2020) to perturb the latent trajectory. Given a noisy latent x_t, the guided noise prediction is expressed as:

$$\hat{\varepsilon}_\theta(x_t, c) = \gamma\varepsilon_\theta(x_t \mid c) - (\gamma - 1)\varepsilon_\theta(x_t \mid \emptyset) \tag{8}$$

where γ is the guidance scale, c is the conditioning prompt, and ε_θ denotes the predicted noise. During the generation process, a higher γ emphasizes alignment with the prompt.

We adopt DDIM to map the selected latent sample x_T back to the intermediate state $x_{t_{\text{inv}}}$, and further map it to $\bar{x}_T$ via DDIM inversion with a guidance gap. Here, t_{inv} is a small integer relative to the full inference steps. This enables local search around semantically meaningful trajectories in the latent space.

In practice, we fix the guidance scale γ during inversion to zero, effectively removing conditional bias in the backward trajectory and allowing for generalization across preference metrics. During re-sampling, the guidance scale is reintroduced to refine toward task-specific signals.

To balance quality and stability, we constrain the number of inversion steps to a small fraction of the total inference steps:

$$t_{\text{inv}} = \lfloor \rho \cdot N_{\text{inf}} \rfloor, \quad \rho \in [0.14, 0.20] \tag{9}$$

where N_{inf} is the total number of inference steps. As discussed in Appendix D, excessive inversion steps may distort the generation quality due to accumulation of noise.

Overall, the complete SOE procedure alternates between global search via dual-seed crossover and local refinement via guided inversion, promoting diversity and controllability simultaneously. The full optimization routine is summarized in Algorithm 1.

Experiments

To evaluate image generation quality under given prompts, all numerical simulations of Algorithm 1 were conducted on a single NVIDIA GeForce RTX 4090D GPU. In the initial setup, $N_s = 10, N_i = 4, Y = 9, \gamma = 0$. First, we assess SOE's performance in text-to-image tasks. Second, we conduct an ablation study on SOE's components and hyperparameter impacts.

Algorithm 1: SOE algorithm

Input: Prompt c, number of iterationsN_i, number of seed per iterationN_s, inversion step Y, inversion guidanceγ

Output: Optimized noise x_{hum}, image $x_{0,best}$

Sample initial Gaussian noise x_T and generate image.
Select x_{sem} and x_{hum} from the generated images by using the evaluation model
Perform Y-step Inversion with guidance scale γ on the x_{sem} and x_{hum} to obtain $x_{\text{sem}-\text{I}}$ and $x_{\text{hum}-\text{I}}$
if Evaluation score of $x_{\text{sem}-\text{I}}$ and $x_{\text{hum}-\text{I}}$ meet the improvement criteria **then**
    $x_{\text{sem}} = x_{\text{sem}-\text{I}}$, $x_{\text{hum}} = x_{\text{hum}-\text{I}}$
end if
for $n_i = 1$ to N_i **do**
    Initial noise samples based on equ.(5)
    Optimize the norm of samples according to equ.(7)
    Generate images from the optimized noise samples.
    Select x_{sem} and x_{hum} from generated images by using the evaluation model
    Perform Y-step Inversion with guidance scale γ on the x_{sem} and x_{hum} to obtain $x_{\text{sem}-\text{I}}$ and $x_{\text{hum}-\text{I}}$
    if Evaluation score of $x_{\text{sem}-\text{I}}$ and $x_{\text{hum}-\text{I}}$ meet the improvement criteria **then**
        $x_{\text{sem}} = x_{\text{sem}-\text{I}}$, $x_{\text{hum}} = x_{\text{hum}-\text{I}}$
    end if
end for
return x_{hum} and corresponding result $x_{0,best}$

Text-To-Image Generation We comprehensively assess the quality of generated images using our proposed approach. Specifically, we conduct both quantitative and qualitative analyses.

Dataset In accordance with the settings used in prior work (Zhou et al. 2024; Bai et al. 2025), we sourced prompts from subsets of three datasets for text-to-image generation: HPDv2(Wu et al. 2023), PickaPic(Kirstain et al. 2023), and Drawbench(Saharia et al. 2022), random sampling 200 prompts each from HPDv2 and PickaPic, and 100 from Drawbench. Experiments were conducted at 512 × 512 and 1024 × 1024 resolutions, leveraging four open-source models: SDv1.4, SDv2.1(Rombach et al. 2022), SDXL(Podell

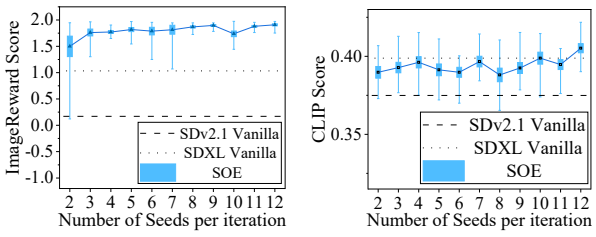

Figure 3: Fixed 4 Iterations: Quantitative Results of SOE on ImageReward and ClipScore with Varying Seeds per Iteration (10 Random Tests on Single Prompt)

Dataset	Method	ClipScore	ImageReward	PickScore	AES	HPSv2	Average Gain
HPDv2	SDv2.1	0.3272	0.3969	21.4301	5.5710	26.0938	-
	Z-Sampling	0.3274	0.4421	21.4119	5.5583	26.6250	0.1317
	Attention excite	0.3293	0.3901	21.4730	5.5717	26.5000	0.0070
	Initno	0.3279	0.3438	21.4976	5.5534	26.3500	-0.1218
	RewardSelect	0.3383	1.1494	21.8837	5.6468	27.4062	2.0149
	SOE	**0.3395**	**1.3405**	**21.9601**	**5.6491**	**27.5156**	**2.5083**
PickaPic	SDv2.1	0.3260	0.3749	20.8608	5.6116	26.0625	-
	Z-Sampling	0.3256	0.4956	20.7767	5.5958	26.7344	0.3397
	Attention excite	0.3265	0.4158	21.0952	5.5871	26.1562	0.1211
	Initno	0.3240	0.4252	21.0694	5.5519	26.3125	0.1370
	RewardSelect	0.3350	1.2546	21.5855	5.6445	27.4844	2.4693
	SOE	**0.3389**	**1.4147**	**21.6473**	**5.6456**	**27.9062**	**2.9276**
Drawbench	SDv2.1	0.3240	0.3113	21.3789	5.4840	26.4375	-
	Z-Sampling	0.3299	0.3589	21.4082	5.4171	26.7812	0.1733
	Attention excite	0.3283	0.2359	21.6314	5.4106	26.2812	-0.2364
	Initno	0.3283	0.4074	21.5260	5.4194	26.5938	0.2804
	RewardSelect	0.3434	1.2328	22.2017	5.4785	27.4531	3.0959
	SOE	**0.3471**	**1.4702**	**22.2417**	**5.4926**	**28.0469**	**3.8969**

Table 1: Quantitative comparison results of SOE with comparative methods using SDv2.1 on Pick-a-Pic, DrawBench, and HPDv2 datasets.

et al. 2023) and SDv3.5 medium(Esser et al. 2024). Specifically, SDv1.4 and SDv2.1 generated 512×512 images, while the remaining models produced 1024×1024 outputs.

Baseline and Evaluation We employ five evaluation metrics: ClipScore(Hessel et al. 2021), ImageRewards(Xu et al. 2023), PickScore(Kirstain et al. 2023), AES(Schuhmann 2022), and HPSv2(Wu et al. 2023) to validate the effectiveness of SOE. As each metric emphasizes different aspects , we calculate an Average Gain (AG) to quantify the improvement of our approach. The formula for AG is:

$$AG = \sum_i \frac{S_i - S_{base}}{S_{base}} \tag{10}$$

where a higher AG value indicates a more significant enhancement. Given that many existing noise optimization methods exhibit strong model dependency, we conducted a comprehensive comparative study across different models. Specifically, we employed seven baseline methods: Vanilla SD, whole-inversion, Zigzag(Bai et al. 2025), Inversion, Attention Excite(Chefer et al. 2023), Initno(Guo et al. 2024a), Golden noise(Zhou et al. 2024), and RewardSelect(Karthik et al. 2023), ensuring fair comparisons across methods.

Qualitative Analysis We demonstrate the performance evolution during the optimization process in Fig.2. The optimized images progressively exhibit higher semantic fidelity and improved image quality throughout the process. Fig. 1 presents a qualitative comparison of generated images by Vanilla SDv2.1, Whole Inversion, Zigzag, Golden noise, RewardSelect, and our proposed SOE on semantically complex prompts. Visual comparisons demonstrate that images generated by SOE exhibit superior semantic fidelity to the key concepts and higher image quality compared to the baselines.

Quantitative Analysis Tables 1 and 2 summarize the quantitative results of SOE and baseline methods across three datasets (HPDv2, PickaPic, and Drawbench) on the SDv2.1 and SDXL models. SOE achieves the highest scores across all evaluation metrics for all datasets, with the overall optimal *AG*. This demonstrates that SOE consistently outperforms baseline methods in both semantic alignment and visual quality. In particular, SOE exhibits significant advantages in semantic alignment compared to RewardSelect, the most competitive baseline method. On the Drawbench dataset, SOE achieves an *AG* score of 3.8969 on SDv2.1, outperforming RewardSelect by 0.8010; on SDXL, its *AG* score is 1.3876, exceeding RewardSelect by 0.5173 and Golden noise by 0.5116. On the PickaPic dataset, SOE's *AG* score on SDv2.1 is 2.9276, surpassing RewardSelect by 0.4583 and Z-sampling by 2.5879. On the HPDv2 dataset, SOE scores 1.5637 (ImageReward) and 30.6562 (HPSv2) on SDXL, and 1.3405 (ImageReward) and 27.5156 (HPSv2) on SDv2.1.

Ablation Study We conducted ablation experiments on the components of SOE and several influencing factors. These ablation studies analyze the number of seeds per iteration, optimization epochs.

Number of Seed per Iteration and Optimization Epoch Fig. 3 illustrates the impact of the number of noise seeds

Dataset	Method	ClipScore	ImageReward	PickScore	AES	HPSv2	Average Gain
HPDv2	SDXL	0.3428	0.9760	22.6731	5.7909	28.7969	-
	Z-Sampling	0.3426	1.0364	22.6271	5.8315	29.8381	0.1024
	Whole-inversion	0.3426	1.0305	22.4848	5.7853	29.7812	0.0802
	RewardSelect	0.3438	1.2617	22.7531	**5.8441**	29.8438	0.3447
	Golden noise	0.3396	0.9677	22.6845	5.8362	29.2344	0.0057
	SOE	**0.3462**	**1.5637**	**22.9218**	5.8373	**30.6562**	**0.6956**
PickaPic	SDXL	0.3396	0.9114	22.3555	5.8959	29.2656	-
	Z-Sampling	0.3413	1.0268	22.3594	5.9605	30.0000	0.1678
	Whole-inversion	0.3429	1.0477	22.2895	5.8917	30.1250	0.1850
	RewardSelect	0.3411	1.2026	22.5529	5.9340	29.8594	0.3595
	Golden noise	0.3378	0.9546	22.3650	**5.9640**	29.2969	0.0551
	SOE	**0.3463**	**1.5372**	**22.6197**	5.9531	**30.9062**	**0.7839**
Drawbench	SDXL	0.3291	0.7474	22.2788	5.4963	27.4062	-
	Z-Sampling	0.3457	0.9862	22.7656	5.6255	29.7656	0.5014
	Whole-inversion	0.3318	0.8983	22.2792	5.5274	28.6406	0.2608
	RewardSelect	0.3485	1.2638	22.9922	5.5526	29.5469	0.8703
	Golden noise	0.3448	1.0070	22.8444	**5.6340**	29.2188	0.5116
	SOE	**0.3539**	**1.5959**	**23.1969**	5.6032	**30.5938**	**1.3876**

Table 2: Quantitative comparison results of SOE with comparative methods using SDXL on Pick-a-Pic, DrawBench, and HPDv2 datasets.

per iteration. Building on earlier experiments validating the effectiveness of SOE over baseline models, we further explore an intriguing question: how many seeds are needed for a weaker early-stage model like SDv2.1 to surpass a stronger model such as SDXL? On the ImageReward metric, SDv2.1 requires only 3 seeds to exceed the average performance of vanilla SDXL, whereas at least 12 seeds are necessary to outperform SDXL on ClipScore. Additionally, experiments on the number of iterations show that increasing the number of rounds improves ImageReward scores, but yields limited gains on ClipScore.

Discussion and Limitations

Although our method demonstrates strong performance across various tasks, no approach is universally optimal in the context of multi-objective optimization. Therefore, several limitations remain. First, SOE relies on both evaluation models and reverse-diffusion-based local optimization. While this study focuses on text-to-image diffusion models, extending to other domains such as video generation, molecular synthesis, or 3D generation will require domain-specific and more robust evaluators. Second, experimental results show that as the improvement of the base diffusion models, the gains from noise optimization tend to diminish. Nevertheless, our method consistently outperforms existing approaches across multiple benchmarks. Finally, similar to other local optimization methods, SOE incurs increased runtime due to extensive sampling operations. Although batch parallel generation can accelerate the process, it consumes more GPU memory. In this study, we adopted serial generation (with FP16 precision in SDv3.5 to reduce memory usage), and future work may explore more efficient parallel strategies.

Conclusion

In this study, we introduce Seed Optimization based on Evolution (SOE), a global noise optimization strategy for text-to-image generation that requires no fine-tuning of the generative model. The approach leverages a critic model to select two seed vectors with complementary advantages, integrating inversion-based local optimization for fine-grained feature exploration and inter-seed sampling for global noise space coverage, thus forming an evolutionary optimization framework. Extensive experiments on HPDv2, PickaPic, and Drawbench datasets demonstrate that SOE, when applied to the SDXL model, significantly outperforms the vanilla SDXL. Compared with state-of-the-art methods such as RewardSelect, SOE maintains high-quality generation while demonstrating more comprehensive noise space exploration through its dual-seed collaborative search mechanism, highlighting the advantage of balancing local feature optimization and global structure search. We hope this work not only provides new perspectives on seed optimization in generative modeling but also encourages future research on balanced multi-objective optimization mechanisms and robust evaluation models across diverse preference dimensions.

Acknowledgments

This research was funded by:

- National Natural Science Foundation of China (Grant No. 52322507);
- National Natural Science Foundation of China (Grant No. 62306220);
- Shaanxi Natural Science Basic Research Project(Grant No. 2025SYS-SYSZD-103);

References

Agarwal, A.; Karanam, S.; Joseph, K.; Saxena, A.; Goswami, K.; and Srinivasan, B. V. 2023. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2283–2293.

Bai, L.; Shao, S.; Zhou, Z.; Qi, Z.; Xu, Z.; Xiong, H.; and Xie, Z. 2025. Zigzag Diffusion Sampling: Diffusion Models can Self-Improve via Self-Reflection. *International Conference on Learning Representations*.

Bai, L.; Sugiyama, M.; and Xie, Z. 2025. Weak-to-Strong Diffusion with Reflection. *arXiv preprint arXiv:2502.00473*.

Ben-Hamu, H.; Puny, O.; Gat, I.; Karrer, B.; Singer, U.; and Lipman, Y. 2024. D-Flow: Differentiating through Flows for Controlled Generation. In *International Conference on Machine Learning*, 3462–3483. PMLR.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4): 1–10.

Chen, C.; Yang, L.; Yang, X.; Chen, L.; He, G.; Wang, C.; and Li, Y. 2024. Find: Fine-tuning initial noise distribution with policy optimization for diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6735–6744.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *The International Journal of Robotics Research*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Eyring, L.; Karthik, S.; Roth, K.; Dosovitskiy, A.; and Akata, Z. 2024. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37: 125487–125519.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A. R.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023a. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations*.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A. R.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023b. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations*.

Garibi, D.; Patashnik, O.; Voynov, A.; Averbuch-Elor, H.; and Cohen-Or, D. 2024. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*, 395–413. Springer.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

Guo, Z.; Liu, J.; Wang, Y.; Chen, M.; Wang, D.; Xu, D.; and Cheng, J. 2024b. Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, 2(2): 136–154.

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

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2020. Classifier-Free Diffusion Guidance. In *In Neural Information Processing Systems*, 6840–6851.

Hong, S.; Lee, K.; Jeon, S. Y.; Bae, H.; and Chun, S. Y. 2024. On exact inversion of dpm-solvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7069–7078.

Karthik, S.; Roth, K.; Mancini, M.; and Akata, Z. 2023. If at First You Don't Succeed, Try, Try Again: Faithful Diffusion-based Text-to-Image Generation by Selection. *arXiv preprint arXiv:2305.13308*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Li, J.; Feng, W.; Chen, W.; and Wang, W. Y. 2024. Reward guided latent consistency distillation. *arXiv preprint arXiv:2403.11027*.

Li, Y.; Keuper, M.; Zhang, D.; and Khoreva, A. 2023. Divide & Bind Your Attention for Improved Generative Semantic Nursing. In *34th British Machine Vision Conference*. BMVA.

Lu, Y.; Wang, Q.; Cao, H.; Wang, X.; Xu, X.; and Zhang, M. 2025. InPO: Inversion Preference Optimization with Reparametrized DDIM for Efficient Diffusion Model Alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28629–28639.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14297–14306.

Miao, B.; Li, C.; Wang, X.; Zhang, A.; Sun, R.; Wang, Z.; and Zhu, Y. 2025. Noise Diffusion for Enhancing Semantic Faithfulness in Text-to-Image Synthesis. *arXiv preprint arXiv:2411.16503*.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6038–6047.

Nguyen, T.; Li, Y.; Ojha, U.; and Lee, Y. J. 2023. Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems*, 36: 9598–9613.

Pan, Z.; Gherardi, R.; Xie, X.; and Huang, S. 2023. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15912–15921.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Qi, Z.; Bai, L.; Xiong, H.; and Xie, Z. 2024. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Samuel, D.; Ben-Ari, R.; Darshan, N.; Maron, H.; and Chechik, G. 2023. Norm-guided latent space exploration for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 57863–57875.

Samuel, D.; Ben-Ari, R.; Raviv, S.; Darshan, N.; and Chechik, G. 2024. Generating images of rare concepts using pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4695–4703.

Schuhmann, C. 2022. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor.

Song, J.; Meng, C.; and Ermon, S. 2023. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238.

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

Xing, Z.; Feng, Q.; Chen, H.; Dai, Q.; Hu, H.; Xu, H.; Wu, Z.; and Jiang, Y.-G. 2024. A survey on video diffusion models. *ACM Computing Surveys*, 57(2): 1–42.

Xiong, J.; Liu, G.; Huang, L.; Wu, C.; Wu, T.; Mu, Y.; Yao, Y.; Shen, H.; Wan, Z.; Huang, J.; et al. 2024. Autoregressive models in vision: A survey. *arXiv preprint arXiv:2411.05902*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Xu, R.; Xi, W.; Wang, X.; Mao, Y.; and Cheng, Z. 2025. StyleSSP: Sampling StartPoint Enhancement for Training-free Diffusion-based Method for Style Transfer. *arXiv preprint arXiv:2501.11319*.

Xue, Q.; Yin, X.; Yang, B.; and Gao, W. 2025. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18826–18836.

Yi, Q.; Chen, X.; Zhang, C.; Zhou, Z.; Zhu, L.; and Kong, X. 2024. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 10: e1905.

Yuksekgonul, M.; Bianchi, F.; Boen, J.; Liu, S.; Lu, P.; Huang, Z.; Guestrin, C.; and Zou, J. 2025. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639: 609–616.

Zhang, G.; Lewis, J. P.; and Kleijn, W. B. 2024. Exact diffusion inversion via bidirectional integration approximation. In *European Conference on Computer Vision*, 19–36. Springer.

Zhang, X.; Yang, L.; Li, G.; Cai, Y.; Xie, J.; Tang, Y.; Yang, Y.; Wang, M.; and Cui, B. 2024. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. *arXiv preprint arXiv:2410.07171*.

Zhou, Z.; Shao, S.; Bai, L.; Xu, Z.; Han, B.; and Xie, Z. 2024. Golden noise for diffusion models: A learning framework. *arXiv preprint arXiv:2411.09502*.