Boosting Text-to-Image Diffusion Models via Core Token Attention-Based Seed Selection

Yunzhe Zhang, Hongfu Liu, Pengyu Hong
Department of Computer Science
Brandeis University, Waltham, MA, USA
{yunzhezhang,hongfuliu,hongpeng}@brandeis.edu

Abstract

Text-to-image diffusion models can synthesize high-quality images, yet the outcome is notoriously sensitive to the random seed: different initial seeds often yield large variations in image quality and prompt–image alignment. We revisit this "seed effect" and show that attention dynamics over prompt core tokens, the content-bearing words, measured during the first few denoising steps, strongly predict final generation quality. Building on this observation, we introduce Attention-Based Seed Selection (ABSS), a training-free, plug-and-play method that ranks seeds for a given prompt by leveraging cross-attention to core tokens during the denoising process. ABSS requires no finetuning and does not alter the initial noise; it scores and ranks all candidate seeds, keeps only the top-k for full generation, and discards the rest, without relying on a fixed accept/reject threshold. Operating purely at inference time, ABSS can serve as a lightweight pre-selection add-on for existing seed-optimization pipelines, enabling additional gains. Across three benchmarks, extensive experiments show that ABSS enables consistent improvements in text–image alignment and visual quality for Stable Diffusion variants, as corroborated by human preference and alignment metrics.

1 Introduction

Text-to-Image Synthesis (T2I) aims to produce realistic and semantically consistent images directly from natural language prompts. Early approaches widely leveraged text-conditioned generative adversarial networks [Reed et al., 2016, Zhang et al., 2017, Xu et al., 2018], conditional variational autoencoders [Sohn et al., 2015], and autoregressive models [Ramesh et al., 2021] to improve semantic alignment and diversity of generated images. Diffusion models [Ho et al., 2020, Song et al., 2021b, Liu et al., 2026] have driven a paradigm shift in generative modeling, establishing themselves as the leading approach through their stable training process and exceptional output quality. In particular, combining diffusion models with large-scale language/vision–language representations has driven major breakthroughs in T2I synthesis [Nichol et al., 2021, Ramesh et al., 2022, Saharia et al., 2022, Rombach et al., 2022, Balaji et al., 2022].

Despite recent and promising advances, diffusion-based T2I systems remain highly sensitive to the seed-defined initial latent, often yielding noticeably variable and unstable results across different random initializations. We refer to initial latents that reliably lead to high-fidelity, prompt-aligned, compositionally grounded, and globally coherent images as good seeds, and those that fail to do so as bad seeds. Even small perturbations to the seed initialization can significantly affect output fidelity, aesthetics, and semantic alignment [Mao et al., 2023, Li et al., 2025]. This problem is particularly evident in earlier diffusion models, but it remains far from solved in modern DiT-based [Peebles and Xie, 2023] and rectified-flow T2I models [Esser et al., 2024]. Although recent models substantially improve visual realism and make many generations appear visually appealing at first glance, bad seeds can still produce images that fail under closer inspection, with subtle yet noticeable errors in object structure, spatial layout, attribute binding, or prompt faithfulness (see Figure 1).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

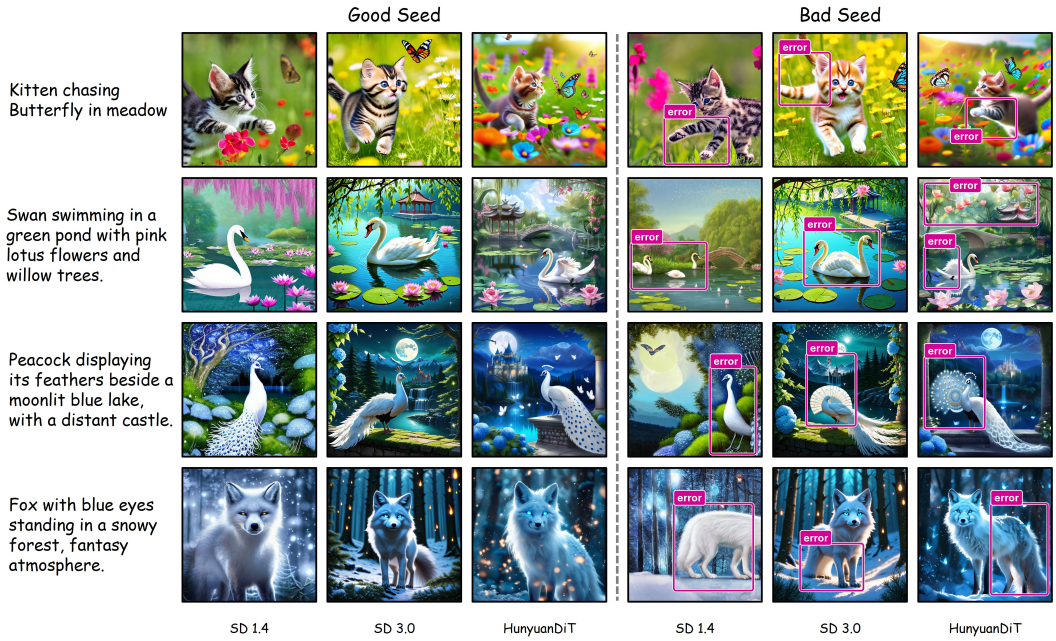

Figure 1: Illustrative examples of generations from good and bad seeds across both earlier and recent diffusion models. In earlier models such as Stable Diffusion 1.4, the gap between good and bad seeds is visually apparent, and errors from poor seeds are easily recognizable. In contrast, recent models such as Hunyuan DiT and Stable Diffusion 3.0 produce significantly higher-quality images, making outputs from bad seeds appear visually convincing at first glance. However, this improvement can obscure underlying artifacts and inconsistencies, which only become noticeable upon closer inspection. As a result, errors are more easily overlooked, potentially leading to more harmful or misleading outcomes. These hidden errors are highlighted with pink rectangles.

To mitigate this issue, prior work modifies the initial random latent determined by the seed to improve generation quality. We refer to studies in this line of research as seed optimization. Approaches include reward-based optimization with preference-guided objectives [Eyring et al., 2024, Miao et al., 2025], attention-guided refinement via cross-/self-attention control [Hong et al., 2023, Chefer et al., 2023, Guo et al., 2024], and controllable rollback techniques using inversion-based backtracking [Bai et al., 2025, Mao et al., 2025, Qi et al., 2024]. In contrast, seed selection operates over a large pool of random seeds and focuses on identifying a subset that produces consistently high-quality outputs. Xu et al. [2025] introduced the concept of "Golden Seeds" by running extensive generation trials on a validation set and selecting seeds that consistently yield superior results, which are then applied to the target set. Similarly, Li et al. [2025] select high-performing seeds from generated datasets produced by a frozen diffusion model and subsequently reuse them for fine-tuning a new diffusion model.

Motivations. The motivation of this paper stems from two key aspects. First, seeds play a substantial role in shaping the quality of outputs. Second, from a practical standpoint, a single text prompt typically requires multiple generated outputs for users to evaluate and compare. Consequently, there is a strong demand for a pool of good seeds, rather than a single seed, that can consistently produce high-quality images. Unlike prior studies on seed optimization or seed-aware training, this work focuses on the **seed selection** problem: given a collection of random seeds, how can we identify those that are likely to generate high-quality outputs?

Contribution. In this paper, we focus on the seed selection problem. Unlike existing seed selection methods that rely on auxiliary datasets, which incurs additional seed curation/fine-tuning costs, our approach distinguishes between good and bad seeds during the denoising process itself, particularly in the early stages. Our key contributions can be summarized as follows:

- We reveal a key insight: early-stage cross-attention on core tokens is a strong predictor of final prompt alignment and image quality, providing the base for deriving a simple yet effective criterion for screening and early stopping of low-quality seeds.

- Based on the above insight, we propose ABSS (Attention-Based Seed Selection), a training-free and plug-and-play procedure for selecting seeds at inference time, leveraging the denoising process itself without requiring any external supervision.
- We validate ABSS through extensive experiments across three benchmarks and multiple Stable Diffusion variants, and further show its effectiveness and generalization when extended to Diffusion Transformer based rectified-flow models.

2 Related Work

Text-to-Image Diffusion Models. T2I diffusion models have rapidly become one of the most powerful generative model families, capable of synthesizing highly diverse and photorealistic images conditioned on natural language descriptions [Saharia et al., 2022, Rombach et al., 2022]. These models typically incorporate pretrained language encoders—such as CLIP [Radford et al., 2021], T5 [Raffel et al., 2020], or more recently large language models [Balaji et al., 2022]—to transform textual inputs into dense representations. The encoded information is then injected into the generative process via cross-attention layers to align text and image semantics, highlighting the importance of token-level representations beyond computational efficiency [Vaswani et al., 2017, Rombach et al., 2022, Kong et al., 2025]. While this architecture has achieved remarkable success in controllable generation, the outputs can be highly sensitive to the initial random seeds, leading to significant quality disparity between good and bad seeds; typical failures include missing objects/parts, inappropriate overlap or misplacement, and even spurious text artifacts in inpainting [Xu et al., 2025, Shen et al., 2025]. Moreover, T2I models are known to suffer compositional errors—e.g., incorrect counts, positions, and attribute binding—yet recent evidence shows that a substantial portion of these failures is in fact seed-dependent [Li et al., 2025, Gokhale et al., 2022, Ban et al., 2025].

Seed Optimization in Text-to-Image Diffusion Models. A prominent line of work refines the seed-defined latent through attention guidance [Hertz et al., 2023], using cross-attention maps to strengthen the representation of prompt subjects. Attend-and-Excite introduces a normalized attention loss that iteratively updates the latent to mitigate subject neglect [Chefer et al., 2023]. Follow-ups further re-weight token attention or enforce coverage and consistency without retraining, targeting seed-specific generation failures [Agarwal et al., 2023, Rassin et al., 2023, Meral et al., 2024, Guo et al., 2024, Qiu et al., 2025]. These methods use prompt-dependent attention signals to guide latent updates, aiming to improve subject presence and semantic alignment during generation. A complementary approach, Golden Noise [Zhou et al., 2025], learns a prompt-conditioned "noise prompt" that transforms random noise into golden seeds using a lightweight NPNet. Its learned transformation provides an alternative mechanism for constructing initial latents before the standard denoising process begins.

Complementary to attention guidance, Z-sampling [Bai et al., 2025] revisits intermediate states through backward and forward transitions during denoising. Such transitions allow the sampler to revise an evolving generation instead of committing to a single uninterrupted trajectory. Ctrl-Z Sampling [Mao et al., 2025] uses a reward signal to control rollback decisions before refinement resumes. Their shared strategy is to intervene along the sampling path, creating additional opportunities to improve the output after initialization. These interventions also introduce extra denoising model evaluations, making the balance between generation quality and computational cost particularly important when comparing inference-time methods.

Different from prior approaches that optimize the initial latent, in this paper we consider the seed selection problem [Xu et al., 2025, Li et al., 2025]. Specifically, we perform early-stage screening of the seed pool, discarding those that are unlikely to yield faithful generations. This selection relies on the core token of the prompt—the core semantic anchor of the description—ensuring that only seeds aligned with the main concept are preserved, leading to more reliable and efficient T2I synthesis.

Impact of the Initial Seeds in SD. A growing body of evidence [Xu et al., 2025, Li et al., 2025] demonstrates that the initial seed—by fixing the starting noise—substantially steers the denoising trajectory and, in turn, the final image. Seeds induce systematic biases in outputs, affecting object arrangements (typical relative placement patterns), global style and tone, subject presence or absence, and even spurious artifacts. Consequently, the same text prompt can yield consistently good or poor results depending solely on the seed. As illustrated in Figure 1, for the same text prompt on SD 1.4 and SD 3.0, varying seeds produce outcomes that are visibly superior or inferior. This observation

has motivated a line of work [Chen et al., 2024, Qu et al., 2023, Zheng et al., 2023, Couairon et al., 2023, Jia et al., 2024, Lian et al., 2024, Kim and Kim, 2026] that tackles generation quality from the seed side—often framed as seed optimization or seed-aware guidance.

3 ABSS: Attention-Based Seed Selection

In this section, we present our method for seed selection in SD. A central challenge is that the quality of a seed, whether it is good or bad, can typically be assessed only after completing the entire diffusion process, which makes seed selection computationally expensive and logically problematic. To address this issue, we propose to monitor the diffusion process itself, with a particular focus on its early stages, to predict whether a given seed has the potential to generate desirable outputs. Specifically, we first analyze the evolution of attention patterns during denoising, which serve as a critical signal shaping the final image. Building on these observations, we then introduce our attention-based seed selection method.

3.1 Observation During Denoising

Intuitively, the entire denoising process can be viewed as a form of continuous image optimization, where each intermediate image is iteratively refined. We therefore conjecture that a successful generation should begin by establishing the main outline—capturing the core tokens that correspond to the head noun(s) denoting the primary subject(s) expected to occupy the central visual mass—followed by progressive refinement of details. By core tokens, we specifically refer to the core subject terms (e.g., *cat* and *dog* in "*a cat and a purple dog*," or *turtle* in "*a fisheye lens view of a turtle in a forest*"), rather than modifiers or contextual descriptors. Appendix A provides more details on defining core tokens and further validates the robustness of ABSS to noisy core-token annotations. In other words, high-quality seeds should prioritize anchoring the dominant subjects before attending to secondary attributes and contextual information. By contrast, low-quality seeds disperse attention prematurely across modifiers and background tokens, which is associated with missing or undersized subjects, incorrect counts, and style-dominated artifacts.

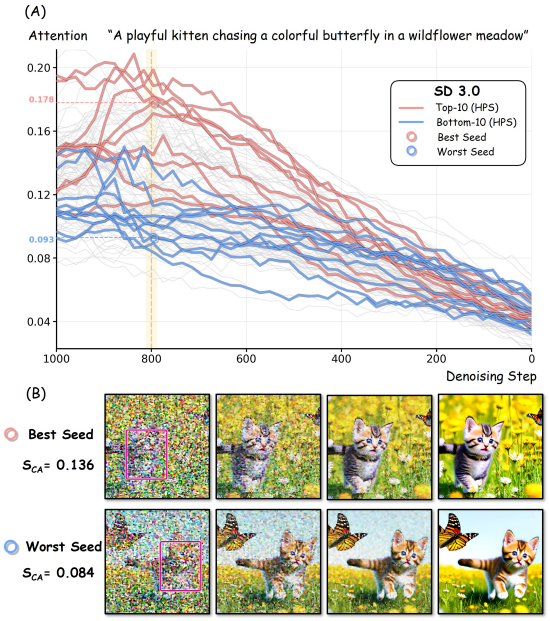

Figure 2: (A) Trends of cross-attention on core token "kitten" for 100 seeds throughout the denoising process. Red/blue curves represent high/low-quality outputs. Notably, as early as $t = 800$, good and bad seeds become roughly separable based on their cross-attention. (B) Intermediate images of the best and worst seed with corresponding cross-attention values within boxes on word kitten at timestep 800.

Motivated by these insights, we consider seed selection by executing only a small number of denoising steps and evaluating whether the process prioritizes the core tokens early on. To achieve this idea, we leverage cross-attention [Guo et al., 2024, Chefer et al., 2023] on the core tokens as the criterion for early-stage seed selection. As illustrated in Figure 2, across diverse timesteps, good seeds (red curves) consistently allocate higher attention weights to core tokens from the very beginning of the denoising trajectory, whereas bad seeds (blue curves) fail.

3.2 Method

Figure 3 illustrates the framework of our proposed Attention-Based Seed Selection (ABSS). ABSS is a lightweight and flexible seed selection method that can be seamlessly integrated into any existing SD pipeline. It identifies promising seeds by evaluating whether the intermediate images effectively capture the core tokens through the cross-attention mechanism. In the following sections, we first define the aggregated cross-attention map, and then describe how it is used to rank and select seeds.

Aggregated Cross-Attention Map. Text–image alignment in SD is achieved through cross-attention [Hertz et al., 2023], which integrates textual semantics into the denoising process. A text prompt y is first encoded by CLIP into a conditional embedding c, formulated as $c = f_{\text{CLIP}}(y)$, where $c \in \mathbb{R}^{n \times d}$ with n denoting the maximum token length and d the embedding dimension. The conditional embedding c is then linearly projected to obtain keys K and values V, while queries Q are derived from UNet activations. For a single cross-attention layer with one head and a seed s sampled from the seed pool, the attention map is computed as

$$A^s = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right). \tag{1}$$

Given a specific seed s, the element $A_t^s[h, w, i]$ represents the attention weight assigned to token i among all tokens at spatial location (h, w) in the intermediate feature map at timestep t. Higher values indicate stronger emphasis on token i.

Note that the UNet architecture in SD has multiple blocks with different resolutions (commonly $64, 32, 16, 8$), which leads cross-attention multi-head. To effectively aggregate prompt-conditioned attention, we define the aggregated cross-attention map: at a chosen spatial resolution with seed s, we collect cross-attention maps from the down/mid/up UNet blocks, stack all heads (and relevant blocks), reshape from the native (batch$\times$heads, q, n) format (with $q = h \cdot w$) to $(-, h, w, n)$, and average over the stacked dimension. This yields a single aggregated map $\bar{A}_t^s \in \mathbb{R}^{h \times w \times n}$. We then apply a temperature-scaled softmax along the token axis to sharpen the distribution,

$$\tilde{A}_t^s[h, w, i] = \frac{\exp\big(\beta\, \bar{A}_t^s[h, w, i]\big)}{\sum_{j=0}^{n-1} \exp\big(\beta\, \bar{A}_t^s[h, w, j]\big)},$$

where β is usually set to be 100. By this means, for every spatial location (h, w) we obtain a well-normalized probability distribution over tokens.

Core-token Sorting: slicing, smoothing, averaging. Given the aggregated attention map $\tilde{A}_t^s[h, w, i]$ for seed s, we focus on a subset of tokens indexed by B. Let $B \subseteq \{0, \ldots, n-1\}$ index the core tokens using the tokenizer's zero-based indexing, excluding special and padding tokens. Special tokens and padding positions are excluded from B. For each $i \in B$, we take the spatial slice, smooth it on the (h, w) grid, and average it:

$$M_{t,i}^s[h, w] = \tilde{A}_t^s[h, w, i], \;\; \widehat{M}_{t,i}^s = G_\sigma *_{\text{refl}} M_{t,i}^s.$$

Here G_σ is a normalized $(2k{+}1) \times (2k{+}1)$ Gaussian kernel, and $*_{\text{refl}}$ denotes 2D discrete convolution with reflection padding at the boundaries. Pooling over space and average all core tokens gives the final core-token concentration at step t:

$$M_t^s(B) \;=\; \frac{1}{|B|\, H\, W} \sum_{i \in B} \sum_{h=1}^{H} \sum_{w=1}^{W} \widehat{M}_{t,i}^s[h, w].$$

Empirically, we use $M_t^s(B)$ in ABSS to pre-screen seeds. Consider the prompt "*A dog walk on the street*" (illustrated in Figure 3). Given a candidate pool $\mathcal{S} = \{s_1, \ldots, s_N\}$, where N is the total number of seeds, we run only the first few denoising steps for each seed to obtain the aggregated cross-attention map $\tilde{A}_t^s[h, w, i]$, and then compute $M_t^s(B)$ for the core token `dog` at the last of these early steps. We rank seeds by this score, retain the top-K to complete the remaining generation, and discard the rest. For the top-K seeds, we continue the remaining denoising steps and obtain K output images. We also adapt our ABSS for Diffusion Transformer based architecture (See Appendix B).

4 Empirical Analysis

We conduct extensive experiments and evaluate the effectiveness and generality of ABSS through three guiding questions: *Q1: How does ABSS compare with existing seed selection and optimization strategies?* We compare ABSS against both seed-selection and seed/noise optimization baselines. For all methods, we evaluate the average k images generation quality. We further analyze their computational costs to assess the trade-off between generation quality and inference-time overhead.

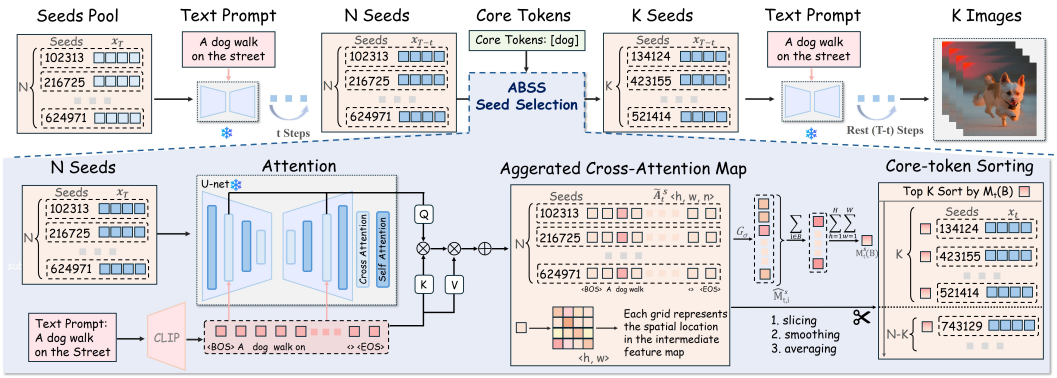

Figure 3: Framework of our proposed Attention-Based Seed Selection (ABSS). The upper pathway illustrates the standard text-to-image generation pipeline with N seeds. After a few denoising steps (t steps), ABSS is applied to evaluate whether the intermediate feature maps effectively capture the core tokens at an early stage. It then selects the top-K intermediate feature maps based on cross-attention and continues the remaining denoising steps with them.

Q2: Can ABSS serve as a plug-in to boost state-of-the-art seed or initial-noise optimization methods? We use ABSS as a seed filter while keeping each downstream optimizer unchanged to assess additive gains. *Q3: Is focusing solely on core tokens sufficient to improve performance?* We ablate non-core tokens and test whether core-token-only yields consistent improvements without additional token-level modeling.

4.1 Experimental setting

Datasets. We evaluate on three prompt suites: *DrawBench*[Saharia et al., 2022], a broad suite covering compositionality, commonsense, spatial relations, and fine-grained attributes; *InitNO* [Guo et al., 2024]; and *Pick-a-Pic* [Kirstain et al., 2023], a large-scale human-preference–oriented benchmark used for alignment studies. These datasets jointly stress both semantic faithfulness and aesthetic preference, enabling a balanced assessment of text–image generation quality.

Metrics. We report four metrics. Three are preference-oriented: HPS (v2.1) [Wu et al., 2023], ImageReward (IR) [Xu et al., 2023], and PickScore [Kirstain et al., 2023]—each trained on large-scale human-preference data and shown to correlate well with human judgments. We also include CLIP [Cherti et al., 2023, Schuhmann et al., 2022, Radford et al., 2021] text–image similarity as a proxy for semantic alignment between the prompt and the generated image. Higher values indicate better performance for all reported metrics. To measure computational cost, we report both the effective number of function evaluations (NFE) and GPU latency. We define NFE as the number of full denoising-model forward passes, where one complete pass through all DiT blocks or U-Net layers counts as one NFE. Thus, DDPM/DDIM inversion [Song et al., 2021a] steps are included in NFE whenever they invoke the full denoising model. GPU latency measures the actual wall-clock inference time, including online overheads such as latent optimization via backpropagation, auxiliary-module execution, evaluation, and seed/noise screening. All experiments were conducted on a Linux-based GPU cluster comprising 40 NVIDIA V100 GPUs with 32 GB of memory. Implementation details are provided in Appendix-C.1.

Backbones. We evaluate ABSS on a diverse collection of text-to-image generative backbones, spanning multiple generations of diffusion models and architectural paradigms. Specifically, we include Stable Diffusion (SD) 1.4, 1.5, 2.0, and 2.1 [Rombach et al., 2022], which adopt the conventional latent diffusion framework with U-Net-based denoisers. These models represent earlier generations of text-to-image synthesis and provide a basis for evaluating seed sensitivity across different model versions. To further assess the generalizability of ABSS to more advanced architectures, we consider Hunyuan-DiT [Li et al., 2024], which employs a diffusion Transformer for image generation, as well as Stable Diffusion 3 Medium and 3.5 Large [Esser et al., 2024], which incorporate multimodal diffusion Transformers and rectified flow formulations. Additionally, we evaluate FLUX.1 [Black Forest Labs, 2024], a large-scale rectified flow Transformer capable of high-quality text-to-image synthesis. Collectively, these backbones encompass substantial differences in network architectures, model capacities, and generative formulations, allowing us to systematically

Table 1: Mean performance on three datasets evaluated by HPS V2, ImageReward (IR), PickScore (PS), and CLIP. Results are averaged over three generated images per prompt under each method's corresponding evaluation setting. For each metric, the best result is highlighted in bold. In the NFE column, $*$ denotes additional module training, and † denotes extra evaluation or gradient-based latent optimization overhead.

Version	Method	*DrawBench*				*InitNO*				*Pick-a-Pic*				NFE
		HPS ↑	IR ↑	PS ↑	CLIP ↑	HPS ↑	IR ↑	PS ↑	CLIP ↑	HPS ↑	IR ↑	PS ↑	CLIP ↑	
1.4	RANDOM	0.2454	-0.2371	20.6041	0.2552	0.2711	0.0612	21.7945	0.2721	0.2425	-0.1964	20.1681	0.2531	50
	NS	0.2444	-0.2315	20.6192	0.2554	0.2716	-0.1182	21.7538	0.2754	0.2436	-0.1875	20.1977	0.2549	333
	GOLDEN	0.2490	-0.1989	20.6341	0.2561	0.2753	0.0998	21.8573	0.2720	**0.2471**	-0.1852	20.2032	0.2551	209†
	AE	0.2469	-0.1978	20.4905	**0.2569**	0.2724	0.1277	21.7440	0.2747	0.2438	-0.1984	20.0388	**0.2565**	75†
	INITNO	0.2315	-0.5618	20.2057	0.2491	0.2637	-0.2067	21.5729	0.2659	0.2382	-0.4419	19.8592	0.2489	100†
	ABSS	**0.2491**	**-0.1972**	**20.6343**	0.2552	**0.2782**	**0.1972**	**21.8881**	**0.2761**	0.2458	**-0.1504**	**20.2047**	0.2542	73
2.1	RANDOM	0.2522	0.0097	20.9931	0.2600	0.2832	0.8369	22.2425	0.2915	0.2634	0.1675	20.6432	0.2668	50
	NS	0.2596	0.0539	21.0366	0.2572	0.2889	0.6919	22.2391	0.2880	0.2703	0.1782	20.7239	0.2695	333
	GOLDEN	0.2643	0.2502	21.1316	0.2639	0.2954	1.0196	22.4292	0.2927	0.2715	**0.3218**	20.7782	0.2713	209†
	ND	**0.2663**	0.2598	21.1283	0.2645	**0.2983**	**1.2138**	22.4425	0.2950	0.2711	0.2784	20.7798	0.2731	550†
	ABSS	0.2616	**0.2778**	**21.1388**	**0.2660**	0.2972	1.0963	**22.4731**	**0.2952**	**0.2728**	0.2866	**20.7842**	**0.2738**	73
Hunyuan DiT	RANDOM	0.2982	0.8133	21.7436	0.2594	0.3451	1.7546	23.7005	0.3058	0.3162	1.1448	22.2266	0.2734	50
	NS	0.2992	0.7814	**21.8136**	0.2617	0.3470	1.7431	23.7872	0.3069	0.3189	1.1250	22.2188	0.2751	333
	GOLDEN	0.2991	0.8026	21.7349	0.2585	0.3484	**1.7598**	23.7366	0.3072	**0.3198**	1.1443	22.2199	0.2738	209†
	NPNET	0.2955	0.7093	21.7486	0.2564	0.3467	1.7561	23.7126	0.3062	0.3148	**1.1456**	22.2203	0.2649	50†*
	ABSS	**0.3002**	**0.8138**	21.7618	**0.2628**	**0.3497**	1.7591	**23.7894**	**0.3078**	0.3190	1.1446	**22.2284**	**0.2760**	71
3.5-L	RANDOM	0.3003	1.0374	22.2218	0.2822	0.3358	1.7713	23.3128	0.2921	0.3208	1.2031	22.5108	0.2836	50
	NS	0.3038	1.0991	21.9421	0.2828	0.3374	1.8234	23.3473	0.2988	0.3221	1.2142	22.5564	0.2833	333
	GOLDEN	0.3039	1.1298	22.0218	0.2819	0.3401	1.7882	23.3689	0.2935	0.3249	1.2153	22.5221	0.2814	209†
	ND	0.2951	1.0658	21.8575	0.2859	0.3369	1.8261	**23.6984**	**0.3048**	0.3231	1.2215	**22.6885**	**0.2864**	550†
	CORE^2	0.2993	**1.1623**	21.2369	0.2863	0.3339	1.8041	23.4675	0.3036	0.3196	1.1572	22.3304	0.2841	50†*
	ABSS	**0.3052**	1.1342	**22.2923**	**0.2869**	**0.3411**	**1.8297**	23.4522	0.2934	**0.3258**	**1.2362**	22.5370	0.2834	71

examine the effectiveness and generalizability of ABSS across both conventional diffusion models and modern Transformer-based generative systems. The implementation of ABSS is publicly available at `https://github.com/Cool-Rayyyy1/ABSS`.

4.2 Comparison with seed selection and optimization strategies (Q1)

We compare with a total of 8 baseline methods to answer Q1, covering both seed-selection and optimization strategies. RANDOM, GOLDEN Seeds (GOLDEN) [Xu et al., 2025], and Noise Selection (NS) [Qi et al., 2024] are used as baselines because they can be applied across different T2I backbones. RANDOM samples distinct seeds uniformly for each prompt, GOLDEN selects a small validation split of prompts for each dataset, ranks candidate seeds by their average validation performance, and then uses the top-ranked seeds for the remaining prompts. NS selects seed according to a stability criterion instead of random sampling. In the main paper, we use four representative backbones for the primary comparison: SD 1.4, SD 2.1, SD 3.5 Large, and Hunyuan-DiT. Results on the remaining backbones with baseline methods, including SD 1.5, SD 2.0, SD 3.0 Medium, and FLUX.1, are provided in Appendix C.2.

For each main backbone, we include representative optimization or refinement baselines that are applicable to the corresponding model family, covering several directions beyond prompt-independent seed selection. On SD 1.4, we compare with Attend-and-Excite (AE) [Chefer et al., 2023], which refines attention maps during denoising for better text-image alignment, and INITNO [Guo et al., 2024], which optimizes the initial noise for each prompt. On SD 2.1, we include Noise Diffusion (ND) [Miao et al., 2025], a noise-based optimization method for improving semantic faithfulness. For recent large-scale backbones, we compare with CORE^2 [Shao et al., 2026] on SD 3.5 Large, which follows an iterative collect-reflect-refine framework, and Golden Noise (NPNET) [Zhou et al., 2025] on Hunyuan-DiT, which learns to optimize diffusion noise. Detailed method settings and the corresponding NFE analysis are provided in Appendix C.1.

Compared with seed-selection baselines, ABSS provides a stronger quality-cost trade-off. RANDOM is efficient but does not exploit prompt-dependent information, while GOLDEN and NS require substantially larger search or validation overheads. In contrast, ABSS performs prompt-specific early screening and only continues the selected seeds to full denoising. As shown in Table 1, ABSS consistently improves over RANDOM across the main backbones and is competitive with or better than GOLDEN and NS on most datasets and metrics. On SD 1.4, ABSS improves the *InitNO* HPS

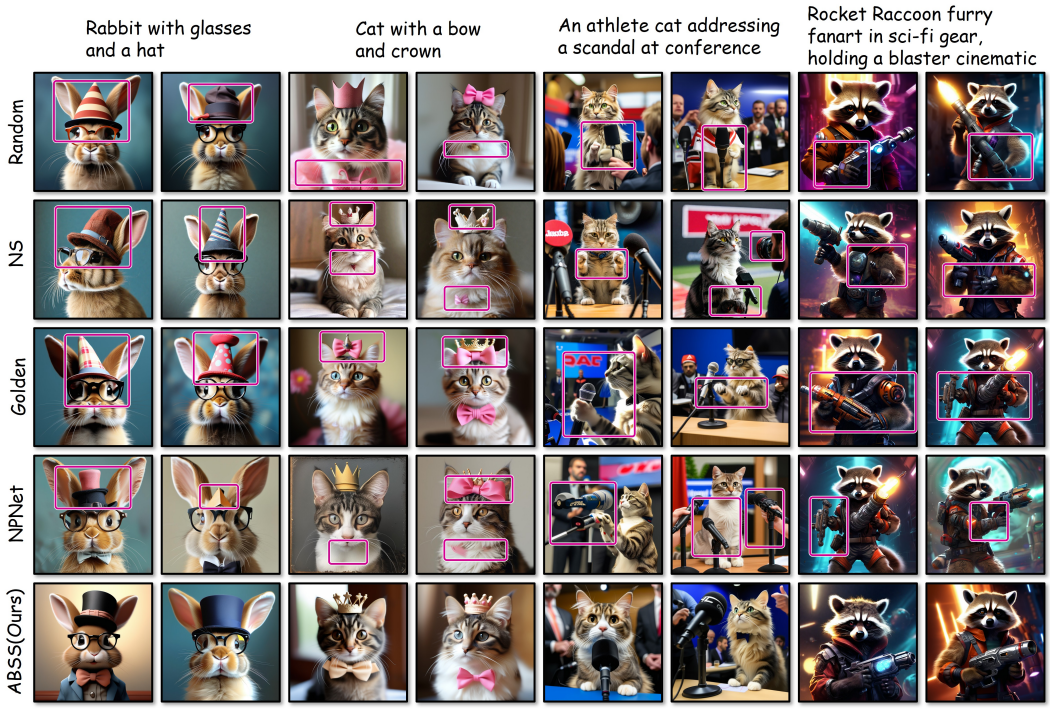

Figure 4: Qualitative comparisons on the Hunyuan-DiT backbone. Images in the same column are generated from the same prompt using different methods. Pink rectangles highlight problematic generated regions. Results with other SD backbones are provided in Appendix C.6.

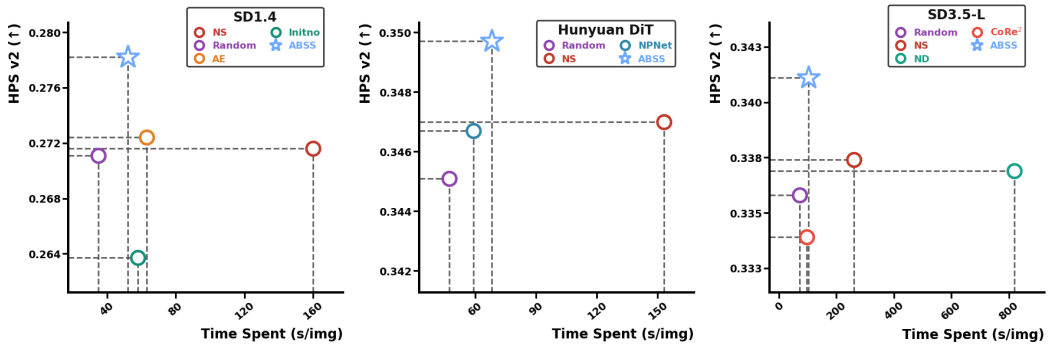

Figure 5: GPU latency comparison of average per-image generation on the *InitNO* dataset across different backbones. ABSS achieves the best quality-latency trade-off, obtaining the highest score with relatively low latency. GOLDEN is omitted because its main overhead comes from validation-set seed search, while its per-image inference cost remains standard 50-step generation.

from 0.2711 to 0.2782 and IR from 0.0612 to 0.1972 over RANDOM. On Hunyuan-DiT, ABSS leads seed-selection methods in CLIP across all datasets and HPS on DrawBench and InitNO. Importantly, this is achieved with only 71–73 NFE, compared with 333 NFE for NS and up to 209† NFE for GOLDEN, indicating that ABSS improves seed selection without introducing heavy search cost.

Compared with optimization and refinement baselines, ABSS achieves stronger overall performance with much lower inference overhead. On SD 1.4, ABSS improves over AE on the *InitNO* from 0.2724 to 0.2782 in HPS and from 0.1277 to 0.1972 in IR, while using 73 NFE instead of 75†. On SD 2.1, although ND obtains slightly higher HPS and IR on InitNO, it requires 550† NFE; in contrast, ABSS uses only 73 NFE and achieves higher PickScore and CLIP across all three datasets.

Our advantage is more pronounced on recent backbones. On Hunyuan-DiT, ABSS outperforms NPNET: it improves *DrawBench* by +0.0047 HPS, +0.1045 IR, and +0.0064 CLIP, and improves *InitNO* PickScore by +0.0768. On SD 3.5 Large, ABSS also outperforms CORE^2 on most metrics, improving *DrawBench* HPS by +0.0059, *InitNO* HPS/IR by +0.0072/+0.0256, and *Pick-a-Pic*



Figure 6: Comparison between RANDOM and ABSS seeds for seed optimization by INITNO on subject mixing and subject neglect issues.

HPS/IR/PickScore by +0.0062/+0.0790/+0.2066. While ND can be better on a few SD 3.5 Large metrics, it requires 550† NFE compared with only 71 NFE for ABSS. Importantly, the wins of ABSS are meaningful because they are achieved with both strong performance and low inference cost. Beyond the NFE comparison in Table 1, Figure 5 further shows this gap in wall-clock latency: ND is almost an order of magnitude slower than ABSS, whereas NPNET and CORE^2 are slightly faster but yield weaker overall quality. These results demonstrate that ABSS provides a stronger practical quality–efficiency trade-off. To further assess the statistical reliability of these improvements, we conduct paired t-tests across prompts, with detailed p-values reported in Appendix C.3 (Figure 8). We also conduct extensive supplementary evaluations on generation diversity, fine-grained failure categories, compositional correctness, and complex prompts, with detailed results and additional ablations presented in Appendix C.4.

4.3 Plug-in to Seed/Initial-Noise Optimization (Q2)

Here we test whether ABSS can be combined with state-of-the-art inference-time optimization to provide an extra boost. We consider two recent methods as baselines. INITNO [Guo et al., 2024] explicitly optimizes the initial noise per prompt under a chosen objective and then samples from the optimized initialization. ATTEND-AND-EXCITE (AE) [Chefer et al., 2023] is an inference-time attention-based guidance method that steers cross-attention toward subject tokens to improve prompt faithfulness without retraining. In our integration, ABSS serves only as a seed filter; baseline pipelines keep their default losses, schedulers, and hyperparameters—we simply replace uniform random seeds with ABSS-selected seeds.

Table 2: Plug-in test with ABSS-selected seeds on *InitNO*.

Method	Version	HPS ↑	IR ↑	PickScore ↑	CLIP ↑
INITNO	1.4	0.2637	-0.2067	21.5729	0.2659
+ ABSS	1.4	**0.2668**	**-0.1760**	**21.5803**	**0.2673**
INITNO	1.5	0.2642	-0.2057	21.5866	0.2651
+ ABSS	1.5	**0.2680**	**-0.1535**	**21.6412**	**0.2684**
AE	1.4	0.2724	0.1277	21.7440	0.2747
+ ABSS	1.4	**0.2803**	**0.3148**	**21.8807**	**0.2784**
AE	1.5	0.2731	0.1494	21.7586	0.2750
+ ABSS	1.5	**0.2818**	**0.3250**	**21.8958**	**0.2789**

Table 2 shows plugging ABSS into both INITNO and AE yields consistent across-the-board improvements in all four metrics, demonstrating strong compatibility with existing state of art latent optimization methods. For INITNO on SD 1.4/1.5, IR improves from -0.2067/-0.2057 to -0.1760/-0.1535 and PickScore increases from 21.5729/21.5866 to 21.5803/21.6412, with HPS/CLIP also increased. For AE on SD 1.4/1.5, the gains are markedly larger: PickScore rises from 21.7440/21.7586 to 21.8807/21.8958 and IR from 0.1277/0.1494 to 0.3148/0.3250, accompanied by clear increases in HPS and CLIP. These results indicate that ABSS supplies better starting latents and particularly strengthens inference-time attention guidance.

In Figure 6, we show that, although INITNO is designed to alleviate subject neglect (the primary subject does not appear) and subject mixing (two or more objects are conflated), certain seeds still exhibit these failures even after optimization, highlighting the limitations of directly refining unfavorable initial noise. For instance, in the horse–pink-chair case: without optimization the horse and chair are entangled; after optimization a pink "chair–horse" emerges; and in the missing-subject case the horse remains absent despite optimization. Our ABSS mitigates these failures by filtering out such bad seeds before optimization, providing more promising initial conditions and thereby strengthening INITNO or other pipelines.

4.4 Ablation on Token Type (Q3)

Beyond core tokens, we ablate alternative token types to answer Q3 and test whether core-token scoring alone is sufficient. For the prompt "*A cat with black hat yawns under the window*," we label *black* as an adjective, *cat* as a core token, *yawns* as a verb, and *under* as a preposition. On *DrawBench* (rich in adjectives and verbs), we find that scoring core tokens yields the best overall performance, improving HPS and IR over scoring alternative token types. More detailed ablation settings and results are provided in Appendix C.5. We also visualize ABSS with different token type focuses. Figure 7 shows that emphasizing *verbs* (e.g., "yawns," licks") promotes motion cues but often harms object details, yielding incomplete subjects and lower HPS. Emphasizing *adjectives* (e.g., "black," pink") spreads color/style broadly—often bleeding into background—while the main subject becomes less faithful. Focusing on *preposition* improves relative placement (e.g., window/plate context) but increases missing or fragmented foreground objects. In contrast, focusing on *core tokens* best preserves subject integrity and semantics, achieving the strongest HPS; thus, ABSS only needs early attention on core tokens.

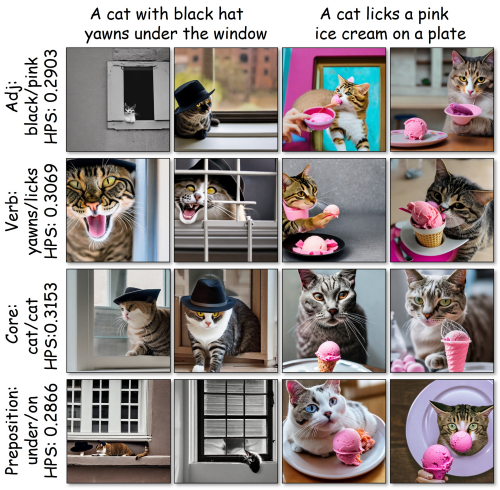

Figure 7: Generated images with different token-type focuses. Focusing on core tokens produces the most coherent images and achieves the highest HPS score.

5 Conclusion

We revisited the seed effect in T2I and showed that early cross-attention to core tokens strongly predicts prompt alignment and image quality. Building on this, we proposed ABSS, a training-free, plug-and-play procedure that globally ranks seeds via a core-token concentration score and works across different Diffusion/Flow based variants. Experiments on three benchmarks demonstrated consistent gains in faithfulness and visual quality, improving human preference and alignment metrics with minimal overhead. ABSS also serves as a lightweight preselection layer that complements seed-optimization and rollback/refinement pipelines. These results positioned attention-based seed selection as a practical primitive for robust, cost-efficient T2I at scale.

6 Limitations

Although ABSS is training-free and applicable across diffusion backbones, it requires access to early attention maps or intermediate attention statistics, potentially limiting its use in closed-source systems or architectures where such signals are inaccessible. Moreover, ABSS assumes that core subject tokens can be reasonably identified. While our robustness experiments demonstrate resilience to noisy core-token annotations, seed ranking may be less reliable for ambiguous prompts or those involving fine-grained subject relations. In particular, we observe limited improvements for highly complex prompts, suggesting that spatial mean-based attention aggregation may fail to capture intricate semantic relationships. Developing more sophisticated attention scoring strategies for such prompts remains an important direction for future work.

Acknowledgment

We gratefully acknowledge the support of the Google TPU Builders Program and Google Tunix library. We also acknowledge computational support from the Brandeis High Performance Computing Cluster. This work was partially supported by the National Science Foundation under Grants DMR-1933525 and OAC-1920147. We thank the anonymous reviewers and the area chair for their valuable feedback and constructive suggestions. The authors declare no competing interests.

References

Aishwarya Agarwal, Srikrishna Karanam, K J Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In *International Conference on Learning Representations*, 2025.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Boqing Gong, Cho-Jui Hsieh, and Minhao Cheng. The crystal ball hypothesis in diffusion models: Anticipating object positions from initial noise. In *International Conference on Learning Representations*, 2025.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 2023.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

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

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 2002.

Chengyou Jia, Minnan Luo, Zhuohang Dang, Guang Dai, Xiaojun Chang, Mengmeng Wang, and Jingdong Wang. Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Kwanyoung Kim and Sanghyun Kim. Model already knows the best noise: Bayesian active noise selection via attention in video diffusion model. In *International Conference on Learning Representations*, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 2023.

Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models – from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Enhancing compositional text-to-image generation with reliable random seeds. In *International Conference on Learning Representations*, 2025.

Zhimin Li, Jianwei Zhang, Qin Lin, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024.

Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *ACM Computing Surveys*, 2026.

Po-Yuan Mao, Shashank Kotyan, Tham Yik Foong, and Danilo Vasconcellos Vargas. Synthetic shifts to initial seed vector exposes the brittle nature of latent-based diffusion models. *arXiv preprint arXiv:2312.11473*, 2023.

Shunqi Mao, Wei Guo, Chaoyi Zhang, Jieting Long, Ke Xie, and Weidong Cai. Ctrl-z sampling: Diffusion sampling with controlled random zigzag explorations. *arXiv preprint arXiv:2506.20294*, 2025.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Boming Miao, Chunxiao Li, Xiaoxiao Wang, Andi Zhang, Rui Sun, Zizhe Wang, and Yao Zhu. Noise diffusion for enhancing semantic faithfulness in text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024.

Weimin Qiu, Jieke Wang, and Meng Tang. Self-cross diffusion guidance for text-to-image synthesis of similar subjects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *ACM International Conference on Multimedia*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 2023.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022.

Christoph Schuhmann et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Shitong Shao, Zikai Zhou, Dian Xie, Yuetong Fang, Tian Ye, Lichen Bai, Bo Han, and Zeke Xie. Improved and accelerated text-to-image generation with collect, reflect, and refine. *Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Fei Shen, Yutong Gao, Jian Yu, Xiaoyu Du, and Jinhui Tang. Imagharmony: Controllable image editing with consistent object quantity and layout. *arXiv preprint arXiv:2506.01949*, 2025.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jia Wang, Jie Hu, Xiaoqi Ma, Hanghang Ma, Yanbing Zeng, and Xiaoming Wei. MagicMirror: A large-scale dataset and benchmark for fine-grained artifacts assessment in text-to-image generation. *arXiv preprint arXiv:2509.10260*, 2025.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 2023.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Zhichao Yang, Tianjiao Gu, Jianjie Wang, Feiyu Lin, Xiangfei Sheng, Pengfei Chen, and Leida Li. Longt2ibench: A benchmark for evaluating long text-to-image generation with graph-structured annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. In *IEEE International Conference on Computer Vision*, 2025.

A Definition of Core Tokens

In ABSS, core tokens are the content-bearing words that specify the primary semantic entities in a prompt, i.e., the main objects or subjects that should dominate the image. For small prompt sets, users can directly specify them by selecting the main subject/object words (typically nouns or short noun phrases). For larger prompt collections, we automatically identify core tokens using a large language model (LLM) as a one-time preprocessing step.

LLM-based core-token identification at scale. For large prompt collections, we identify core tokens automatically using an LLM applied once per prompt. Concretely, we use GPT-5.1 with a simple instruction to extract up to three core words that denote the main objects or subjects expected to dominate the image, while labeling other content words into coarse categories (adjectives, verbs, or prepositions). This procedure is fully automatic and incurs negligible overhead compared to diffusion sampling, since it is run only once per prompt. Moreover, our results show that ABSS is robust to moderate noise in this step: even approximate core-token identification already yields consistent improvements, suggesting that the method does not rely on perfectly accurate annotation. In our implementation, processing 100 prompts with the LLM takes slightly over one minute and produces highly accurate outputs under manual inspection, further indicating that the preprocessing is both cheap and reliable.

Prompt template used for LLM extraction. We use the following prompt template for each text prompt:

Instruction. Assuming you are an expert in text-to-image prompting and visual semantics, given a text-to-image prompt, identify up to three core words that denote the main objects or subjects that would occupy most of the image. Output these as `core_tokens`. Then label the remaining content words into `adjectives`, `verbs`, and `prepositions` (ignore articles and stopwords).

Prompt 1. `Kneeling cat knight portrait finely detailed armor intricate design silver silk cinematic lighting`
Output 1 (JSON).

```
{"core_tokens": ["cat", "knight", "armor"],
 "adjectives": ["kneeling", "portrait", "finely", "detailed", "intricate",
                "silver", "silk", "cinematic"],
 "verbs": [],
 "prepositions": []}
```

Prompt 2. `A galactic eldritch squid towering over the planet Earth stars galaxies and nebulas in the background photorealistic 8k`
Output 2 (JSON).

```
{"core_tokens": ["squid", "Earth", "nebulas"],
 "adjectives": ["galactic", "eldritch", "towering", "planet",
                "photorealistic", "8k"],
 "verbs": [],
 "prepositions": ["over", "in"]}
```

We then map the extracted core words to tokenizer indices using the underlying text encoder tokenizer (e.g., CLIP BPE), and compute core-token cross-attention statistics accordingly.

Robustness to Noisy Core-Token Annotations. We further evaluate the robustness of ABSS to noisy core-token annotations. Since ABSS only relies on a small set of main subject tokens, it is expected to be relatively insensitive to moderate annotation errors. To verify this, we intentionally corrupt 50% of the core-token annotations and evaluate ABSS on FLUX.1 using the *InitNO* dataset. As shown in Table 3, ABSS still improves over random seed selection across all metrics even when half of the core-token annotations are corrupted, suggesting that its effectiveness does not rely on perfectly annotated core tokens.

Table 3: Robustness of ABSS to noisy core-token annotations on FLUX.1 using the *InitNO* dataset. We intentionally corrupt 50% of the core-token annotations. Even under this noisy setting, ABSS still improves over random seed selection across all metrics.

Setting	HPS ↑	IR ↑	PS ↑	CLIP ↑
RANDOM	0.3364	1.5945	23.8038	0.2967
ABSS (50% noisy)	**0.3382**	**1.5958**	**23.8134**	**0.2974**

B Extend ABSS to DiT based architecture

Diffusion Transformers (DiTs) have recently emerged as a strong alternative to UNet-based diffusion backbones. In particular, DiT [Peebles and Xie, 2023] demonstrates that scaling transformer capacity and training can yield highly competitive—and often state-of-the-art—generation quality, making transformer-based diffusion architectures increasingly popular in modern text-to-image systems. Given this trend, it is important to understand whether our attention-based seed scoring principle is specific to UNets, or whether it generalizes to DiT-style backbones.

In the main paper, we show that ABSS is effective on UNet-based Stable Diffusion across both the 1.x and 2.x series, where cross-attention directly reveals how image latents focus on core text tokens in early denoising. We further show that the same principle extends to DiT-style backbones: although DiT lacks UNet cross-attention blocks, its in-context conditioning jointly processes image and text tokens via self-attention, enabling an analogous image-to-core-text signal for seed scoring. Specifically, we investigate ABSS on SD 3.0 and FLUX to assess robustness across modern architectures and training objectives. Although FLUX and SD 3.0 is trained with a flow-matching objective, ABSS only relies on early-step attention under text conditioning, making it largely architecture-driven rather than objective-specific. Below we formalize the ABSS score under DiT conditioning.

Aggregated Self-Attention Map. Unlike UNet-based Stable Diffusion, DiT patchifies the latent/image into M patch tokens and concatenates them with the N text tokens, then applies standard self-attention over the joint sequence. Although DiT has no explicit cross-attention blocks, its self-attention matrix still contains a well-defined image-to-text attention signal.

Concretely, DiT represents the current latent state at timestep t as the combination of M patch image tokens $x_t^{(l-1)} \in \mathbb{R}^{M\times d}$ and the conditional prompt as N text tokens $c \in \mathbb{R}^{N\times d}$. These tokens are concatenated into a single sequence

$$X_t^{(l-1)} = \begin{bmatrix} x_t^{(l-1)} \\ c \end{bmatrix} \in \mathbb{R}^{(M+N)\times d},$$

where indices $1,\dots,M$ correspond to image patch tokens and $M+1,\dots,M+N$ correspond to text tokens. The self-attention matrix at timestep t and layer (l) is

$$A_t^{(l)} = \mathrm{softmax}\left(\frac{Q_t^{(l)}(K_t^{(l)})^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{(M+N)\times(M+N)},$$

where $Q_t^{(l)}, K_t^{(l)}$ are the query/key projections of $X_t^{(l-1)}$ and $A_t^{(l)}[u,v]$ is the attention weight from token u to token v. In particular, $A_t^{(l)}[i,\ M+j]$ with $i \in \{1,\dots,M\}$ and $j \in \{1,\dots,N\}$ measures how strongly the i-th image patch token attends to the j-th text token, analogous to the token distribution at a spatial location (h,w) in the UNet case.

Instead of stacking self-attention maps across multiple layers, we empirically find that hooking a single intermediate transformer layer $l^\star$ yields the best ABSS performance. Given a seed s and timestep t, we denote the hooked self-attention matrix by $\tilde{A}_t^s \triangleq \tilde{A}_t^{s,(l^\star)}$ and use its image-to-text entries $\tilde{A}_t^s[i,\ M+j]$ as the attention signal for seed scoring.

Core-token sorting: averaging, smoothing, and pooling. Given the hooked self-attention matrix $\tilde{A}_t^s \in \mathbb{R}^{(M+N)\times(M+N)}$ for seed s at timestep t, we first average over image tokens to obtain a scalar

score for each text token $i \in \{1, \dots, N\}$:

$$M_t^s[i] \;=\; \frac{1}{M} \sum_{j=1}^{M} \tilde{A}_t^s[j,\, M+i].$$

We then smooth this token-level score vector along the text-token axis using a normalized Gaussian kernel:

$$\widehat{M}_t^s \;=\; G_\sigma *_{\mathrm{refl}} M_t^s,$$

where G_σ is a normalized 1D Gaussian kernel and $*_{\mathrm{refl}}$ denotes discrete convolution with reflection padding. Finally, given the core-token index set $B \subseteq \{1, \dots, N\}$, we define the core-token concentration score as

$$M_t^s(B) \;=\; \frac{1}{|B|} \sum_{i \in B} \widehat{M}_t^s[i],$$

which we use in ABSS to pre-screen and rank candidate seeds for DiT-based models.

C Setup, Supplementary Experiments and Analyses

C.1 Detailed Generation and Evaluation Settings for Table 1

ABSS Attention extraction and sampling settings. We provide additional implementation details for reproducing Table 1. For SD 1.4/1.5, we aggregate cross-attention maps across all layers and heads at a spatial resolution of 16×16, while for SD 2.0/2.1 we use 24×24. For transformer-based backbones, we extract attention from a single representative transformer block: the 12th block for FLUX.1, SD 3.0, and Hunyuan-DiT, and the 18th block for SD 3.5-Large. For image generation, we follow the official inference configurations of each backbone as provided on Hugging Face, including the number of inference steps, guidance scale, and image resolution. Specifically, SD 1.x, SD 2.x, and the transformer-based backbones use resolutions of 512×512, 768×768, and 1024×1024, respectively. The inference steps and guidance scales are set according to the corresponding backbone-specific configurations rather than using a unified setting across all models.

Comparison method settings and averaging protocol. We summarize the method-specific settings for comparison methods as follows. For a fair comparison, all shared generation configurations, including image resolution, classifier-free guidance scale, and the number of inference steps, are aligned with ABSS. We also report a coarse NFE estimate for each method, where one full denoising-model forward pass is counted as one NFE.

1. GOLDEN.

   We use the first 100 prompts from *InitNO* and *DrawBench*, and the first 50 prompts from *Pick-a-Pic* as a validation set to extract the "golden seeds"; all remaining prompts are used for evaluation. Specifically, GOLDEN ranks candidate seeds by their average HPS-v2 score on the validation set, and then applies the top-ranked seeds to all test prompts in the same dataset. The validation search cost is amortized over the reported test generations. For example, on *DrawBench*,

   $$\text{NFE} = 50 + \frac{100 \times 10 \times 50}{100 \times 3} \approx 216.7\,\dagger\,.$$

   Averaged across datasets, we report the coarse NFE of GOLDEN as $209^\dagger$, where † indicates that the method involves additional HPS-v2 evaluation overhead during validation-set seed ranking.

2. NS. For a fair comparison with the ABSS setting using a 10-seed pool, we use $K{=}10$ noise candidates per prompt instead of the default candidate number of 100. Following the official implementation, each candidate is ranked by its DDIM inversion stability: we first run 50-step DDIM sampling from z_T to z_0, then perform 50-step DDIM inversion back to z'_T, and compute $\cos(z_T, z'_T)$ as the ranking score. We then select the top-3 candidates and directly report their already generated images. Thus, the coarse NFE per reported image is

   $$\text{NFE} = \frac{10 \times (50 + 50)}{3} \approx 333.3.$$

3. INITNO. We follow the official INITNO implementation, optimizing the initial noise mean and log-variance with Adam using learning rate 1×10^{-2}. We use the default attention thresholds τ_{cross}=0.2 and τ_{self}=0.3, with at most 5 restart rounds of 10 optimization steps. According to the pseudo-code in INITNO, each denoising optimization step invokes one full denoising-model forward pass. Thus, the maximum optimization cost is 5×10 NFEs, and the final image generation uses another 50 denoising steps. Since the threshold-based stopping criterion may terminate the optimization early, the actual latency can be lower in practice. For simplicity, we report the coarse NFE as

$$\text{NFE} = 5 \times 10 + 50 = 100^{\dagger},$$

where † indicates additional latent/noise optimization overhead beyond standard denoising.

4. AE. We follow the official AE latent optimization setting, using attention maps at 16×16 resolution. Latent updates are applied during the first 25 denoising steps with scale factor 20. At zero-based sampling steps 0, 10, and 20, iterative refinement is triggered when subject-token attention falls below the respective target thresholds of 0.05, 0.5, and 0.8, with at most 20 refinement iterations per triggered step. Final images are decoded after the standard 50-step denoising process. For a coarse NFE estimate, we count 50 denoising forward passes and 25 additional attention/loss forward passes, excluding any extra passes introduced by threshold-triggered iterative refinement:

$$\text{NFE} = 50 + 25 = 75^{\dagger}.$$

For simplicity, we do not include the threshold-triggered iterative refinement steps in this coarse NFE estimate, as they are only invoked when the attention loss does not satisfy the specified thresholds. Here, † indicates additional latent-optimization overhead beyond standard denoising.

5. ND. We do not follow the default ND setting, as its original configuration uses 50 optimization epochs and 50 noise candidates, which incurs very high computational cost. Instead, for a fairer comparison, we reduce the setting to 10 optimization epochs and 10 noise candidates. We otherwise follow the official ND implementation, which uses VQAScore [Lin et al., 2024] with clip-flant5-xl as the outer objective. At each optimization epoch, ND updates the noise based on the selected candidate, and the final image is generated with the same 50-step denoising setting as ABSS. The coarse NFE is therefore computed as

$$\text{NFE} = 10 \times 50 + 50 = 550^{\dagger},$$

where the first term denotes the 10 optimization epochs and the second term denotes the final 50-step image generation. Here, † indicates additional overhead from gradient-cache updates and VQAScore evaluation, which is not fully captured by the coarse NFE.

6. CORE^2. We follow the official CORE^2 implementation on SD 3.5-Large with the released noise model checkpoint. The refinement module uses a LoRA-based PromptSD35Net with 28 LoRA slots and rank 64. We set the weak-to-strong guidance scale to 1.5 and apply the refinement branch at every denoising step. For each prompt, we sample 3 images with different random seeds. Although the refinement branch is applied during inference, it does not introduce an additional full denoising-model forward pass through all blocks. Therefore, the number of full denoising steps remains 50, and we report its coarse NFE as $50^{\dagger*}$, where † means the refine step within each denoising step and $*$ indicates additional training cost for the Collect-and-Reflect refinement module.

7. NPNET. We follow the official NPNET inference pipeline and use the released pretrained noise-prompt model for each backbone. For Hunyuan-DiT, we use the DiT branch with the released `dit.pth` checkpoint to predict the golden initial noise. For each prompt, we generate 3 golden-noise samples with deterministic seeds and use the generated images as the NPNET baseline. During inference, NPNET only applies a lightweight auxiliary module once at the beginning to predict or adjust the initial noise, while the subsequent denoising process still follows the standard 50-step generation. Since this auxiliary module does not require an additional full denoising-model forward pass, we report its coarse NFE as $50^{\dagger*}$, where $*$ indicates the use of a separately trained auxiliary module whose training cost is not counted in NFE and † represents the extra initial NPNet noise optimization during inference.

8. ABSS. We use a seed pool of 10 candidates per prompt and select the top-3 seeds for final image generation. For SD 1.x and SD 2.x, attention maps are collected at the 10th denoising step across

all layers and heads, using spatial resolution 16×16 for SD 1.x and 24×24 for SD 2.0/2.1. Since this requires a full forward pass, the coarse NFE per reported image is

$$\text{NFE} = \frac{10 \times 10 + 3 \times 40}{3} \approx 73.3.$$

For FLUX.1, Hunyuan-DiT, SD 3.0, and SD 3.5-Large, we use truncated forward screening by extracting attention from an intermediate block. For example, with attention collected at the 12th block out of 30 total blocks, the coarse NFE becomes

$$\text{NFE} = \frac{10 \times (9 + 12/30) + 3 \times 40}{3} \approx 71.3.$$

Accordingly, we report the coarse NFE of ABSS as approximately 73 for the U-Net backbones and 71 for the DiT-based backbones.

C.2 Additional Results

Table 4: Additional results on the remaining backbones. We report mean top-k performance with $k = 3$ from a 10-seed pool; the best result among RANDOM, GOLDEN, and ABSS is bolded for each metric.

		DrawBench				*InitNO*				*Pick-a-Pic*			
Version	**Method**	HPS ↑	IR ↑	PS ↑	CLIP ↑	HPS ↑	IR ↑	PS ↑	CLIP ↑	HPS ↑	IR ↑	PS ↑	CLIP ↑
1.5	RANDOM	0.2428	-0.2651	20.5737	0.2569	0.2702	-0.0286	21.7658	0.2712	0.2446	**-0.1732**	20.2062	0.2541
	GOLDEN	**0.2488**	-0.1678	**20.6277**	0.2558	0.2754	0.0189	21.8361	0.2723	0.2461	-0.2261	**20.2213**	0.2546
	ABSS	0.2484	**-0.1666**	20.6194	**0.2576**	**0.2757**	**0.1191**	**21.8381**	**0.2735**	**0.2465**	-0.2119	20.2203	**0.2551**
2.0	RANDOM	0.2504	0.0214	20.9348	0.2662	0.2814	0.8255	22.1335	0.2926	0.2612	0.1038	20.5830	0.2689
	GOLDEN	0.2601	**0.2014**	**21.0551**	0.2668	0.2936	1.0140	22.3182	0.2942	**0.2705**	**0.2732**	**20.7832**	0.2703
	ABSS	**0.2604**	0.1558	21.0299	**0.2669**	**0.2942**	**1.0489**	**22.3478**	**0.2955**	0.2702	0.2578	20.7765	**0.2714**
3.0	RANDOM	0.2963	0.9782	22.0641	0.2862	0.3405	1.8090	23.6124	0.3052	0.3110	1.0498	22.0072	0.2750
	GOLDEN	0.2987	1.0203	**22.1714**	0.2878	0.3422	1.8089	23.6061	**0.3065**	0.3129	1.0630	22.0813	**0.2779**
	ABSS	**0.2991**	**1.0292**	22.1453	**0.2881**	**0.3440**	**1.8091**	**23.6398**	0.3057	**0.3144**	**1.0757**	**22.0831**	0.2753
FLUX	RANDOM	0.2971	0.7093	22.0118	0.2624	0.3364	1.5945	23.8038	0.2967	0.3094	0.7997	22.0647	0.2568
	GOLDEN	0.2986	0.6618	21.9796	0.2616	0.3382	1.6125	23.8249	0.2960	0.3088	**0.8161**	22.0442	0.2572
	ABSS	**0.2996**	**0.7273**	**22.0242**	**0.2641**	**0.3408**	**1.6517**	**23.8505**	**0.2980**	**0.3114**	0.8060	**22.0716**	**0.2581**

We additionally provide results on the remaining backbones, including SD 1.5, SD 2.0, SD 3.0 Medium, and FLUX.1, to complement the main-paper comparisons on SD 1.4, SD 2.1, SD 3.5 Large, and Hunyuan-DiT. As shown in Table 4, ABSS generally improves over RANDOM and remains competitive with or better than GOLDEN across most datasets and metrics. On SD 1.5, ABSS improves IR over RANDOM on *DrawBench* and *InitNO*, from -0.2651 to -0.1666 and from -0.0286 to 0.1191, respectively. On SD 2.0, ABSS achieves the best HPS and CLIP on *DrawBench*, and further improves *InitNO* performance, increasing IR from 0.8255 to 1.0489 over RANDOM.

The gains are also consistent on stronger backbones. On SD 3.0 Medium, ABSS obtains the best HPS and IR on *DrawBench*, improving IR from 0.9782 to 1.0292, and achieves the best HPS, IR, and PickScore on *Pick-a-Pic*. The improvement is particularly strong on FLUX.1, where ABSS outperforms both RANDOM and GOLDEN on 11 out of 12 metrics. For example, on *DrawBench*, ABSS improves HPS from 0.2971 to 0.2996, IR from 0.7093 to 0.7273, PickScore from 22.0118 to 22.0242, and CLIP from 0.2624 to 0.2641 over RANDOM. These additional results further support that ABSS generalizes across different Stable Diffusion versions and recent non-U-Net backbones.

C.3 Additional Evidence for Baseline Comparisons for Q1

To assess the stability of our conclusions, we provide additional statistical and diagnostic evidence for ABSS beyond the main comparisons. Specifically, we validate ABSS from three complementary perspectives: (i) paired significance testing on the overall metrics, (ii) ranking agreement with the HPS-induced seed ordering, and (iii) sensitivity to the evaluation timestep t.

Paired t-test on Table 1. We further conduct paired t-tests on the entries in Table 4 and parts of entries in Table 1 across prompts. For each backbone and metric, we perform a paired t-test between ABSS and the corresponding baseline using per-prompt scores, and report the resulting p-values in

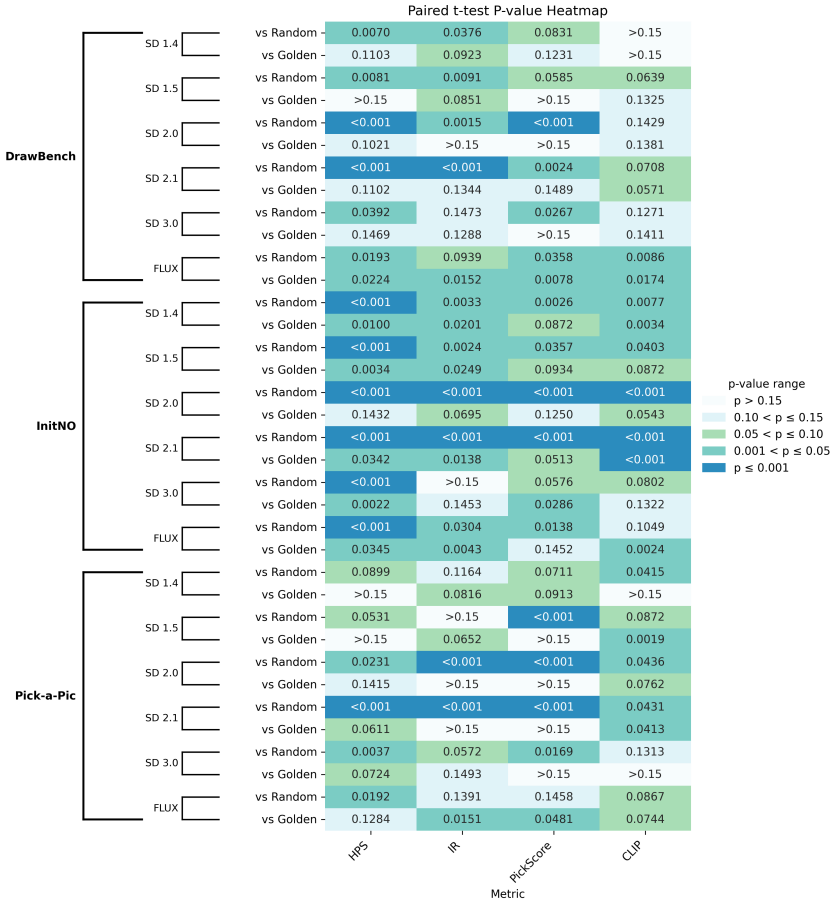

Figure 8: Paired t-test p-values for the entries in Table 4 and additional comparisons (computed across prompts).

Figure 8. For readability, values below 10^{-3} are reported as < 0.001 (omitting additional digits), and values above 0.15 are omitted since they do not indicate sufficient statistical significance. This analysis provides a quantitative measure of whether the observed improvements are consistent across prompts.

Overall, the improvements of ABSS over the RANDOM baseline on *DrawBench* and *Pick-a-Pic* are relatively stable across backbones and metrics, with many comparisons yielding statistically significant p-values. In contrast, comparisons against the GOLDEN baseline are largely significant and reliable but exhibit a few less consistent cases. This behavior is expected: "golden seeds" are typically obtained via substantially more compute-intensive procedures, whereas ABSS is training-free and incurs negligible additional computation at inference time. Therefore, achieving comparable or frequently significant performance against GOLDEN provides strong evidence that ABSS is a reliable and cost-effective alternative.

Notably, ABSS is particularly effective on *InitNO*. Across all *InitNO* entries in Figure 8 (12 rows $\times$ 4 metrics = 48 tests), 41/48 p-values are below 0.10 (and 33/48 are below 0.05), indicating that the gains are highly consistent across prompts. Additionally, the newest DiT backbones (SD 3.0 and FLUX) show strong performance across the three datasets, with FLUX in particular exhibiting a high fraction of significant comparisons, further supporting the robustness of ABSS on modern diffusion and flow backbones.

Ranking agreement with HPS. To further validate ABSS based on the observations in Figure 2, we compare the ABSS seed-ranking list with the ground truth HPS ranking using (i) the overlap rate—the proportion of seeds common to both lists—and (ii) Normalized Discounted Cumulative Gain (NDCG) [Järvelin and Kekäläinen, 2002], which assesses how closely an ordering matches the ideal one while applying position-based discounts. ABSS achieves an NDCG above 0.9 and an

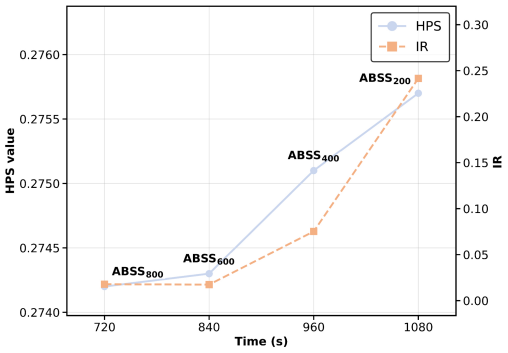

Figure 9: Performance and running time of ABSS on *InitNO* with different timesteps.

Table 5: Generation quality and diversity on Hunyuan-DiT across three datasets. Diversity results are reported as mean ± standard deviation over 10 independent test cases.

	DrawBench			*InitNO*			*Pick-a-Pic*		
Method	HPS ↑	MSS ↓	Vendi ↑	HPS ↑	MSS ↓	Vendi ↑	HPS ↑	MSS ↓	Vendi ↑
Random	0.2982	0.5331±0.0059	2.7180±0.0143	0.3451	0.5446±0.0070	2.7007±0.0160	0.3162	0.5562±0.0062	2.6740±0.0151
ABSS	**0.3002**	0.5348±0.0059	2.7153±0.0152	**0.3497**	0.5476±0.0067	2.6935±0.0152	**0.3190**	0.5547±0.0058	2.6778±0.0114

overlap rate of 0.68, indicating that it effectively prioritizes valuable seeds and closely approximates the ground truth ranking.

Effect of evaluation timestep t. We also investigate ABSS at different timesteps and observe that performance metrics generally improve as denoising progresses. As illustrated in Figure 9, HPS increases from 0.2742 to 0.2757 and IR from 0.0178 to 0.2416 as the evaluation timestep decreases from $t = 800$ to $t = 200$, corresponding to progressively later stages of denoising. This trend suggests that later stages provide more informative signals for seed selection, but require additional computation before screening can be performed. Evaluating at $t = 800$ enables earlier seed selection at lower computational cost, offering a practical trade-off between generation quality and efficiency.

C.4 Additional Evaluations and Analyses

To further investigate the reliability and generalizability of ABSS, we conduct additional evaluations covering output diversity, fine-grained compositional correctness, and complex prompts. These analyses complement the main experimental results and provide further evidence of the effectiveness of early core-token attention for seed selection.

Generation Quality and Diversity. An important consideration in seed selection is whether improving generation quality reduces output diversity. To investigate this, we compare ABSS with random seed selection on Hunyuan-DiT across three datasets, using 10 independent test cases. We evaluate diversity using Mean Self-Similarity (MSS, lower is better) and Vendi Score [Friedman and Dieng, 2023] (higher is better), alongside HPS for generation quality. As shown in Table 5, ABSS consistently improves HPS while maintaining diversity metrics close to those of random selection. The differences in MSS and Vendi Score are small relative to their reported standard deviations and show no consistent direction across datasets. These findings suggest that the quality improvements of ABSS are not accompanied by systematic diversity reduction.

Fine-Grained Failure-Category Analysis. To assess whether the effectiveness of ABSS extends beyond coarse subject grounding, we construct a 200-prompt evaluation set covering four categories: attribute binding, spatial relations, object structure, and local artifacts, with 50 prompts per category. Prompts are selected from T2I-CompBench [Huang et al., 2023], GenEval [Ghosh et al., 2023], and MagicBench [Wang et al., 2025] according to their corresponding categories. We evaluate Hunyuan-DiT using the same pool-10/top-3 selection setting and average results over two independent test cases. Attribute binding is evaluated using BLIP-VQA [Li et al., 2022], relations and structure follow the corresponding benchmark protocols, and local artifacts are assessed using MagicAssessor-7B [Wang et al., 2025]. As shown in Table 6, ABSS improves attribute binding, relational correctness, and

local-artifact quality, while achieving comparable structural correctness. These results indicate that the benefits of early core-token attention are not restricted to detecting missing subjects.

Table 6: Fine-grained failure-category evaluation on Hunyuan-DiT. Higher values indicate better performance.

Method	Attribute ↑	Relation ↑	Structure ↑	Artifact ↑
Random	0.6404	0.1821	**0.3116**	0.8833
ABSS	**0.6482**	**0.2008**	0.3114	**0.9200**

Compositional Correctness. We further evaluate ABSS on Comp90 [Li et al., 2025] using FLUX.1, sampling 50 numerical and 50 spatial prompts. Following the benchmark protocol, we use GPT-4o to evaluate compositional correctness and compare ABSS against RANDOM and GOLDEN. As shown in Table 7, ABSS achieves the best performance across numerical accuracy, numerical error, and spatial accuracy. These results provide additional evidence that seed selection based on early core-token attention can improve compositional faithfulness.

Table 7: Compositional correctness on Comp90 using FLUX.1. Results cover numerical counting and spatial relations.

	Numerical		Spatial
Method	Acc-G ↑	MAE-G ↓	Acc-G ↑
Random	41.3	1.27	19.6
Golden	50.5	1.08	35.2
ABSS	**52.1**	**0.94**	**37.6**

Evaluation on Long and Complex Prompts. We further investigate the applicability of ABSS to more challenging text descriptions by sampling 200 long and complex prompts from LongT2IBench [Yang et al., 2026]. These prompts contain richer semantic interactions and compositional requirements than simple subject-centric descriptions. We evaluate FLUX.1 under the same seed-selection setting as our main experiments. Table 8 shows that ABSS improves all four metrics over random seed selection and remains competitive with GOLDEN. Although the gains are relatively modest, these findings suggest that the early attention signal remains useful for longer prompts, while also leaving room for improved modeling of complex semantic relationships.

Table 8: Evaluation on 200 long and complex prompts from LongT2I using FLUX.1.

Method	HPS ↑	IR ↑	PS ↑	CLIP ↑
Random	0.3017	0.7461	22.2667	0.2759
Golden	0.3031	0.7630	**22.2888**	0.2786
ABSS	**0.3036**	**0.7679**	22.2877	**0.2798**

C.5 Additional Ablations on Token Types for Q3

Table 9: Performance of ABSS with different token types on *DrawBench* using SD 1.4.

Method	HPS ↑	IR ↑	PS ↑	CLIP ↑
ABSS_{adj}	0.2440	-0.2374	20.8490	0.2617
$\text{ABSS}_{\text{verb}}$	0.2438	-0.2272	20.8345	0.2609
$\text{ABSS}_{\text{core}}$	**0.2473**	**-0.2195**	**20.8702**	**0.2619**

Experimental setting. We conduct this ablation on *DrawBench* using SD 1.4. *DrawBench* prompts typically contain richer and more diverse modifiers (e.g., adjectives) and actions (verbs) than the other benchmarks, making it a suitable testbed for studying whether ABSS should score only core

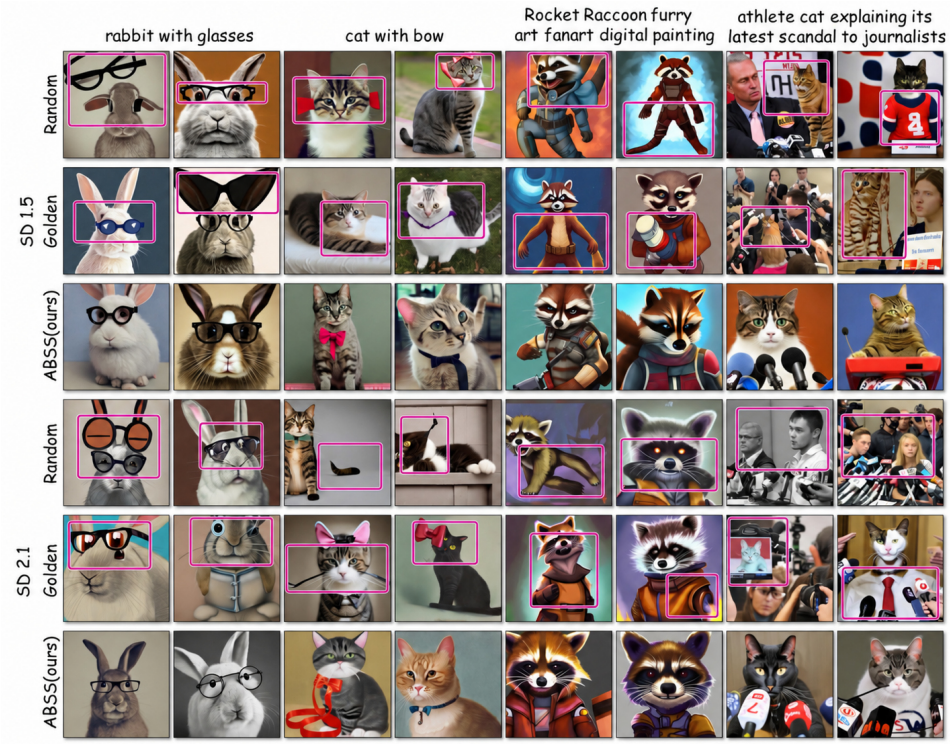

Figure 10: Additional qualitative comparisons on earlier Stable Diffusion backbones. Images in the same column are generated from the same prompt using RANDOM, GOLDEN, and ABSS. Pink rectangles highlight problematic generated regions.

subject tokens or alternative token types. We follow the same generation and evaluation settings as in Appendix Section C.1, except that we use a larger seed pool (N=100) and retain top-k=50 seeds per prompt to better reveal the general trend and reduce randomness. In Table 9, ABSS_{adj} denotes scoring adjective tokens only, $\text{ABSS}_{\text{verb}}$ denotes scoring verb tokens only, and $\text{ABSS}_{\text{core}}$ denotes scoring core tokens (our default). Metrics are averaged over the selected top-k generations.

Results. As shown in Table 9, core-token scoring achieves the best overall performance across metrics. Compared with scoring adjectives or verbs alone, $\text{ABSS}_{\text{core}}$ consistently improves HPS and IR (and also yields slightly better PickScore and CLIP). This supports our main hypothesis that early-stage attention to the dominant subjects provides the most reliable signal for seed quality, whereas focusing on modifiers or actions can over-emphasize secondary attributes and context, leading to weaker semantic grounding of the primary objects.

C.6 Additional Qualitative Results on Earlier SD Backbones.

We provide additional qualitative comparisons between ABSS, RANDOM, and GOLDEN on earlier Stable Diffusion backbones. These examples complement the main-paper visual results on Hunyuan-DiT and show that ABSS can reduce common generation errors across different model versions. Problematic regions are highlighted with pink rectangles.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe ABSS as a training-free seed selection method based on early cross-attention signals. The stated claims are supported by the experimental results across datasets, model versions, and metrics.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of ABSS in Section 6, including its dependence on early attention signals, core-token extraction, and additional seed-screening overhead.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not present formal theoretical results, theorems, or proofs. The proposed method is motivated by empirical observations during denoising and validated through experiments and ablation studies.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the key information needed to reproduce the main experiments, including datasets, model versions, seed-pool construction, selection protocol, inference settings, evaluation metrics, and comparison method settings. Additional implementation details and reproducibility information are provided in the appendix and released code.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide public access to our code and experimental prompts through a GitHub repository, along with instructions for reproducing the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the necessary experimental details, including datasets, model versions, inference settings, seed-pool size, evaluation metrics, and baseline configurations. ABSS is training-free, so no optimizer or training hyperparameters are required.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report averaged results over multiple runs or generated samples and provide statistical significance analysis in Appendix Section C.3. We also include robustness experiments showing that ABSS still improves performance when core-token annotations are mistakenly corrupted.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute resources and efficiency information used in our experiments, including GPU type, memory, GPU latency, and NFE. We also discuss the additional early-denoising cost introduced by candidate seed screening in ABSS.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe that the research reported in this paper conforms to it. The reported experiments use existing public benchmarks, pretrained diffusion models, and automated evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the broader impacts of improving text-to-image generation, including potential positive effects such as better controllability and efficiency, as well as potential negative uses such as generating misleading or harmful visual content. Since ABSS is a seed selection method rather than a new generative model, these impacts mainly arise through its use with existing diffusion systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new pretrained generative model, scraped dataset, or other high-risk asset.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite and credit the existing datasets, pretrained diffusion models, evaluation metrics, and baseline methods used in our experiments. We follow their corresponding licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release the implementation of ABSS and the core-token annotations used in our experiments as new assets. Documentation and reproduction instructions are provided alongside the released repository, while the prompts are from public benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or human-subject studies. The core-token annotations are used for method implementation and robustness analysis, rather than collected through a participant study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or human-subject studies, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We describe the use of an LLM for core-token extraction and annotation in Appendix Section A. We also evaluate the robustness of ABSS to noisy or mistaken core-token annotations.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.