Ctrl-Z Sampling: Scaling Diffusion Sampling with Controlled Random Zigzag Explorations

Shunqi Mao, Wei Guo, Chaoyi Zhang, Jieting Long, Ke Xie, and Weidong Cai

School of Computer Science, The University of Sydney, Australia
{shunqi.mao, wei.guo, chaoyi.zhang, jieting.long, kxie0655, tom.cai}@sydney.edu.au

Abstract. Diffusion models generate conditional samples by progressively denoising Gaussian noise, yet the denoising trajectory can stall at visually plausible but low-quality outcomes with conditional misalignment or structural artifacts. We interpret this behavior as local optima in a surrogate quality landscape: Once early denoising commits to a suboptimal global structure, later steps mainly sharpen details and seldom correct the underlying mistake. While existing inference-time approaches explore alternative diffusion states via re-noising with fixed strength or direction, they exhibit limited capacity to escape steep quality plateaus. We propose Controlled Random Zigzag Sampling (`Ctrl-Z` Sampling), a scalable sampling strategy that detects plateaus in quality landscape via a surrogate score, and allocates exploration only when a plateau is detected. Upon detection, `Ctrl-Z` Sampling rolls back to noisier states, samples a set of alternative continuations, and updates the trajectory when a candidate improves the score, otherwise escalating the exploration depth to escape the current plateau. The proposed method is model-agnostic and broadly compatible with existing diffusion frameworks. Experiments show that `Ctrl-Z` Sampling consistently improves generation quality over other inference-time scaling samplers across different NFE budgets, offering a scalable compute-quality trade-off. Code available at: `https://github.com/ShunqiM/Ctrl-Z-Sampling`.

Keywords: Diffusion Models · Sampling Strategies · Inference-Time Scaling

1 Introduction

Diffusion models are a powerful class of generative models, achieving state-of-the-art performance across a wide range of generative tasks, including image [45,48], video [3,23], text [34], audio [16,31], 3D object [46], along with a growing number of domains [4, 6, 27, 64]. These models gradually transform Gaussian noise into data samples through an iterative denoising process [21, 55, 56]. Modern variants are typically designed to accept conditional inputs, enabling guided generation based on labels, texts, or other structured signals [2, 59, 67].

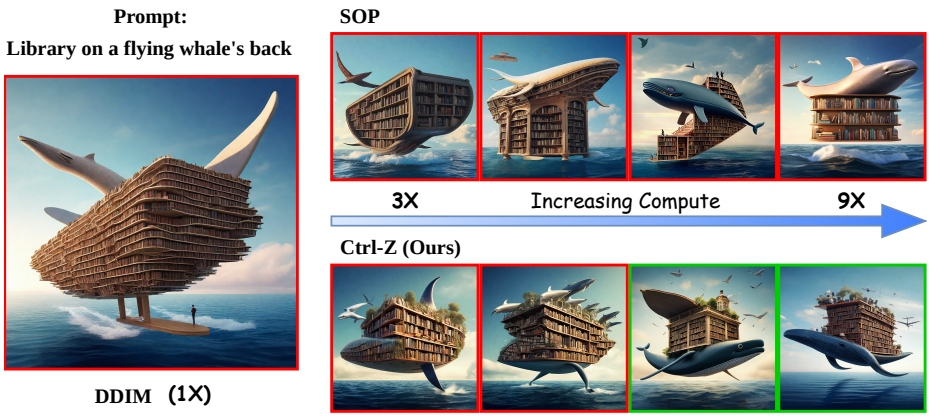

Fig. 1: We introduce `Ctrl-Z` Sampling, a diffusion sampling strategy that scales with compute to improve generation quality. Under comparable budgets, it outperforms inference-scaling baselines such as Search-over-Path (SOP). For the prompt *"Library on a flying whale's back"*, SOP produces misaligned compositions (*red*), while `Ctrl-Z` generate prompt-consistent (*green*) output as compute increase from 3× to 9× NFEs.

Despite their strong generative performance, diffusion models often exhibit semantic misalignment or global inconsistency in conditional generation. Taking image generation as an example, the denoising process is an iterative refinement procedure on the data manifold. In practice, this refinement can stagnate in suboptimal regions where the current sample already looks locally plausible but remains semantically flawed (*e.g.*, missing objects, incorrect relations, or anatomically implausible details). We refer to this sampling stagnation or collapse as converging to a *local optima* in a conceptual sample quality space, where quality reflects how faithful, structurally correct, and well aligned the image is with the text condition. This notion of optimality is defined by a quality objective rather than the diffusion model likelihood. Such failures are commonly triggered by misaligned initial noise, which misguides the trajectory toward outputs that are visually compelling but misaligned with the intended condition [63].

Some existing works mitigate these failures by reinforcing conditional information during sampling. Classifier-free guidance (CFG) [22] amplifies the conditioning signal, while other approaches alternate conditional denoising with unconditional inversion steps [1, 40] to trade off plausibility for improved conditioning. These strategies can strengthen conditional influence, yet their intervention is often loosely controlled and may be insufficient when progress stalls on broad plateaus. Recent test-time scaling methods allocate additional inference compute via score-guided exploration, for example applying small mutations to the current state or re-noising for a fixed number of steps before re-denoising [19, 41, 47, 68]. However, exploration is typically shallow, so escaping steep or broad local optima often requires substantially increasing the exploration budget. Moreover, exploration is commonly applied at every step or at

preset intervals, which can spend compute broadly while still failing to trigger sufficiently non-local moves when stagnation occurs. As a result, these methods can be impractical for local inference settings.

In this paper, we propose Controlled Random Zigzag Sampling (`Ctrl-Z` Sampling), a diffusion sampling strategy that alleviate above issues and delivers consistent quality gains across different NFE budgets (as shown in Figure 1). Specifically, `Ctrl-Z` Sampling escapes local optima through randomized backward exploration into adaptively noisier timesteps. `Ctrl-Z` Sampling detects potential local maxima using a stagnation criterion on the preference scores from a reward model, which we use as a surrogate for conceptual sample quality. Upon detection, it injects noise and reverts to an earlier, noisier timestep to explore beyond the current quality plateau. The reward model evaluates each candidate state, and `Ctrl-Z` Sampling accepts only those that yield more promising future trajectories. When no improvement is found, `Ctrl-Z` Sampling retreats to deeper noise levels to increase the chance of escape. Overall, `Ctrl-Z` Sampling induces a controlled zigzag trajectory that alternates between forward refinement and backward exploration, improving both condition alignment and visual quality.

`Ctrl-Z` Sampling is broadly applicable across diffusion backbones. Experiments on text-to-image benchmarks show that `Ctrl-Z` Sampling consistently improves generation quality across multiple metrics under low-budget NFEs, providing a controllable quality-compute trade-off via its exploration strength. This suggests `Ctrl-Z` Sampling as a practical test-time scaling alternative for single-device, per-sample inference, where large candidate pools or extreme NFE budgets are often infeasible. To summarize, our key contributions are:

- We interpret conditional diffusion sampling as a hill-climbing-like process in a surrogate sample-quality space, and empirically show that existing strategies can stall on broad plateaus that exhibit conditional misalignment or global inconsistency due to insufficient exploration depth.
- We propose Controlled Random Zigzag Sampling (`Ctrl-Z` Sampling), a reward-guided diffusion sampler with controlled explorations for adaptive deeper exploration, enabling escape from local optima during generation.
- We conduct extensive experiments demonstrating that `Ctrl-Z` Sampling substantially improves text-to-image generation quality. Results show that, unlike search methods relying on numerous shallow trials across many directions, `Ctrl-Z` Sampling achieves better exploration efficiency by making fewer, progressively deeper steps in the generation landscape.

2 Related Works

2.1 Diffusion Models Inference Techniques

Diffusion models generate data by gradually transforming Gaussian noise into structured samples through iterative denoising [21, 55, 56], with both U-Net-based [45, 48, 50] and Transformer-based architectures [33, 44] achieving strong results across modalities. However, training or finetuning remains costly, and

constrained by computational demands [12,58]. To address these challenges, various inference-time techniques have been proposed to enhance conditional generation [9]. Classifier-free guidance (CFG) [8,22] contrasts conditional and unconditional predictions to improve alignment, though at the risk of fidelity loss. Others guide generation by modifying attention maps during sampling [5,14,24,36], but these can struggle with abstract or global conditions. While pruning techniques for diffusion models facilitate efficient inference by reallocating saved compute to additional denoising steps for enhanced generation quality [13,32,53], they do not improve per-step denoising quality independently. More recent methods exploit the sensitivity of generation to noise initialization, leveraging information leakage from noisy priors [11,38] and developing strategies to optimize or select noise vectors [17, 42, 51, 57, 63, 70], or inject low-frequency condition cues [18]. However, guided optimization of noise latents can be inefficient and may also lead to degenerate solutions by collapsing the denoising process. In contrast, `Ctrl-Z` Sampling improves conditional alignment without retraining or costly optimization by adaptively enhancing exploration strength, enabling escape from broad local maxima and improving generation quality at inference.

2.2 Diffusion Sampling Strategies

DDPM [21] introduces diffusion sampling as a stochastic reverse process, which DDIM [56] later reformulates as a deterministic ODE for faster inference. Building on this, recent methods such as DPM-Solver [39], UniPC [69], and AYS [49] improve both efficiency and quality through high-order ODE solvers and adaptive step sizes, while some other work incorporates directional guidance [7, 15, 60, 66]. Later works improve sampling quality via latent-space inversion or zigzag explorations [1, 40, 65]. In parallel, guided search by reward models has been explored via particle-based sampling [29, 35] and beam search [43]. More recent efforts extend these toward test-time scaling in diffusion models [19, 41, 47, 54, 68], aiming to effectively improve sampling performance with more function evaluations. However, these approaches primarily focus on enlarging the candidate pool through shallow, local perturbations, often neglecting the role of exploration depth and thus remaining vulnerable to local optima. In contrast, concurrent works [26, 68] adopt DFS-style search strategies that roll back diffusion states to significantly earlier, noisier timesteps, increasing exploration depth but requiring costly re-denoising from these deeper noise levels. In contrast, `Ctrl-Z` Sampling mitigates this by applying controlled exploration only upon detecting stagnation, and adaptively increases inversion depth until improvement is found, avoiding unnecessary computation elsewhere. Additional relevant sampling and scaling methods are discussed in the Supp. Section A.

3 Preliminaries

3.1 Diffusion Process

Diffusion models are generative frameworks that synthesize data by progressively denoising a sample drawn from a Gaussian prior. Let $\mathcal{N}$ represent the standard

normal distribution and $\mathcal{D}$ denote the target data distribution. Starting from a noise sample $x_T \sim \mathcal{N}$, a sample $x_0 \in \mathcal{D}$ is generated through a sequence of T denoising steps, conditioned on x_T and auxiliary prompt c: $x_0 = \Phi(x_T, c)$, where Φ denotes the sampling procedure, which may be stochastic or deterministic.

3.2 Deterministic Sampling with DDIM

We adopt the DDIM sampling framework [56], which enables efficient and deterministic generation by removing stochasticity from the denoising process. Given a predefined noise schedule $\{\beta_t\}_{t=1}^T$, let $\alpha_t = \prod_{i=1}^t (1-\beta_i)$ denote the cumulative product. The generation process is decomposed into T successive mappings:

$$x_0 = \Phi^1 \circ \Phi^2 \circ \cdots \circ \Phi^T(x_T, c), \tag{1}$$

where each mapping Φ^t transforms x_t into x_{t-1} using the predicted noise $\epsilon_\theta^t(x_t, c)$:

$$x_{t-1} = \Phi^t(x_t, c) = \sqrt{\alpha_{t-1}} \cdot \frac{x_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^t(x_t, c)}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta^t(x_t, c). \tag{2}$$

Note the σ term in DDIM is zeroed and omitted for determinism. The unscaled first term in Equation (2) is a Tweedie-style posterior-mean estimation [10] of the clean sample corresponding to x_t, which we denote as:

$$\hat{x}_0^{t-1}(x_t, c) = \frac{x_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^t(x_t, c)}{\sqrt{\alpha_t}}. \tag{3}$$

This intermediate estimate serves as a useful proxy for the final output and is used in our method to assess semantic alignment via a reward model. Although Φ^t formally returns only x_{t-1}, we compute and reuse $\hat{x}_0^{t-1}$ jointly for efficiency.

3.3 Latent Inversion

DDIM defines a deterministic denoising trajectory but lacks an explicit inversion re-noising mechanism. To reintroduce noise in a controlled manner, we define an inversion operator $\Psi(x_t, \Delta)$, which simulates a forward transition from x_t to a noisier latent state (or pixel-space state, for pixel diffusion models) $x_{t+\Delta}$ using:

$$x_{t+\Delta} = \Psi(x_t, \Delta; \epsilon) = \sqrt{\frac{\alpha_{t+\Delta}}{\alpha_t}} \cdot x_t + \sqrt{1 - \frac{\alpha_{t+\Delta}}{\alpha_t}} \cdot \epsilon, \tag{4}$$

where $\alpha_t = \prod_{i=1}^t (1-\beta_i)$ is the cumulative noise schedule inherited from the original diffusion process, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ a random noise.

This operator enables transitions into higher-noise manifold of the latent space while retaining denoising progress. It provides a mechanism for test-time exploration that perturbs the generation trajectory without completely discarding previously accumulated semantic structure.

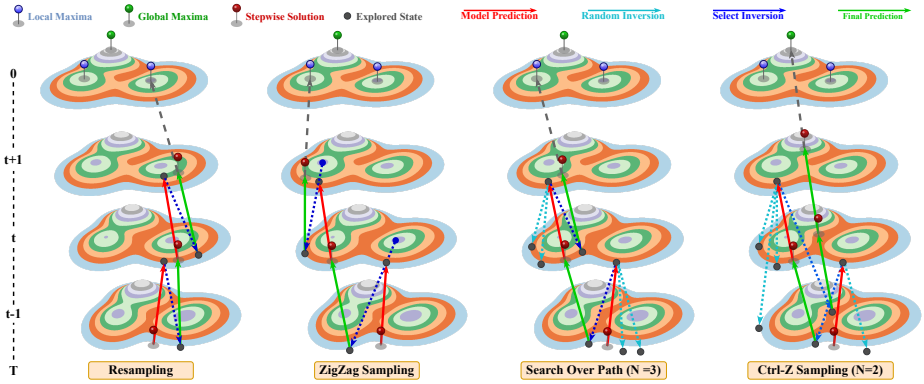

Fig. 2: Illustration of different sampling strategies during the diffusion process. The diffusion trajectory is depicted as ascending a rugged, conceptual quality landscape for sample quality and condition alignment, measured by a surrogate score. Each strategy begins with a standard denoising update (***red***), followed by either a single or multiple inversion explorations. The selected/executed inversions are marked in ***blue***, while discarded ones are in ***cyan***. The accepted inversion is followed by a forward conditional denoising step (***green***). Subsequent steps are omitted and shown in *gray*. Unlike prior methods, the proposed `Ctrl-Z` Sampling adaptively increases exploration strength through larger inversion steps when local perturbations fail to reveal improved trajectories, enabling escape from broader local maxima. N denotes the number of inversion candidates per exploration stage.

4 Methods

In this section, we first develop the intuition for `Ctrl-Z` Sampling through a hill-climbing-like view in surrogate quality space in Section 4.1, and then present its technical details; the full procedure is summarized in Algorithm 1.

4.1 Denoising Trajectories in Surrogate Quality Space

As illustrated in Figure 2, we analyze deterministic sampling trajectories through a surrogate quality objective. Conceptually, we posit a quality functional $Q(x, c)$ and approximate it with an off-the-shelf scorer $R(x, c)$. At step t, sampling yields a clean prediction $\hat{x}_0^{(t)}$ and a score $r_t = R(\hat{x}_0^{(t)}, c)$, forming a trajectory $\{r_t\}_t$ over timesteps. In practice, the score trajectory often plateaus: the sample remains visually plausible under the base model, yet fails to improve under R. We refer to such plateaus as *local optima in surrogate quality space*, meaning that within a bounded exploration budget the nearby continuations explored from the current state do not yield sufficient score gains. From this perspective, standard deterministic sampling behaves like a greedy local search that can become trapped on broad plateaus (as visualized in the trajectory analysis in Supp. Figure 8).

In the illustrated example, Resampling [40] applies a random, shallow random inversion to perturb the state and probe nearby alternatives, which may be in-

Algorithm 1 `Ctrl-Z` Sampling

Input: Denoising operator Φ^t, clean sample estimate $\hat{x}_0^{t-1}(x_t, c)$, inversion operator $\Psi(x_t, \Delta; \epsilon)$, condition c, total steps T, reward model R, exploration window λ, accept threshold δ, max inversion depth $d_{\max}$, max number of candidates N.
Output: Final image estimate x_0
Sample Gaussian noise x_T
Initialize reward score $r_{\text{prev}} \leftarrow -\infty$
for $t = T$ **to** 1 **do**
  $x_{t-1} \leftarrow \Phi^t(x_t, c)$; $\hat{x}_0^{t-1} \leftarrow \hat{x}_0^{t-1}(x_t, c)$ # Equation (2, 3)
  if $t > T - \lambda$ **then**
    $r \leftarrow R([c], \hat{x}_0^{t-1})$
    if $r \geq r_{\text{prev}} + \delta$ **then**
      $r_{\text{prev}} \leftarrow r$
    else
      inversion_step $\leftarrow 1$, best_score $\leftarrow r$, best_state $\leftarrow x_{t-1}$
      while inversion_step $\leq d_{\max}$ **do**
        $\Delta \leftarrow \min(\text{inversion\_step}, T - t)$
        for $i = 1$ **to** N **do**
          Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
          $\tilde{x}_{t+\Delta} \leftarrow \Psi(x_t, \Delta; \epsilon)$ # Inversion step
          for $k = t + \Delta$ **to** t **do**
            $\tilde{x}_{k-1} \leftarrow \Phi^k(\tilde{x}_k, c)$
          end for
          Estimate $\hat{x}_0^{t-1} \leftarrow \hat{x}_0^{t-1}(\tilde{x}_t, c)$
          $r_{\text{cand}} \leftarrow R(c, \hat{x}_0^{t-1})$
          if $r_{\text{cand}} >$ best_score **then**
            best_score $\leftarrow r_{\text{cand}}$, best_state $\leftarrow \tilde{x}_{t-1}$
          end if
        end for
        if best_score $\geq r_{\text{prev}} + \delta$ **then**
          break from search
        end if
        inversion_step $\leftarrow$ inversion_step $+ 1$
      end while
      $x_{t-1} \leftarrow$ best_state, $r_{\text{prev}} \leftarrow$ best_score
    end if
  end if
  $x_t \leftarrow x_{t-1}$
end for
return x_0

sufficient when the plateau is broad. Z-Sampling [1] performs inversion biased by an unconditional trajectory and then re-applies conditional denoising, repeatedly re-injecting the prompt signal. SOP [41] extends Resampling by scoring multiple candidates, but uses a fixed exploration depth, which limits its ability to escape wider plateaus. In contrast, `Ctrl-Z` Sampling adaptively adjusts the strength of latent perturbation based on reward feedback. When shallow exploration fails, `Ctrl-Z` Sampling incrementally expands its search radius, enabling escape from both narrow and broad local maxima while preserving output quality.

4.2 Controlled Random Zigzag Sampling

To enable conditional diffusion models to escape local optimal, we propose a reward-guided sampling strategy that incorporates controlled backward exploration. Our approach detects potential suboptimal states using a reward model, then dynamically perturbs the latent trajectory (*i.e.*, the sequence of intermediate latent states x_t produced by DDIM sampling) through guided inversion

steps of increasing strength. This process forms a zigzag trajectory in the latent space, alternating between conditional refinement and noise-space exploration.

4.3 Local Maxima Detection

At each selected step t, the clean sample estimate $\hat{x}_0^{t-1}$ is computed from the current latent state x_t using Equation (3), and its quality is evaluated by a reward model $R(c, \hat{x}_0^{t-1})$ that assigns a scalar score to the candidate state. A local maximum is detected when the current reward fails to improve over the most recently accepted score by at least δ:

$$R(c, \hat{x}_0^{t-1}) < r_{\text{prev}} + \delta, \tag{5}$$

where δ denotes the acceptance threshold. By default, $\delta = 0$ enforces non-decreasing reward over time. δ governs the minimal improvement required for a candidate to be accepted. A positive δ suppresses trivial gains and promotes robustness in candidate selection, while $\delta \leq 0$ allows more permissive updates, including those with marginal or slightly lower reward values, which may carry a higher risk of degraded outputs. This criterion provides a simple yet effective signal for identifying local maxima during sampling. However, relying on a fixed offset from the prior best can rigidly penalize high-reward plateaus, especially when δ is large. Designing more flexible criteria such as relative thresholds or global reward schedulers could mitigate such cases and improve adaptability. We leave such extensions to future work.

4.4 Controlled Latent Inversion

To escape from a detected local maximum, we apply the inversion operator $\Psi(x_t, \Delta; \epsilon)$ (defined in Equation (4)) to perturb the current latent x_t toward a higher-noise state $x_{t+\Delta}$. The step size $\Delta \in \mathbb{N}$ corresponds to discrete timestep indices in the DDIM schedule, where smaller values induce mild perturbations and larger values enable broader exploration. This controlled noise injection serves as a test-time search mechanism that reintroduces uncertainty while preserving structural information acquired during prior denoising, thereby enabling the model to explore nearby alternative trajectories effectively.

4.5 Candidate Selection and Adaptive Inversion Step

After applying the inversion, the model resumes forward denoising from $x_{t+\Delta}$ to x_{t-1} using standard conditional DDIM steps $\Phi^k(\cdot, c)$ for $k = t + \Delta, \ldots, t$. The resulting candidate $\hat{x}_0^{t-1}$ is then evaluated by the reward model. At each iteration, N candidates are generated using distinct noise vectors ϵ, and the one with the highest reward is selected. If its score exceeds the most recently accepted reward by at least a threshold δ, the update is accepted.

Otherwise, the inversion step size Δ is incrementally increased (*e.g.*, from 1 to $d_{\max}$), enabling progressively stronger perturbations to search for better

trajectories. This defines a greedy yet adaptive hill-climbing strategy: the search proceeds until a satisfactory candidate is found or the maximum inversion depth is reached. If no improvement beyond the threshold is identified, the best candidate observed during the search is retained, promoting stability while often yielding results comparable to or better than the default forward step.

The inversion depth can be unbounded with $d_{max} = \infty$, and the candidate budget N per step guarantees termination. The candidate evaluation process is inherently parallelizable since each candidate trajectory is conditionally independent until selection, thus can be efficiently distributed across computational resources, supporting scalable inference with minimal overhead.

Empirically, at early timesteps (*e.g.*, $t = 600/1000$), the DDIM update already assigns substantial weight (above 0.02) to the clean estimate $\hat{x}_0^{t-1}$ via Equation (2), meaning x_{t-1} is strongly influenced by $\hat{x}_0^{t-1}$. As a result, much of the image's low-frequency structure is established early, making later corrections less effective. Therefore, `Ctrl-Z` Sampling restricts exploration to the first λ steps from the start of sampling (the high-noise regime, $t \in \{T, \ldots, T-\lambda+1\}$), where trajectory perturbations can still meaningfully influence global structure.

5 Experiments and Results

5.1 Experiment Settings

We evaluate our method on three representative text-to-image benchmarks: **Pick-a-Pic** [30], **DrawBench** [50], and **T2I-CompBench** [25] covering real-world diversity, compositional complexity, and large-scale open-world scenarios. Evaluations are conducted using four metrics: **HPSv2** [61], **PickScore** [30], **ImageReward (IR)** [62], and **Aesthetic Score (AES)** [52]. All experiments are conducted on SD-2.1-base [14] and Hy-DiT [37]. We use $T = 50$ inference steps with CFG (scale 5.5), an exploration window $\lambda = 40$, $N = 4$ candidates, $d_{\max} = 3$, $\delta = 0$, and ImageReward as the reward model. We compare `Ctrl-Z` **Sampling** against standard **DDIM**, **Resampling** [40], **Z Sampling** [1], and **SOP** [41] under settings with comparable number of function evaluations (NFEs), which is reported as the average number of denoiser forward passes per denoising step. Unless specified otherwise, `Ctrl-Z` uses the above default setting; in the main results we denote this default as `Ctrl-Z`‡ and the low-NFE variant as `Ctrl-Z`†, where $\lambda = 30$, $N = 2$, and $d_{\max} = 3$. More details are available in the Supplementary Section B.

5.2 Quantitative Evaluations

The quantitative results in Tables 1 and 2 show consistent improvements over baselines brought by `Ctrl-Z` Sampling. In particular, our method achieves significant gains on ImageReward across both Pick-a-Pic and DrawBench. Compared to SOP, which also relies on ImageReward, `Ctrl-Z` Sampling achieves higher HPSv2 and PickScore, indicating stronger improvements in human-aligned image

Table 1: Quantitative results of sampling methods on Pick-a-Pic and DrawBench, evaluated using four human-aligned metrics. Results are reported for both Stable Diffusion 2.1 and Hunyuan-DiT, along with average number of function evaluations (NFEs). † and ‡ indicate `Ctrl-Z` variants with different exploration parameters, demonstrating performance under varying NFEs.

	Pick-a-Pic↑				DrawBench↑				NFEs ↓
Method	HPS v2	AES	PickScore	IR	HPS v2	AES	PickScore	IR	
Stable Diffusion 2.1									
DDIM	25.34	5.649	20.67	0.194	24.90	5.410	21.39	0.046	1.00
Resampling	26.00	5.648	20.78	0.322	25.49	5.385	21.52	0.205	2.00
Z-Sampling	26.29	5.653	20.71	0.346	25.95	5.465	21.47	0.296	3.00
SOP-1	26.34	5.684	20.85	0.735	25.78	5.426	21.64	0.637	3.00
SOP-4	27.23	5.700	**21.12**	1.113	26.72	5.469	**21.89**	1.008	9.00
`Ctrl-Z`†	26.44	5.686	20.88	0.720	26.15	5.476	21.68	0.650	2.77
`Ctrl-Z`‡	**27.34**	**5.705**	21.02	**1.138**	**26.73**	**5.501**	21.84	**1.025**	7.72
Hunyuan-DiT									
DDIM	30.22	6.324	22.15	1.071	28.90	5.930	22.47	0.827	1.00
Resampling	30.20	6.338	22.14	1.112	28.70	5.912	22.49	0.904	2.00
Z-Sampling	30.10	**6.398**	22.18	1.111	28.60	**5.997**	22.49	0.899	3.00
SOP-1	30.56	6.333	22.19	1.232	28.77	5.902	22.50	1.102	3.00
SOP-4	30.81	6.303	22.22	**1.444**	29.49	5.942	22.58	1.246	9.00
`Ctrl-Z`†	30.69	6.310	22.24	1.214	29.27	5.906	22.51	1.067	2.85
`Ctrl-Z`‡	**30.91**	6.325	**22.30**	1.441	**29.79**	5.908	**22.62**	**1.319**	8.79

Table 2: Quantitative results of different methods on T2I-CompBench.

	Stable Diffusion 2.1 ↑					**Hunyuan-DiT ↑**				
Method	Color	Shape	Texture	Spatial	Numeracy	Color	Shape	Texture	Spatial	Numeracy
DDIM	46.27	41.01	46.06	13.80	46.44	66.46	44.32	60.06	24.18	53.99
Resampling	48.67	40.99	47.46	14.21	46.76	68.86	48.58	60.86	25.29	54.06
Z-Sampling	50.38	42.19	48.84	13.65	47.77	68.80	48.51	60.46	22.89	54.27
SOP-1	53.85	45.00	51.53	18.25	50.81	71.01	49.11	63.33	21.00	54.51
SOP-4	59.64	48.91	60.56	16.97	52.85	73.74	53.74	64.85	21.78	**57.58**
`Ctrl-Z`†	58.65	47.10	57.75	18.55	51.83	71.10	51.59	63.76	21.13	56.35
`Ctrl-Z`‡	**61.26**	**53.97**	**62.24**	**19.29**	**53.73**	**73.77**	**54.99**	**66.45**	**25.43**	56.69

quality. The smaller gain on ImageReward itself is likely due to SOP's always-on search strategy, which repeatedly optimizes the surrogate ImageReward score at each step and thus over-optimizes the reward model without consistently improving overall quality. Table 2 shows that `Ctrl-Z` Sampling performs well on CompBench, whose prompts emphasize object relationships and visual attribute binding. `Ctrl-Z` Sampling adaptively increases exploration strength to revisit and escape early-stage plateaus that lock in coarse suboptimal layout. In contrast, prior methods rely on fixed perturbations and struggle to revise suboptimal

trajectories once coarse structures are formed. We leave some further analysis on the results in Supplementary Section C.1.

Importantly, `Ctrl-Z` provides a simple test-time scaling scheme with a controllable compute–quality trade-off. By tuning exploration depth and width, it can operate in a low-cost regime at roughly 3× NFEs that exceeds other baselines across most metrics, while increasing the budget to about 7–9× NFEs yields further improvements. Compared to SOP at similar budgets, `Ctrl-Z` achieves comparable overall performance with fewer NFEs, and its adaptive depth escalation enables deeper exploration when shallow candidates fail. At comparable NFE budgets, `Ctrl-Z` benefits from adaptive inversion depth, enabling deeper trajectory revisions than fixed-depth exploration and resulting in consistently stronger performance across both small and large compute regimes.

Results on both SD2.1 and Hy-DiT confirm the effectiveness of `Ctrl-Z` Sampling across U-Net and Transformer–based diffusion models, showcasing its compatibility with major frameworks. While our experiments focus primarily on latent diffusion, the hill-climbing intuition naturally extends to pixel-space models such as EDM [28], where denoising similarly seeks regions of less noise, higher quality, and better alignment with the input conditions. We expect `Ctrl-Z` Sampling to remain effective in such settings, leaving direct validation to future work due to computational constraints.

5.3 Ablations

The following sections include ablation experiments on the effectiveness of various design and hyper-parameters used in `Ctrl-Z` Sampling. Additional ablation experiments regarding the max inversion steps, accept threshold, the exploration initiation criteria, and the effect of exploration window can be found at the Supplementary Section C.

Scaling Effect of Exploration Depth and Width. We investigate how the maximum exploration depth $d_{\max}$ and the number of candidates N per inversion step affect generation quality across evaluation metrics (We fix λ as 30 here for efficiency concerns). As shown in Figure 3, increasing $d_{\max}$ improves performance across most metrics, which is consistent with enabling revisions to suboptimal early trajectory decisions, while increasing N raises the chance of identifying a higher-scoring alternative at each triggered step. Overall, depth and width complement each other in improving alignment and perceived quality.

We observe that deeper exploration can reduce reliance on large candidate sets. When $d_{\max}$ is small, increasing N tends to help by providing more local alternatives per step. When $d_{\max}$ is large, further increasing N yields less consistent gains, suggesting diminishing returns from widening the candidate set once the method can already escape via deeper inversions. Under comparable NFEs, deeper-but-narrower settings often outperform wider-but-shallow settings, as indicated by points with similar colors in Figure 3. This supports the importance of adaptive depth escalation for compute-efficient scalable inference.

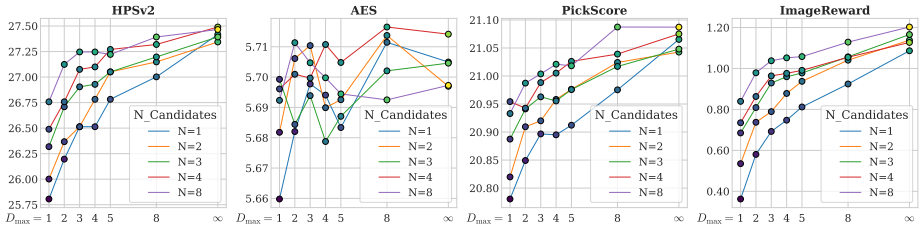

Fig. 3: Effect of exploration depth ($d_{\max}$) and candidate width (N). Each marker aggregates results over the same prompts and seeds for a fixed ($d_{\max}, N$). Lighter colors indicate higher realized average NFEs. `Ctrl-Z` Sampling benefits from both deeper and wider exploration, with increased depth sometimes yielding stronger gains than additional candidates under similar NFEs. Comparisons are best viewed zoomed in.

As exploration budgets increase, improvements are consistent across most metrics. AES is a partial exception, since it measures aesthetic preference that is not directly optimized by ImageReward-based selection, and therefore may correlate less with the exploration objective that measures the conditional alignment and structure correctness. While larger-scale settings are beyond the scope of this study due to resource constraints, the observed trends suggest that `Ctrl-Z` Sampling can scale effectively by trading off depth and width.

Choice of Reward Model. We evaluate the impact of different reward models using all employed metrics and also CLIPScore [20], with results presented in Table 3. Compared to the DDIM baseline reported in Table 1, `Ctrl-Z` Sampling improves most metrics under a wide range of surrogate reward choices, indicating that its effectiveness is not tied to a specific scorer. Crucially, for each surrogate reward, the remaining evaluation metrics are held out from the exploration loop. The predominantly positive changes on these non-optimized measures across both benchmarks indicate that Ctrl-Z Sampling improves generation quality beyond simply maximizing the selected scorer.

In particular, using AES as the reward model increases aesthetic scores, but has limited impact on alignment-sensitive metrics since it does not evaluate condition faithfulness. Among CLIP-derived scorers, PickScore tends to outperform CLIPScore when used for exploration, while overall PickScore and ImageReward yield stronger results than other tested objectives.

We adopt ImageReward in the main experiments due to its broad applicability, efficiency, and reduced benchmark coupling, since PickScore is trained on data closely related to Pick-a-Pic. Notably, the choice of R also affects realized NFEs by changing the trigger and acceptance dynamics, yielding different compute–performance profiles. While [41] suggests that combining multiple reward models can further improve performance, we leave this to future work.

When to Explore. This section evaluates different strategies for triggering zigzag exploration. We compare the proposed reward-based initiation mecha-

Table 3: Ablation results on exploration guided by different reward models and exploration initiation criteria. PS and IR denote PickScore and ImageReward, respectively. Best results are **bolded**; second-best are underlined.

	Pick-a-Pic↑				DrawBench↑				NFEs ↓
Method	HPS v2	AES	PS	IR	HPS v2	AES	PS	IR	
Reward Model for Controlled Exploration									
CLIPScore	26.59	5.614	20.97	0.573	26.29	5.333	21.70	0.483	11.75
HPS v2	27.39	5.723	21.05	0.503	**27.14**	5.503	21.80	0.478	3.44
AES	26.29	**6.147**	20.87	0.306	25.54	**5.902**	21.62	0.231	11.17
PickScore	**27.51**	5.763	**21.69**	0.589	27.12	5.497	**22.38**	0.573	7.01
ImageReward	27.34	5.705	21.02	**1.138**	26.73	5.501	21.84	**1.025**	7.72
Exploration Initiation Criteria									
Always ($p = 1.0$)	**27.86**	**5.737**	**21.10**	**1.317**	**27.17**	5.482	**21.91**	**1.174**	16.72
Random ($p = 0.5$)	27.00	5.713	21.03	1.062	26.71	5.469	21.83	0.906	7.81
Reward-Based	27.34	5.705	21.02	1.138	26.73	5.501	21.84	1.025	7.72

nism, which activates exploration upon detecting a reward plateau, against two baselines: *Always*, which performs exploration at every denoising step, and *Random*, which triggers exploration at each step with probability $p = 0.5$.

As shown in Table 3, the Always strategy yields the highest generation quality but incurs substantial computational overhead, averaging 16.72× NFEs. In contrast, the reward-based strategy attains competitive performance at much lower cost (7.72× NFEs), offering a stronger quality–compute trade-off. The Random strategy provides modest quality gains with a similar computational budget (7.81× NFEs), though the probability can be flexibly tuned to interpolate performance between never and always exploring. Overall, the proposed strategy delivers improved sampling quality relative to cost and remains more efficient than the Always strategy. Notably, all three strategies are effective within the `Ctrl-Z` Sampling framework and can be selected based on the desired trade-off between performance and inference cost.

5.4 Sample Diversity Analysis

Reward-guided inference-time scaling can potentially improve precision by concentrating probability mass on a smaller set of high-scoring outputs, at the cost of reduced sample diversity. We therefore evaluate distributional quality and diversity on a COCO subset with 2k images using Stable Diffusion 2.1. Table 4 reports FID, precision, recall, and LPIPS under the default DDIM sampler and two Ctrl-Z configurations with increasing inference budgets.

As the inference budget increases, Ctrl-Z improves both distributional quality and precision. Importantly, these gains are not accompanied by systematic diversity degradation: LPIPS remains slightly above the DDIM baseline, while recall stays close to the baseline across both Ctrl-Z configurations. This suggests that the precision improvements primarily arise from better visual quality

Table 4: Sampling diversity analysis on COCO-2k with Stable Diffusion 2.1. Ctrl-Z improves distributional quality and precision while preserving diversity close to the DDIM baseline.

Method	FID↓	Precision↑	Recall↑	LPIPS↑	NFEs↓
DDIM	40.22	71.20	55.85	75.00	1.00
`Ctrl-Z`†	40.10	73.95	54.80	75.71	2.71
`Ctrl-Z`‡	39.98	75.10	55.20	75.54	7.54

and prompt alignment, rather than from a monotonic precision–recall trade-off caused by mode concentration.

This behavior is consistent with the design of `Ctrl-Z`. Unlike methods that draw a large pool of independently completed samples and apply final best-of-(N) reranking, `Ctrl-Z` performs bounded trajectory-level selection only when the reward trajectory reaches a plateau. Its exploration therefore revises locally suboptimal denoising paths rather than globally filtering outputs toward a narrow set of reward-preferred modes. Within the evaluated compute range, the improvements in precision primarily reflect better visual quality and prompt alignment, while preserving intra-prompt diversity comparable to standard DDIM sampling.

5.5 Qualitative Evaluations

Figure 4 presents qualitative comparisons between baseline methods and the proposed `Ctrl-Z` Sampling. `Ctrl-Z` Sampling produces outputs that are both visually coherent and semantically aligned with input prompts, as reflected in examples involving spatial relations, numeracy, color, and action understanding. In contrast, baseline methods often produce results that are either semantically misaligned with the prompt or lack visual coherence. While SOP can improve image quality by incurring a larger candidate pool, we observe failures in challenging cases. The fixed exploration strength makes SOP vulnerable to low-frequency errors in the latent space, leading to local reward gains from object presence. Consequently, the trajectory often converges to visually plausible but semantically imprecise states (*e.g.*, "cat and vase" or "blue and bear"). In contrast, `Ctrl-Z` Sampling adaptively increases exploration strength, allowing the model to escape such traps and generate more aligned content with distinct low-frequency components. These findings underscore the effectiveness of `Ctrl-Z` Sampling in escaping suboptimal generations through adaptive exploration. Further qualitative examples and analysis can be found at Supp. Section F.

6 Conclusion

We present `Ctrl-Z` Sampling, an inference-time scaling sampling strategy for diffusion models that improves text-to-image generation by escaping plateaus and

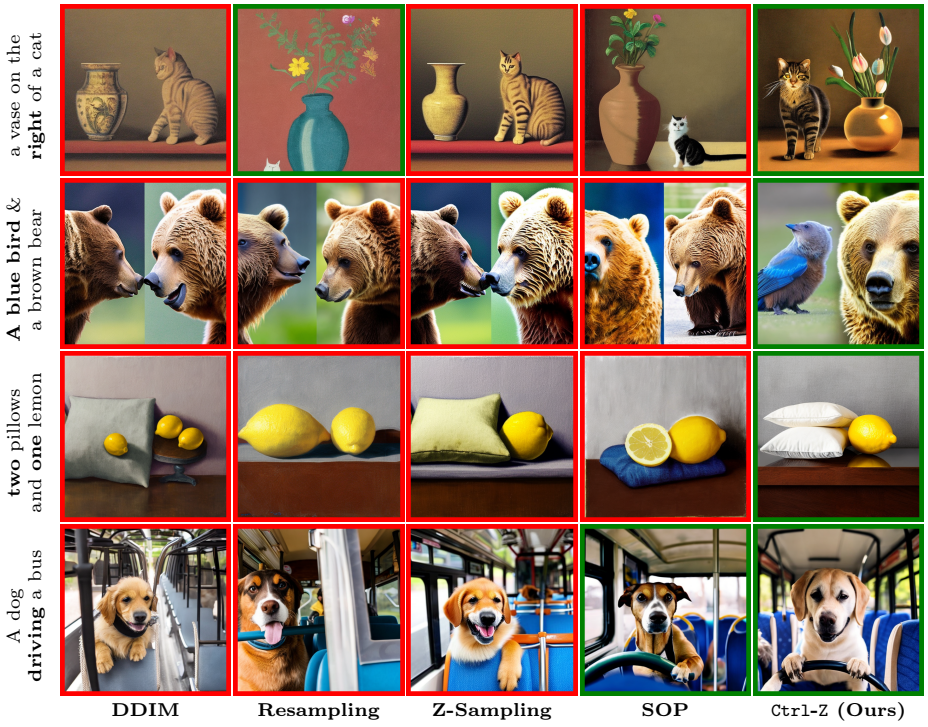

Fig. 4: Qualitative comparison of different sampling methods with SD-2.1-base. Generated images that align with the input condition (shown on the left) are highlighted with *green bounding boxes*, while erroneous or suboptimal images are marked in *red*.

local optima in the surrogate quality landscape. `Ctrl-Z` Sampling detects stagnation and performs controlled inversion with adaptively increased depth using a surrogate quality score. Experiments on text-to-image benchmarks show consistent gains in conditioning alignment and visual fidelity in different NFE regimes, offering a practical compute–quality trade-off without modifying the base denoiser. Future work includes validating `Ctrl-Z` Sampling in broader, higher-budget inference-time scaling settings and developing reward-model-agnostic schedules that jointly adapt exploration initiation and strength.

References

1. Bai, L., Shao, S., Zhou, Z., Qi, Z., Xu, Z., Xiong, H., Xie, Z.: Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
2. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 843–852 (2023)

3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Chamberlain, B., Rowbottom, J., Gorinova, M.I., Bronstein, M., Webb, S., Rossi, E.: Grand: Graph neural diffusion. In: International Conference on Machine Learning (ICML). pp. 1407–1418 (2021)
5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM SIGGRAPH **42**(4), 1–10 (2023)
6. Chen, C., Ding, H., Sisman, B., Xu, Y., Xie, O., Yao, B.Z., Tran, S.D., Zeng, B.: Diffusion models for multi-task generative modeling. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
7. Choi, Y., Park, J., Kim, H., Lee, J., Park, S.: Fair sampling in diffusion models through switching mechanism. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2024)
8. Chung, H., Kim, J., Park, G.Y., Nam, H., Ye, J.C.: CFG++: Manifold-constrained classifier free guidance for diffusion models. arXiv preprint arXiv:2406.08070 (2024)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 8780–8794 (2021)
10. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**, 1602 – 1614 (2011)
11. Everaert, M.N., Fitsios, A., Bocchio, M., Arpa, S., Süsstrunk, S., Achanta, R.: Exploiting the signal-leak bias in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4025–4034 (2024)
12. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 79858–79885 (2023)
13. Fang, G., Ma, X., Wang, X.: Structural pruning for diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 16716–16728 (2023)
14. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
15. Ghosh, A., Lyu, H., Zhang, X., Wang, R.: Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
16. Guo, W., Wang, H., Ma, J., Cai, W.: Gotta hear them all: Towards sound source aware audio generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 40, pp. 21495–21503 (2026)
17. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9380–9389 (2024)
18. Guttenberg, N.: Diffusion with offset noise. CrossLabs Research Blogs (January 2023)

19. He, H., Liang, J., Wang, X., Wan, P., Zhang, D., Gai, K., Pan, L.: Scaling image and video generation via test-time evolutionary search. arXiv preprint arXiv:2505.17618 (2025)
20. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 6840–6851 (2020)
22. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
23. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 8633–8646 (2022)
24. Hong, S.: Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 66743–66772 (2024)
25. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 78723–78747 (2023)
26. Jain, V., Sareen, K., Pedramfar, M., Ravanbakhsh, S.: Diffusion tree sampling: Scalable inference-time alignment of diffusion models. arXiv preprint arXiv:2506.20701 (2025)
27. Janner, M., Du, Y., Tenenbaum, J.B., Levine, S.: Planning with diffusion for flexible behavior synthesis. In: International Conference on Machine Learning (ICML) (2022)
28. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
29. Kim, S., Kim, M., Park, D.: Test-time alignment of diffusion models without reward over-optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2025)
30. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 36652–36663 (2023)
31. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
32. Li, S., Hu, T., van de Weijer, J., Khan, F.S., Liu, T., Li, L., Yang, S., Wang, Y., Cheng, M.M., Yang, J.: Faster diffusion: Rethinking the role of the encoder for diffusion model inference. Advances in Neural Information Processing Systems (NeurIPS) **37**, 85203–85240 (2024)
33. Li, T., Sun, Q., Fan, L., He, K.: Fractal generative models. arXiv preprint arXiv:2502.17437 (2025)
34. Li, X., Thickstun, J., Gulrajani, I., Liang, P.S., Hashimoto, T.B.: Diffusion-lm improves controllable text generation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 4328–4343 (2022)
35. Li, X., Uehara, M., Su, X., Scalia, G., Biancalani, T., Regev, A., Levine, S., Ji, S.: Dynamic search for inference-time alignment in diffusion models. arXiv preprint arXiv:2503.02039 (2025)

36. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Divide & bind your attention for improved generative semantic nursing. In: British Machine Vision Conference (BMVC) (2023)
37. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., Chen, D., He, J., Li, J., Li, W., Zhang, C., Quan, R., Lu, J., Huang, J., Yuan, X., Zheng, X., Li, Y., Zhang, J., Zhang, C., Chen, M., Liu, J., Fang, Z., Wang, W., Xue, J., Tao, Y., Zhu, J., Liu, K., Lin, S., Sun, Y., Li, Y., Wang, D., Chen, M., Hu, Z., Xiao, X., Chen, Y., Liu, Y., Liu, W., Wang, D., Yang, Y., Jiang, J., Lu, Q.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024)
38. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5404–5411 (2024)
39. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 5775–5787 (2022)
40. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11461–11471 (2022)
41. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., Xie, S.: Scaling inference time compute for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2523–2534 (June 2025)
42. Mao, J., Wang, X., Aizawa, K.: Guided image synthesis via initial image editing in diffusion model. In: Proceedings of the 31st ACM International Conference on Multimedia (MM). pp. 5321–5329 (2023)
43. Oshima, Y., Suzuki, M., Matsuo, Y., Furuta, H.: Inference-time text-to-video alignment with diffusion latent beam search. arXiv preprint arXiv:2501.19252 (2025)
44. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4195–4205 (2023)
45. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
46. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
47. Ramesh, V., Mardani, M.: Test-time scaling of diffusion models via noise trajectory search. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
49. Sabour, A., Fidler, S., Kreis, K.: Align your steps: optimizing sampling schedules in diffusion models. In: International Conference on Machine Learning (ICML) (2024)
50. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 36479–36494 (2022)

51. Samuel, D., Ben-Ari, R., Raviv, S., Darshan, N., Chechik, G.: Generating images of rare concepts using pre-trained diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2024)
52. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 25278–25294 (2022)
53. Seo, H., Jeong, W., sun Seo, J., Chun, S.Y.: Skrr: Skip and re-use text encoder layers for memory efficient text-to-image generation. In: International Conference on Machine Learning (ICML) (2025)
54. Singhal, R., Horvitz, Z., Teehan, R., Ren, M., Yu, Z., McKeown, K., Ranganath, R.: A general framework for inference-time scaling and steering of diffusion models. In: International Conference on Machine Learning (ICML) (2025)
55. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML). pp. 2256–2265. pmlr (2015)
56. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
57. Tang, Z., Peng, J., Tang, J., Hong, M., Wang, F., Chang, T.H.: Inference-time alignment of diffusion models with direct noise optimization. arXiv preprint arXiv:2405.18881 (2024)
58. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8228–8238 (2024)
59. Wang, Z., Xia, X., Chen, R., Yu, D., Wang, C., Gong, M., Liu, T.: Lavin-dit: Large vision diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 20060–20070 (2025)
60. Watson, D., Chan, W., Ho, J., Norouzi, M.: Learning fast samplers for diffusion models by differentiating through sample quality. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022)
61. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
62. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 15903–15935 (2023)
63. Xu, K., Zhang, L., Shi, J.: Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. arXiv preprint arXiv:2405.14828 (2024)
64. Xu, M., Yu, L., Song, Y., Shi, C., Ermon, S., Tang, J.: Geodiff: A geometric diffusion model for molecular conformation generation. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022)
65. Xu, Y., Deng, M., Cheng, X., Tian, Y., Liu, Z., Jaakkola, T.S.: Restart sampling for improving generative processes. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
66. Yin, X., Xu, C., Zhou, H., Wei, B., Cai, W.: Accelaes: Accelerating diffusion transformers for training-free aesthetic-enhanced image generation. arXiv preprint arXiv:2603.12575 (2026)

67. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (2023)
68. Zhang, X., Lin, H., Ye, H., Zou, J., Ma, J., Liang, Y., Du, Y.: Inference-time scaling of diffusion models through classical search. arXiv preprint arXiv:2505.23614 (2025)
69. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 49842–49869 (2023)
70. Zhou, Z., Shao, S., Bai, L., Xu, Z., Han, B., Xie, Z.: Golden noise for diffusion models: A learning framework. arXiv preprint arXiv:2411.09502 (2024)