Training-Free Adaptive Diffusion with Bounded Difference Approximation Strategy

Hancheng Ye[1,*], **Jiakang Yuan**[2,*], **Renqiu Xia**[3], **Xiangchao Yan**[1], **Tao Chen**[2], **Junchi Yan**[3], **Botian Shi**[1], **Bo Zhang**[1,‡]

[1]Shanghai Artificial Intelligence Laboratory
[2]School of Information Science and Technology, Fudan University
[3]School of Artificial Intelligence, Shanghai Jiao Tong University

yehancheng@pjlab.org.cn, jkyuan22@m.fudan.edu.cn, zhangbo@pjlab.org.cn

Abstract

Diffusion models have recently achieved great success in the synthesis of high-quality images and videos. However, the existing denoising techniques in diffusion models are commonly based on step-by-step noise predictions, which suffers from high computation cost, resulting in a prohibitive latency for interactive applications. In this paper, we propose *AdaptiveDiffusion* to relieve this bottleneck by adaptively reducing the noise prediction steps during the denoising process. Our method considers the potential of skipping as many noise prediction steps as possible while keeping the final denoised results identical to the original full-step ones. Specifically, the skipping strategy is guided by the *third*-order latent difference that indicates the stability between timesteps during the denoising process, which benefits the reusing of previous noise prediction results. Extensive experiments on image and video diffusion models demonstrate that our method can significantly speed up the denoising process while generating identical results to the original process, achieving up to an average $2 \sim 5\times$ speedup without quality degradation. The code is available at https://github.com/UniModal4Reasoning/AdaptiveDiffusion.

1 Introduction

Recently, Diffusion models [1, 11, 30, 33] have emerged as a powerful tool for synthesizing high-quality images and videos. Their capability to generate realistic and detailed visual content has made them a popular choice in various applications, ranging from artistic creation to data augmentation, *e.g.*, Midjourney, Sora [4], etc. However, the conventional denoising techniques employed in these models involve step-by-step noise predictions, which are computationally intensive and lead to significant latency, *e.g.*, taking tens seconds for SDXL [32] to generate a high-quality image of 1024x1024 resolutions. Diffusion acceleration as an effective technique has been deeply explored recently, which mainly focuses on three paradigms: (1) reducing sampling steps [39, 13, 23, 35, 24], (2) optimizing model architecture [15, 42, 9, 27] and (3) parallelizing inference [16, 38].

Currently, most strategies are designed based on a fixed acceleration mode for all prompt data. However, in our experiments, it is observed that different prompts may require different steps of noise prediction to achieve the same content as the original denoising process, as presented in Fig. 1. Here, we compare the denoising paths using two different prompts for SDXL [32], both of which preserve rich content against the original full-step generation results. The *denoising path* denotes a bool-type sequence, where each element represents whether to infer the noise from the noise prediction model. It can be observed that Prompt 2 needs more steps to generate an almost lossless image than Prompt

*Equal contribution, ‡Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

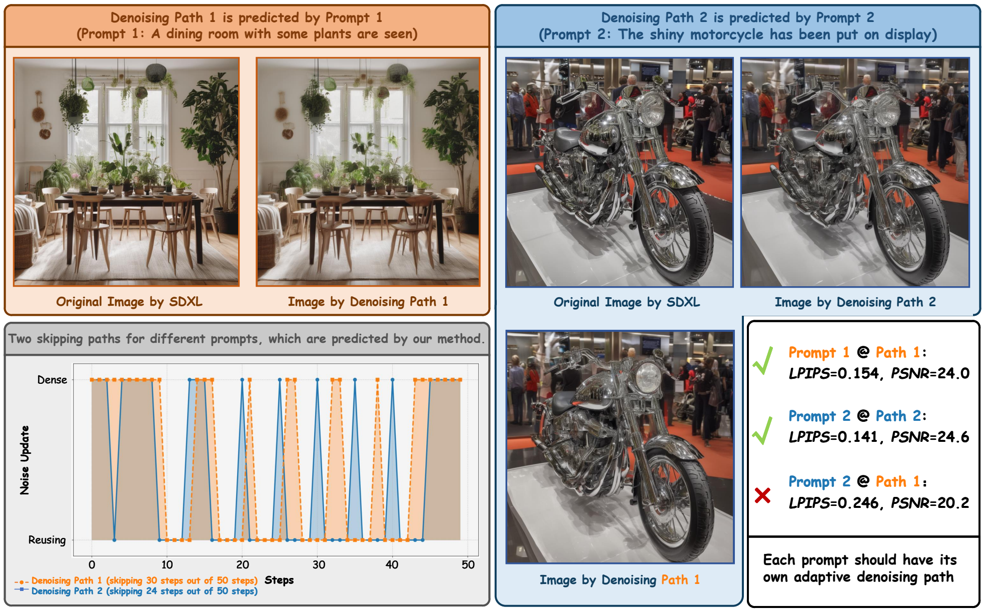


Figure 1: Different prompts may have different denoising paths to generate the high-quality image. For Prompt 1, we only need 20 steps out of 50 steps for noise predictions to generate an almost lossless image, while for Prompt 2, we need 26 steps out of 50 steps to achieve an almost lossless image.

1 (A lower LPIPS [51] value means more similarity between two images generated by the original strategy and our strategy). Therefore, it is necessary to explore a **prompt-adaptive** acceleration paradigm to consider the denoising diversity between different prompts.

Motivated by this observation, in this paper, we deeply dive into the skipping scheme for the noise prediction model and propose AdaptiveDiffusion, a novel approach that adaptively accelerates the generation process according to different input prompts. The fundamental concept behind AdaptiveDiffusion is to adaptively reduce the number of noise prediction steps according to different input prompts during the denoising process, and meanwhile maintain the quality of the final output. The key insight driving our method is that **the redundancy of noise prediction is highly related to the *third*-order differential distribution between temporally-neighboring latents**. This relation can be leveraged to design an effective skipping strategy, allowing us to decide when to reuse previous noise prediction results and when to proceed with new calculations. Our approach utilizes the *third*-order latent difference to assess the redundancy of noise prediction at each timestep, reflecting our strategy's dependency on input information, thus achieving a prompt-adaptive acceleration paradigm.

Extensive experiments conducted on both image and video diffusion models demonstrate the effectiveness of AdaptiveDiffusion. The results show that our method can achieve up to a 5.6x speedup in the denoising process with better preservation quality. This improvement in acceleration quality opens up new possibilities for the application of diffusion models in real-time and interactive environments.

In summary, AdaptiveDiffusion represents a substantial advancement in adaptively efficient diffusion, offering a practical solution to the challenge of high computational costs associated with sequentially denoising techniques. The main contribution is threefold: (1) To our best knowledge, our method is the first to explore the adaptive diffusion acceleration from the step number reduction of noise predictions that makes different skipping paths for different prompts. (2) We propose a novel approach, namely AdaptiveDiffusion, which develops a plug-and-play criterion to decide whether the noise prediction should be inferred or reused from the previous noise results. (3) Extensive experiments conducted on various diffusion models [32, 33, 52, 44] and tasks [7, 21, 45, 8] demonstrate the superiority of our AdaptiveDiffusion to the existing acceleration methods in the trade-off among efficiency, performance and generalization ability.

2 Related Works

2.1 Diffusion Models

Diffusion models [1, 11, 30, 33, 10] have achieved great success and served as a milestone in content generation. As a pioneer, Denoising Diffusion Probabilistic Models (DDPMs) [11] generate higher-

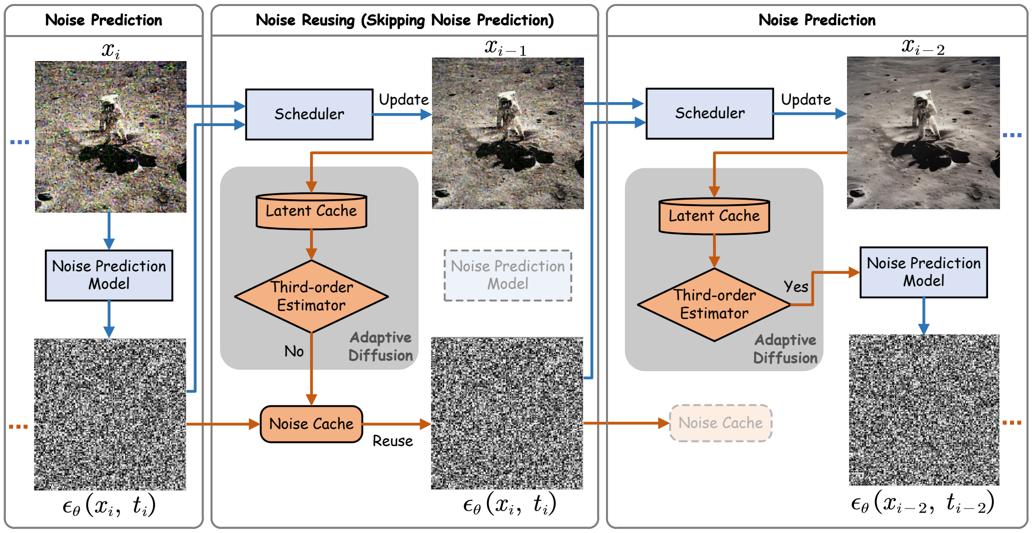


Figure 2: Denoising process of the proposed AdaptiveDiffusion: We design a third-order estimator (Refer to Sec. 3.3 for details), which can find the redundancy between neighboring timesteps, and thus, the noise prediction model can be skipped or inferred according to the indicate from the estimator, achieving the adaptive diffusion process. Note that the timestep and text information embeddings are not shown for the sake of brevity.

quality images compared to generative adversarial networks (GANs) [17, 6] through an iterative denoising process. To improve the efficiency of DDPM, Latent Diffusion Models (LDMs) [33] perform forward and reverse processes in a latent space of lower dimensionality which further evolves into Stable Diffusion (SD) family [36, 32]. Recently, video diffusion models [3, 52, 2, 44] have attracted increasing attention, especially after witnessing the success of Sora [4]. Stable Video Diffusion (SVD) [3] introduces a three-stage training pipeline and obtains a video generation model with strong motion representation. I2VGen-XL [52] first obtains a model with multi-level feature extraction ability, then enhances the resolution and injects temporal information in the second stage. Despite the high quality achieved by diffusion models, the inherent nature of the reverse process which needs high computational cost slows down the inference process.

2.2 Accelerating Diffusion Models

Current works accelerate diffusion models can be divided into the following aspects. *(1) Reducing Sampling Steps* [39, 23, 24, 49, 40, 25, 20, 37, 28, 29, 26]. DDIM [39] optimizes sampling steps by exploring a non-Markovian process. Further studies [23, 24, 49] such as DPM-Solver [23] propose different solvers for diffusion SDEs and ODEs to reduce sampling steps. Another way to optimize sampling steps is to train few-step diffusion models by distillation [40, 25, 20]. Among them, consistency models [40, 25] directly map noise to data to enable one-step generation. Other works [20, 37] explore progressive distillation or adversarial distillation to effectively reduce the reverse steps. Besides, some works [26, 42, 29] introduce early stop mechanism into diffusion models. *(2) Optimizing Model Architecture* [9, 19, 46, 27, 5, 47, 22, 41]. Another strategy to accelerate diffusion models is to optimize model efficiency to reduce the cost during inference. Diff-pruning [9] compresses diffusion models by employing Taylor expansion over pruned timesteps. DeepCache [27] notices the feature redundancy in the denoising process and introduces a cache mechanism to reuse pre-computed features. *(3) Parallel Inference* [16, 38, 12]. The third line lies in sampling or calculating in a parallel way. ParaDiGMS [38] proposes to use Picard iteration to run multiple steps in parallel. DistriFusion [16] introduces displaced patch parallelism by reusing the pre-computed feature maps. Compared with the existing paradigms designing a fixed acceleration mode for all input prompts, our method highlights the adaptive acceleration manner with a plug-and-play criterion based on the high-order latent differential distribution, which allows various diffusion models with different sampling schedulers to achieve significant speedup with a negligible performance drop and deployment cost.

3 The Proposed Approach

3.1 Preliminary

Reverse Denoising Process. Diffusion models [11, 39] are designed to learn two processes with noise addition (known as the forward process) and noise reduction (known as the reverse process).

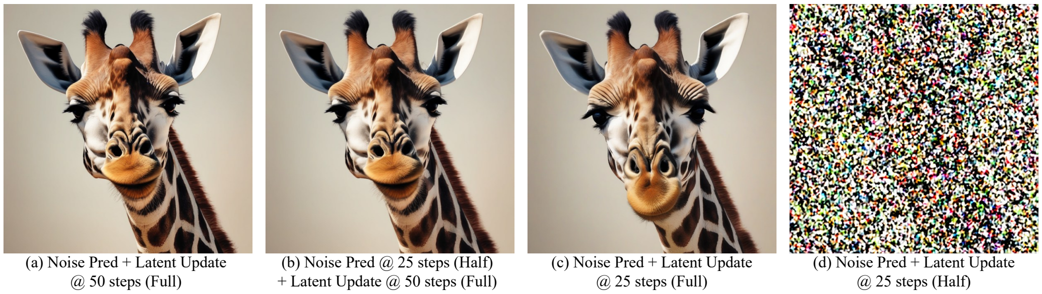

(a) Noise Pred + Latent Update @ 50 steps (Full)
(b) Noise Pred @ 25 steps (Half) + Latent Update @ 50 steps (Full)
(c) Noise Pred + Latent Update @ 25 steps (Full)
(d) Noise Pred + Latent Update @ 25 steps (Half)

Figure 3: Different update strategies. (a) The default SDXL [32] samples 50 steps of noise prediction followed by the latent update process. (b) Our AdaptiveDiffusion skips 25 steps of noise prediction according to the *third*-order estimator, while the latent is fully updated at all 50 steps. (c) SDXL samples 25 steps of the noise prediction and latent update process. (d) The default SDXL skips 25 steps of both noise prediction and latent update from its sampled 50 steps.

During the inference stage, only the reverse denoising process is adopted that starts from the Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and iteratively denoises the input sample under the injected condition to get the final clean image(s) x_0, where T is the predefined number of denoising steps. Specifically, given an intermediate noisy image x_i at timestep i ($i = 1, ..., T$), the noise prediction model ϵ_θ (*e.g.*, UNet [34]) takes x_i, timestep t_i and an additional condition c (*e.g.*, text, image, and motion embeddings, etc) as input to approximate the noise distribution in x_i. The update from x_i to x_{i-1} is determined by different samplers (*a.k.a*, schedulers) that can be generally formulated as Eq. (1):

$$x_{i-1} = f(i-1) \cdot x_i - g(i-1) \cdot \epsilon_\theta(x_i, t_i), \quad i = 1, \ldots, T, \tag{1}$$

where $f(i)$ and $g(i)$ are step-related coefficients derived by specific samplers [23, 50]. The computation in the update process mainly involves a few element-wise additions and multiplications. Therefore, the main computation cost in the denoising process stems from the inference of noise prediction model ϵ_θ [16].

Step Skipping Strategy. As proved by previous works [27, 42, 15], features between consecutive timesteps present certain similarities in distribution, thus there exists a set of redundant computations that can be skipped. Previous works usually skip either the whole update process or the partial computation within the noise prediction model at redundant timesteps. However, as visualized in Fig. 3, the latent update process in those redundant timesteps may be important to the lossless image generation. Besides, the calculation redundancy of the noise prediction model within each denoising process is still under-explored. To reduce the computation cost from the noise prediction model, we consider directly reducing the number of noise prediction steps from the original denoising process, which will be proved more effective and efficient to accelerate the generation with almost no quality degradation in Sec. 3.3. Given a certain timestep i to skip, our skipping strategy for the update process can be formulated using Eq. (2):

$$\begin{aligned} x_i &= f(i) \cdot x_{i+1} - g(i) \cdot \epsilon_\theta(x_{i+1}, t_{i+1}), \\ x_{i-1} &= f(i-1) \cdot x_i - g(i-1) \cdot \epsilon_\theta(x_{i+1}, t_{i+1}). \end{aligned} \tag{2}$$

3.2 Error Estimation of the Step Skipping Strategy

To validate the effectiveness of our step-skipping strategy, we theoretically analyze the upper bound of the error between the original output and skipped output images. For simplicity, we consider skipping one step of noise prediction, *e.g.*, the i-th timestep. Specifically, we have the following theorem about the one-step skipping.

Theorem 1. *Given the skipping timestep i ($i > 0$), the original output x_{i-1}^{ori} and the skipped output x_{i-1}, then the following in-equation holds:*

$$\varepsilon_{i-1} = \|x_{i-1} - x_{i-1}^{ori}\| = \mathcal{O}(t_i - t_{i+1}) + \mathcal{O}(x_i - x_{i+1}). \tag{3}$$

The proof can be found in Appendix A.2.1. From Eq. (3), it can be observed that the error of the noise-skipped output is upper-bounded by the first-order difference between the previous two outputs. Similarly, the error of the continuously skipped output is upper-bounded by the accumulative differences between multiple previous outputs, which can be found in Appendix A.2.2 and A.2.3.

Therefore, it can be inferred that as the difference between the previous outputs x is continuously minor, it is possible to predict that the noise prediction at the next timestep can be skipped without damaging the output. This inspires us to utilize the distribution of previous outputs to indicate the skip potential of the next-step noise prediction, as detailed in the following section.

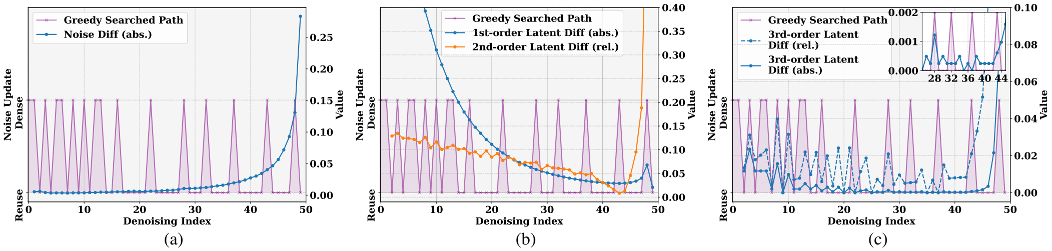


Figure 4: The relation between order differential distributions and the searched optimal skipping path for one prompt. (a) The 1st-order noise differential distribution of the original full-step generation shows no relation with the optimal skipping path. (b) The 1st latent differential distribution indicates the distribution of the optimal skipping path but with no explicit mapping with skipping decisions, while the relative 2nd-order latent differential distribution shows a certain skipping signal in its fluctuation, but this signal is buried in the unstable magnitude. (c) The relative 3rd-order latent differential distribution shows a clearer signal for skipping decisions.

3.3 Third-order Estimation Criterion

Observations. In this section, we take SDXL [32] with Euler sampling scheduler as an example to describe the effectiveness of the proposed third-order estimator. Before deriving the third-order estimation criterion, one thing is to calculate the optimal skipping path from the given timestep number, so that we can evaluate the effectiveness of our proposed estimator. Considering the explosive searching cost within a large search space (*e.g.*, searching the optimal path of skipping N steps within T steps for one prompt requires C_T^N search time cost), we design a greedy search algorithm to approximate the optimal skipping path under different skipping targets, which can be found in Alg. (1).

Here we randomly take the prompt "*A bustling 18th-century market scene with vendors, shoppers, and cobblestone streets, all depicted in the detailed oil painting style.*" as an example to visualize the optimal skipping path searched by the greedy search algorithm under the predefined skipping target constraint. Meanwhile, as the skipping error is upper-bounded by the constraint of the first-order latent difference, it inspires us to explore the relationship between the first-order difference distribution and the ideal skipping path.

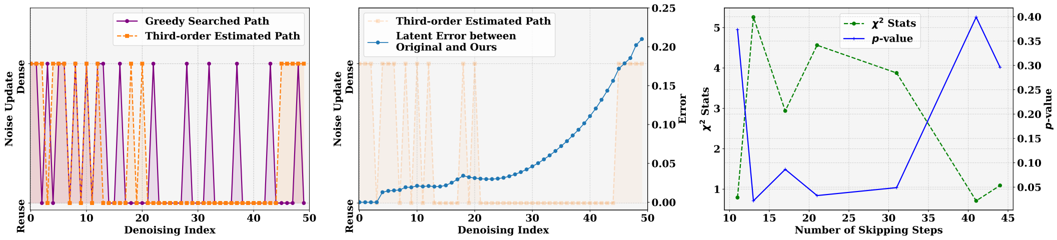
As shown in Fig. 4, we visualize two types of first-order difference (one in Fig. 4a representing the noise difference distribution, another in Fig. 4b representing latent difference distribution) in the original full-step diffusion to compare with the skipping path. It can be observed that both two distributions of first-order differences are smooth across the denoising process, showing little relationship with the optimal skipping path. A similar insight can be observed in the distribution of the second-order latent difference, as shown in Fig. 4b.

However, when considering the third-order latent difference distribution, it presents a significant fluctuation in the original full-step denoising process. As shown in Fig. 4c, the distribution of the skipping path is related to the distribution of the third-order latent difference, especially in the early denoising process (around 15 timesteps), where the noise densely updates when the third-order latent difference increases and can be skipped when the third-order difference decreases. As for the later denoising process, when most third-order differences are relatively minor, most noise prediction steps are also skipped. According to the first-order latent difference presented in Fig. 4b, the differences between consecutive latent in the early denoising process are significantly larger than those in the later denoising process. Therefore, the precise importance estimation of noise predictions in the early process is much more important and the third-order difference distribution can intuitively serve as the indicator of the noise prediction strategy. The theoretical relation between the third-order derivative of latents and the optimal skipping scheme is analyzed in Appendix A.2.4.

Criterion. Based on the above empirical observation of the relationship between the optimal skipping path and the high-order difference distributions, the third-order estimator is proposed to indicate the potential of skipping the noise computation. Specifically, the criterion is formulated as follows:

$$\xi\left(x_{i-1}\right) = \left\|\Delta^{(3)} x_{i-1}\right\| \geq \delta \|\Delta x_i\|, \tag{4}$$

where $\xi\left(x_{i-1}\right)$ is the indicator that takes x_{i-1} and previous latents x_i, x_{i+1}, x_{i+2} as input to estimate whether the next noise prediction can be skipped. If $\xi(x_{i-1})$ returns False, then the noise from the previous step will be reused to update x_{i-1}. $\Delta^{(3)} x_{i-1}$ denotes the third-order latent difference at timestep $i-1$, *i.e.*, $\Delta^{(3)} x_{i-1} = \Delta^{(2)} x_{i-1} - \Delta^{(2)} x_i = \Delta x_{i-1} - 2\Delta x_i + \Delta x_{i+1}$, and Δx_i is defined as the difference between x_i and x_{i+1} ($i = 0, ..., T-1$). δ is a hyperparameter thresholding the



(a) (b) (c)

Figure 5: The effectiveness of the proposed *third*-order estimator. (a) The third-order estimated skipping path shares a similar distribution with the optimal skipping path. (b) The latent error between the full-step update path and the estimated skipping path. (c) The χ^2 stats and p-value between the greedy searched paths and the third-order estimated paths at different skipping targets.

relative scale of $\Delta^{(3)}x_{i-1}$. The reason for selecting Δx_i is that $\Delta^{(3)}x_{i-1}$ actually describes the distance between $(\Delta x_{i-1} + \Delta x_{i+1})/2$ and Δx_i. Therefore, it is natural to utilize the relative distance against Δx_i to indicate the stability of the denoising process. Fig. 4c present a strong relation between $\|\Delta^{(3)}x_{i-1}/\Delta x_i\|$ (the blue dashed line) and the optimal skipping path.

Effectiveness of the Third-order Estimator. To validate the effectiveness of the proposed third-order estimator, we compare our third-order estimated path with the optimal skipping path searched by the greedy algorithm, which is shown in Fig. 5a. It can be observed that the distribution of our estimated path is largely similar to the optimal skipping path. The reason for continuous skipping in the later denoising process is that the third-order difference keeps approaching zero as illustrated in Fig. 4c. The accumulative error caused by skipping noise predictions is described in Fig. 5b, where it is observed that the error starts increasing quickly after continuously skipping the noise predictions. Thus, it is vital to introduce another hyperparameter, *i.e.*, the maximum step number of continuous skipping C_{max}, to control the accumulative error. Hyperparameter analyses are described in Sec. 4.3.

Furthermore, we analyze the statistical correlation between the estimated path and the optimal path to test whether the designed criterion is significantly correlated to the optimal skipping criterion. As shown in Fig. 5c, we compute the χ^2 stats and p-values under different step numbers of skipping. The results indicate that when the skipping steps are moderate, the estimated skipping path and the optimal skipping path are significantly correlated. For those targeting at small and large numbers of skipping steps, the correlation is statistically insignificant. The test details can be found in Appendix A.3. The overall skipping algorithm is shown in Alg. (2) of Appendix A.2.5.

4 Experiments

4.1 Experimental Setup

Models. We conduct experiments in three prompt-based settings including text-to-image (T2I), image-to-video (I2V), and text-to-video (T2V) generation tasks. In addition, we also test the effectiveness of AdaptiveDiffusion on the conditional image generation task. For the T2I task, we use Stable Diffusion-v1-5 (SD-1-5) [33] and Stable Diffusion XL (SDXL) [32] and evaluate on three different sampling schedulers (*i.e.*, DDIM [39], DPM-Solver++ [24], and Euler). For the I2V and T2V tasks, we utilize I2VGen-XL [52] and ModelScopeT2V [44] respectively. Note that we use ZeroScope-v2 instead of the original ModelScopeT2V model to generate watermark-free videos. For conditional image generation, we use LDM-4 [33] as the baseline model.

Benchmark Datasets. Following [27], we use ImageNet [7] and MS-COCO 2017 [21] to evaluate the results on class-conditional image generation and T2I tasks, respectively. For the I2V task, we randomly sample 100 prompts and reference images in AIGCBench [8]. For the T2V task, we conduct experiments on a widely-used benchmark MSR-VTT [45] and sample one caption for each video in the validation set as the test prompt. More details can be found in Appendix A.3.

Comparison Baselines. We compare AdaptiveDiffusion against DeepCache and Adaptive DPM-Solver in both generation quality and efficiency. Deepcache [27] caches high-level features of UNet to update the low-level features at each denoising step, thus reducing the computational cost of UNet. The latter [23] dynamically adjusts the step size by combining different orders of DPM-Solver.

Evaluation Metrics. For all tasks, we evaluate our proposed method in both quality and efficiency. We report MACs, latency, and speedup ratio to verify the efficiency. For the image generation task, following previous works [17, 16, 18], we evaluate image quality with commonly-used metrics, *i.e.*,

Table 1: Quantitative results on MS-COCO 2017 [21].

Model	Scheduler	Method	PSNR ↑	LPIPS ↓	FID ↓	MACs (T)	Mem (MB)	Latency (s)	Speedup Ratio
SD-1-5	DDIM	Deepcache	19.05	0.199	6.96	26	3940	1.7	1.58×
		Ours	**21.74**	**0.131**	**5.22**	**22**	**3850**	**1.5**	**1.77**×
	DPM++	Adaptive DPM	19.0	0.195	6.38	31	3896	3.5	1.25×
		Ours	**23.2**	**0.092**	**4.06**	**25**	**3852**	**2.8**	**1.57**×
	Euler	Deepcache	18.89	0.240	7.51	26	3894	1.6	1.57×
		Ours	**20.60**	**0.157**	**6.02**	**21**	**3848**	**1.3**	**1.98**×
SDXL	DDIM	Deepcache	21.9	0.221	7.34	**162**	14848	**7.4**	**1.75**×
		Ours	**25.4**	**0.141**	**5.13**	186	**14460**	7.8	1.66×
	DPM++	Deepcache	21.3	0.255	8.48	**162**	14800	**7.6**	**1.74**×
		Ours	**26.1**	**0.125**	**4.59**	190	**14454**	8.0	1.65×
	Euler	Deepcache	22.0	0.223	7.36	**162**	14796	**6.3**	**2.16**×
		Ours	**24.33**	**0.168**	**6.11**	174	**14458**	6.7	2.01×

Peak Signal Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS), and Fréchet Inception Distance (FID). For the video generation task, we use per-frame PSNR and LPIPS to measure the quality of generated videos. Besides, Fréchet Video Distance (FVD) [43] is also used to quantify the temporal coherence and quality of each frame. *Note that since our method achieves adaptive acceleration results, all reported metrics of our method are averaged across all prompts.*

Implementation Details. We conduct all experiments on RTX 3090 GPUs. For SD-1-5 and SDXL models, the original sampling timesteps T are set as 50, and two hyperparameters are set as: $\delta = 0.01$, $C_{\text{max}} = 4$. For LDM-4, $T = 250, \delta = 0.005, C_{\text{max}} = 10$. For I2VGen-XL and ModelScopeT2V, $T = 50, \delta = 0.007, C_{\text{max}} = 4$. More details of other methods are listed in Appendix A.3.

4.2 Main Results

4.2.1 Results on Image Generation

We first evaluate our method on T2I generation. As shown in Tab. 1, compared to DeepCache and Adaptive DPM methods, AdaptiveDiffusion achieves both higher quality and efficiency in various settings. For example, AdaptiveDiffusion achieves 0.092 LPIPS on SD-v1-5 [33], generating almost lossless images compared to those generated by the full-step denoising process. Meanwhile, the averaged speedup ratio across all testing prompts achieves 2.01× when using SDXL and Euler sampling scheduler. In addition, when comparing the generation performance between different models (*e.g.*, SD-1-5 and SDXL [32])) and schedulers (*e.g.*, DDIM, DPM-Solver++, and Euler), AdaptiveDiffusion shows stronger generalization capability to adapt to different settings.

Table 2: Quantitative results on ImageNet 256×256 [7].

Model	Method	PSNR ↑	LPIPS ↓	FID ↓	Mem (MB)	MACs (T)	Latency (s)	Speedup Ratio
LDM-4	Deepcache	24.98	0.098	5.22	3030	**5.6**	**1.4**	**6.35**×
	Ours	**25.79**	**0.090**	**4.91**	**2912**	6.5	1.6	5.56×

To further verify the effectiveness of our method on different image generation tasks, we also conduct experiments on LDM-4 for the conditional image generation task. As shown in Tab. 2, due to the larger timestep setting in the original denoising process, the slow change of latent change allows us to achieve faster generation with almost lossless quality at nearly 5.6x speedup. Compared with Deepcache, the AdpativeDiffusion obtains a better trade-off between generation quality and efficiency.

4.2.2 Results on Video Generation

We further conduct experiments on more tough tasks, *e.g.*, video generation tasks including I2V and T2V generation. As shown in Tab. 3, AdaptiveDiffusion can generate videos of lossless frames and similar video quality with a significant speedup against the original models' full-step generated videos. Specifically, our proposed method can achieve a significantly higher quality in single frame evaluation which can be seen by LPIPS and PSNR (*e.g.*, +6.38dB PSNR compared to DeepCache using I2VGen-XL). On the other hand, AdaptiveDiffusion can ensure temporal consistency as the original models due to the property of lossless acceleration which can be reflected on FVD.

4.3 Ablation Studies

Ablation Study on Skipping Threshold. As shown in the upper part of Tab. 4, we first analyze the effect of the skipping threshold δ. It can be observed that when the skipping threshold is relatively small, there will be fewer skipping steps, resulting in a relatively small yet still clear acceleration with

Table 3: Quantitative results on video generation tasks.

Model	Method	PSNR ↑	LPIPS ↓	FVD ↓	Mem (GB)	MACs (T)	Latency (s)	Speedup Ratio
I2VGen-XL	Deepcache	20.82	0.212	916.0	28.0 (+58.4%)	**1878**	**77**	**2.00×**
	Ours	**27.20**	**0.088**	**274.5**	**17.7 (+0.3‰)**	1983	80	1.93×
ModelScopeT2V	Deepcache	18.24	0.343	3383.0	14.2 (+50.8%)	1034	45	1.37×
	Ours	**27.70**	**0.081**	**470.6**	**8.7 (-7.8%)**	**948**	**42**	**1.46×**

Table 4: Ablation studies on hyperparameters using SDXL [32]. We conduct the ablation studies using 50-step Euler sampling scheduler for SDXL on COCO2017.

Hyper-params	Value	PSNR ↑	LPIPS ↓	FID ↓	MACs (T)	Latency (s)	Speedup Ratio
δ (C_{max} = 4)	0.005	34.3	0.023	0.96	262	9.7	1.38×
	0.008	28.6	0.079	3.22	217	7.7	1.74×
	0.01	24.3	0.168	6.11	174	6.7	2.01×
	0.015	21.0	0.282	9.49	119	5.1	2.64×
	0.02	21.0	0.289	9.79	106	4.4	3.06×
C_{max} (δ = 0.01)	4	24.3	0.168	6.11	174	6.7	2.01×
	6	23.3	0.217	7.74	147	5.7	2.37×
	8	22.7	0.256	9.43	135	5.2	2.59×
	10	22.1	0.307	11.91	123	4.7	2.86×

a much higher preservation quality. With the threshold gradually increasing, the speedup ratio will be largely improved but at the cost of quality degradation. It can be seen that the image quality does not significantly change with large thresholds (*i.e.*, 0.015 and 0.02). This is because the pre-defined maximum skipping steps prevent the further increase of skip steps. In this paper, we set the skipping threshold to 0.01 which is a better trade-off between the generation quality and inference speed.

Ablation Study on Maximum Skipping Steps. Further, we conduct ablation studies on the maximum skipping steps $C_{\max}$, as shown in the lower part of Tab. 4. With the increase of the maximum skipping steps, the quality of generated images continuously decreases. The reason is that when the timestep t_i approaches 0, a large number of denoising steps will be skipped due to the minor value of the third-order latent difference when the max-skip-step is relatively large. The phenomenon reveals that $C_{\max}$ can effectively prevent the continuous accumulation of generated image errors and ensure image quality.

Analysis on Sampling Steps. To evaluate the effectiveness of AdaptiveDiffusion on few-step sampling. It can be seen from Tab. 5 that AdaptiveDiffusion can further accelerate the denoising process under the few-step settings. Note that the hyperparameters (*i.e.*, δ, and $C_{\max}$) should be slightly adjusted according to the original sampling steps due to the varying updating magnitudes in different sampling steps. Specifically, a higher threshold δ and lower max-skip-step $C_{\max}$ can make better generation quality when reducing the sampling steps.

Table 5: Study on few-step sampling. Acceleration results with different original sampling steps using SDXL [32] on COCO2017.

Steps	PSNR ↑	LPIPS ↓	FID ↓	MACs (T)	Latency (s)	Speedup Ratio
50 steps	24.3	0.168	6.11	174	6.7	2.01×
25 steps	32.9	0.047	1.62	128	5.8	1.31×
15 steps	19.9	0.122	5.06	70	3.1	1.38×
10 steps	29.4	0.169	8.28	53	2.5	1.21×

4.4 Visualization Results

Generation Comparisons. To demonstrate the effectiveness of AdaptiveDiffusion more intuitively, we show some visualization results. Fig. 6 shows the results of the text-to-image generation task, it can be seen that AdaptiveDiffuion can better maintain image quality compared to Deepcache with nearly equal acceleration. Since AdaptiveDiffusion can adaptively determine which steps can be skipped, unimportant steps that have little impact on the final generation quality will be skipped during inference. Besides, to demonstrate the generalization of AdaptiveDiffusion on different tasks, we provide video generation results in Fig. 7. More generalization results can be found in Appendix A.4

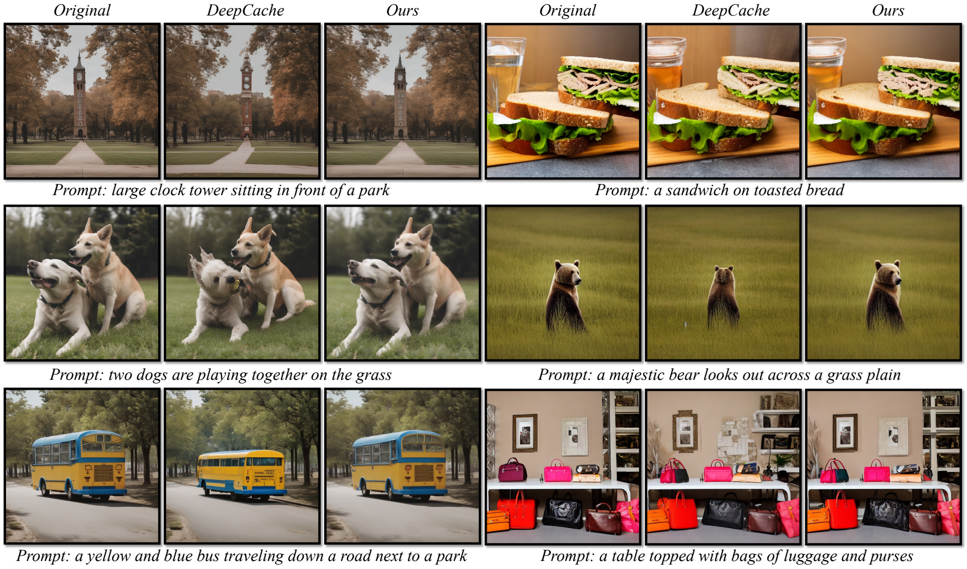

Original
DeepCache
Ours
Original
DeepCache
Ours
Prompt: large clock tower sitting in front of a park
Prompt: a sandwich on toasted bread
Prompt: two dogs are playing together on the grass
Prompt: a majestic bear looks out across a grass plain
Prompt: a yellow and blue bus traveling down a road next to a park
Prompt: a table topped with bags of luggage and purses

Figure 6: Qualitative results of text-to-image generation task using SDXL and SD-1-5 on MS-COCO 2017 benchmark. Left: SDXL, Right: SD-1-5.

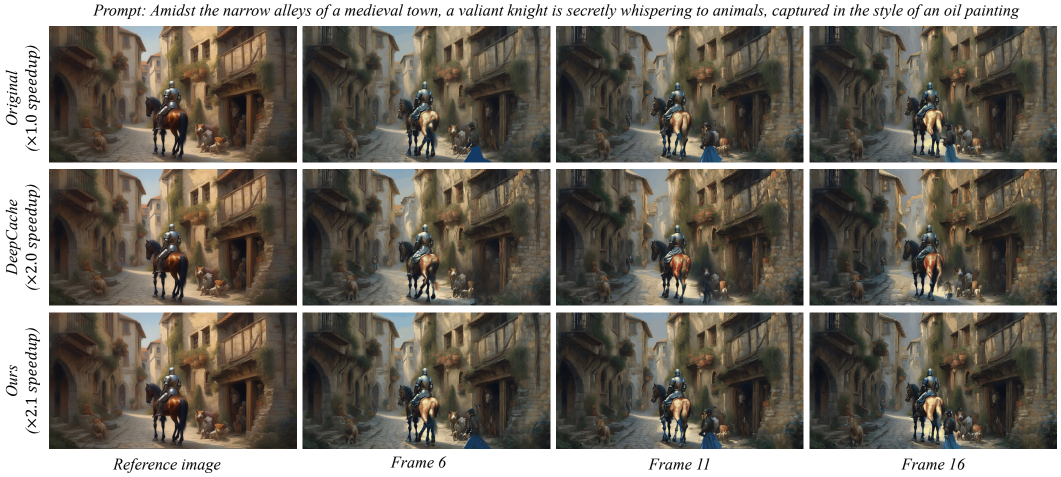

Prompt: Amidst the narrow alleys of a medieval town, a valiant knight is secretly whispering to animals, captured in the style of an oil painting
Original (×1.0 speedup)
DeepCache (×2.0 speedup)
Ours (×2.1 speedup)
Reference image
Frame 6
Frame 11
Frame 16

Figure 7: Qualitative results of image-to-video generation task using I2VGen-XL on AIGCBench.

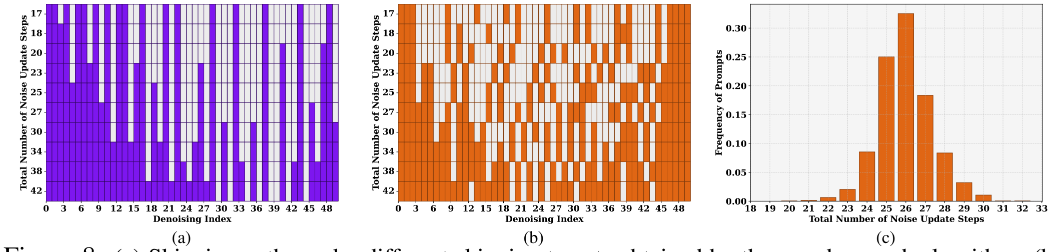


Figure 8: (a) Skipping paths under different skipping targets obtained by the greedy search algorithm. (b) Skipping paths under different skipping thresholds by the third-order estimator. (c) The frequency distribution of the skipping number of noise update steps for SDXL generating images on MS-COCO 2017 benchmark.

Denoising Path Comparisons. Fig. 8 illustrates the distribution of skipping paths at different skipping schemes. It can be observed from Fig. 8a and 8b that when the number of noise update steps keeps decreasing (more blank grids in the horizontal lines), both greedy searched paths and third-order estimated paths tend to prioritize the importance of early and late denoising steps. From Fig. 8c, we find that most prompts in the MS-COCO 2017 benchmark only need around 26 steps of noise update to generate an almost lossless image against the 50-step generation result.

5 Conclusion

In this paper, we explore the training-free diffusion acceleration and introduce AdaptiveDiffusion, which can dynamically select the denoising path according to given prompts. Besides, we perform the error analyses of the step-skipping strategy and propose to use the third-order estimator to indicate the computation redundancy. Experiments are conducted on MS-COCO 2017, ImageNet, AIGCBench and MSR-VTT, showing a good trade-off between high image quality and low inference cost.

6 Acknowledgements

The research was supported by the National Key R&D Program of China (Grant No. 2022ZD0160104), the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ1100102), Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0102), Shanghai Rising Star Program (Grant No. 23QD1401000), National Key Research and Development Program of China (No. 2022ZD0160101), Shanghai Natural Science Foundation (No. 23ZR1402900), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), National Natural Science Foundation of China (No. 62071127 and 62101137).

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[5] Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2061–2070, 2023.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(4):100152, 2023.

[9] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. *Advances in neural information processing systems*, 36, 2024.

[10] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9935–9946, 2023.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[12] Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, and Kurt Keutzer. Streamdiffusion: A pipeline-level solution for real-time interactive generation. *arXiv preprint arXiv:2312.12491*, 2023.

[13] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[14] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[15] Lijiang Li, Huixia Li, Xiawu Zheng, Jie Wu, Xuefeng Xiao, Rui Wang, Min Zheng, Xin Pan, Fei Chao, and Rongrong Ji. Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7105–7114, 2023.

[16] Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Ming-Yu Liu, Kai Li, and Song Han. Distrifusion: Distributed parallel inference for high-resolution diffusion models. *arXiv preprint arXiv:2402.19481*, 2024.

[17] Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. Gan compression: Efficient architectures for interactive conditional gans. In *CVPR*, 2020.

[18] Muyang Li, Ji Lin, Chenlin Meng, Stefano Ermon, Song Han, and Jun-Yan Zhu. Efficient spatially sparse inference for conditional gans and diffusion models. In *NeurIPS*, 2022.

[19] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *NeurIPS*, 2023.

[20] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[22] Zhijian Liu, Zhuoyang Zhang, Samir Khaki, Shang Yang, Haotian Tang, Chenfeng Xu, Kurt Keutzer, and Song Han. Sparse refinement for efficient high-resolution semantic segmentation. *arXiv preprint arXiv:2407.19014*, 2024.

[23] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

[24] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

[25] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[26] Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022.

[27] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[28] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[29] Taehong Moon, Moonseok Choi, EungGu Yun, Jongmin Yoon, Gayoung Lee, and Juho Lee. Early exiting for accelerated inference in diffusion models. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023.

[30] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

[31] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

[32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[35] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2021.

[36] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024.

[37] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

[38] Andy Shih, Suneel Belkhale, Stefano Ermon, Dorsa Sadigh, and Nima Anari. Parallel sampling of diffusion models. *NeurIPS*, 2023.

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020.

[40] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

[41] Haotian Tang, Shang Yang, Zhijian Liu, Ke Hong, Zhongming Yu, Xiuyu Li, Guohao Dai, Yu Wang, and Song Han. Torchsparse++: Efficient training and inference framework for sparse convolution on gpus. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 225–239, 2023.

[42] Shengkun Tang, Yaqing Wang, Caiwen Ding, Yi Liang, Yao Li, and Dongkuan Xu. Deediff: Dynamic uncertainty-aware early exiting for accelerating diffusion model generation. *arXiv preprint arXiv:2309.17074*, 2023.

[43] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[44] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[45] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[46] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22552–22562, 2023.

[47] Hancheng Ye, Chong Yu, Peng Ye, Renqiu Xia, Yansong Tang, Jiwen Lu, Tao Chen, and Bo Zhang. Once for both: Single stage of importance and sparsity search for vision transformer compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5578–5588, 2024.

[48] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[49] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *ICLR*, 2022.

[50] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models. 2022.

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[52] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

A Appendix

Due to the nine-page limitation of the manuscript, we provide more details and visualizations from the following aspects:

A.1 Limitations and Broader Impacts

Currently, AdaptiveDiffusion is only evaluated on image generation and video generation tasks. A long-term vision is to achieve lossless acceleration on any modalities generation such as 3D and speech. Besides, although AdaptiveDiffusion can improve the speed on few-step (*e.g.*, 10 steps) settings, it may not work for extreme few-step generation for the rapid changes of the latent features.

A.2 Method Explanations

A.2.1 Error Estimation Induced by Single-step Skipping

In Sec. 3.2, the error between the latents of the continuous noise update and one-step skipping of noise update is stated to be upper-bounded of the difference between the previous two output latents. Here, we provide the detailed proof of this statement.

Assumptions We first make the assumptions that:
(1) $\epsilon_\theta(x, t)$ is Lipschitz w.r.t to its paramters x and t;
(2) The 1st-order difference $\mathit{\Delta} x_i = x_i - x_{i+1}$ exists and is continuous for $0 \leq i \leq T - 1$.

Proof Take i-th step as an example to perform the one-step skipping of noise prediction, we can obtain the following update formulations.

$$\begin{aligned} x_{i+1} &= f(i+1) \cdot x_{i+2} - g(i+1) \cdot \epsilon_\theta(x_{i+2}, t_{i+2}); \\ x_i &= f(i) \cdot x_{i+1} - g(i) \cdot \epsilon_\theta(x_{i+1}, t_{i+1}); \\ x_{i-1} &= f(i-1) \cdot x_i - g(i-1) \cdot \epsilon_\theta(x_{i+1}, t_{i+1}); \\ x_{i-2} &= f(i-2) \cdot x_{i-1} - g(i-2) \cdot \epsilon_\theta(x_{i-1}, t_{i-1}); \end{aligned} \tag{5}$$

Here, the i-th step of noise prediction is replaced by reusing the $i+1$-th noise prediction. Then, the error caused by skipping the i-th noise update, $\varepsilon_{i-1} = \|x_{i-1} - x_{i-1}^{ori}\|$, can be derived as follows.

$$\begin{aligned} \|\varepsilon_{i-1}\| &= \|g(i-1) \cdot [\epsilon_\theta(x_{i+1}, t_{i+1}) - \epsilon_\theta(x_i, t_i)]\| \\ &= \|g(i-1) \cdot [\epsilon_\theta(x_{i+1}, t_{i+1}) - \epsilon_\theta(x_i, t_{i+1}) + \epsilon_\theta(x_i, t_{i+1}) - \epsilon_\theta(x_i, t_i)]\| \\ &\leq \|g(i-1) \cdot \mathcal{O}(t_i - t_{i+1})\| + \|g(i-1) \cdot \mathcal{O}(x_i - x_{i+1})\| \\ &= \mathcal{O}(t_i - t_{i+1}) + \mathcal{O}(x_i - x_{i+1}). \end{aligned} \tag{6}$$

The proof utilizes the property that $\|\epsilon_\theta(x_i, t_{i+1}) - \epsilon_\theta(x_{i+1}, t_{i+1})\|$ and $\|\epsilon_\theta(x_i, t_i) - \epsilon_\theta(x_i, t_{i+1})\|$ are upper-bounded by $\mathcal{O}(x_i - x_{i+1})$ and $\mathcal{O}(t_i - t_{i+1})$ respectively according to the Lipschitz continuity.

A.2.2 Error Estimation Induced by Two-step Skipping

We further estimate the situation of consecutive two-step skipping of the noise prediction model. Assuming that noise prediction is skipped at the i-th and $(i-1)$-th steps, then we have the following formulation.

$$\begin{aligned} x_{i+1} &= f(i+1)\cdot x_{i+2} - g(i+1)\cdot \epsilon_\theta(x_{i+2}, t_{i+2});\\ x_i &= f(i)\cdot x_{i+1} - g(i)\cdot \epsilon_\theta(x_{i+1}, t_{i+1});\\ x_{i-1} &= f(i-1)\cdot x_i - g(i-1)\cdot \epsilon_\theta(x_{i+1}, t_{i+1});\\ x_{i-2} &= f(i-2)\cdot x_{i-1} - g(i-2)\cdot \epsilon_\theta(x_{i+1}, t_{i+1}). \end{aligned} \tag{7}$$

The error, $\varepsilon_{i-2} = \|x_{i-2} - x_{i-2}^{ori}\|$, can be derived as follows.

$$\begin{aligned} \varepsilon_{i-2} &= \|f(i-2)\cdot(x_{i-1} - x_{i-1}^{ori}) - g(i-2)\cdot[\epsilon_\theta(x_{i+1}, t_{i+1}) - \epsilon_\theta(x_{i-1}^{ori}, t_{i-1})]\|\\ &= \|f(i-2)\cdot g(i-1)\cdot[\epsilon_\theta(x_i, t_i) - \epsilon_\theta(x_{i+1}, t_{i+1})]\\ &\quad - g(i-2)\cdot[\epsilon_\theta(x_{i+1}, t_{i+1}) - \epsilon_\theta(x_{i-1}^{ori}, t_{i-1})]\|\\ &= \|h^2(i-1)\cdot[\epsilon_\theta(x_i, t_i) - \epsilon_\theta(x_i, t_{i+1}) + \epsilon_\theta(x_i, t_{i+1}) - \epsilon_\theta(x_{i-1}^{ori}, t_{i-1})]\\ &\quad - g(i-2)\cdot[\epsilon_\theta(x_{i+1}, t_{i+1}) - \epsilon_\theta(x_{i-1}^{ori}, t_{i-1})]\|\\ &\le \|h^2(i-1)\cdot\mathcal{O}(t_i - t_{i+1})\| + \|h^2(i-1)\cdot\mathcal{O}(x_i - x_{i+1})\|\\ &\quad + \|g(i-2)\cdot\mathcal{O}(x_{i-1} - x_{i-1}^{ori})\|\\ &= \mathcal{O}(t_i - t_{i+1}) + \mathcal{O}(x_i - x_{i+1}) + \mathcal{O}(t_{i-1} - t_i) + \mathcal{O}(x_{i-1} - x_i). \end{aligned} \tag{8}$$

From the above derivations, it can be observed that the skipping error is clearly related to and upper-bounded by the accumulation of previous latent differences. Here $h^2\left(i-1\right)$ is defined as $h^2\left(i-1\right) := g\left(i-1\right)f\left(i-2\right)$. The above conclusion of skipping error's upper bound can be easily extended to any situation where finite-step skipping is used.

A.2.3 Error Estimation Induced by k-step Skipping

Take i-th step as an example to perform the k-step ($k \ge 2$) skipping of noise prediction, we can obtain the following update formulations.

$$\begin{aligned} x_i &= f(i)\cdot x_{i+1} - g(i)\cdot\epsilon_\theta(x_{i+1}, t_{i+1});\\ x_{i-1} &= f\left(i-1\right)\cdot x_i - g\left(i-1\right)\cdot\epsilon_\theta\left(x_{i+1}, t_{i+1}\right);\\ x_{i-2} &= f\left(i-2\right)\cdot x_{i-1} - g\left(i-2\right)\cdot\epsilon_\theta\left(x_{i+1}, t_{i+1}\right);\\ &\ \ \vdots\\ x_{i-k} &= f\left(i-k\right)\cdot x_{i-k+1} - g\left(i-k\right)\cdot\epsilon_\theta\left(x_{i+1}, t_{i+1}\right).\\ \Rightarrow \varepsilon_{i-k} &= \left\|x_{i-k} - x_{i-k}^{ori}\right\|\\ &= \left\|f\left(i-k\right)\left(x_{i-k+1} - x_{i-k+1}^{ori}\right) - g\left(i-k\right)\left[\epsilon_\theta\left(x_{i+1}, t_{i+1}\right) - \epsilon_\theta\left(x_{i-k+1}, t_{i-k+1}\right)\right]\right\|\\ &\leqslant f\left(i-k\right)\varepsilon_{i-k+1} + g\left(i-k\right)\left\|\epsilon_\theta\left(x_{i+1}, t_{i+1}\right) - \epsilon_\theta\left(x_{i-k+1}, t_{i-k+1}\right)\right\|\\ &\leqslant \sum_{m=1}^{k-1}\left\|h^{k-m+1}\left(i-m\right)\cdot\mathcal{O}\left(t_{i-m+1} - t_{i-m+2}\right)\right\|\\ &\quad + \sum_{m=1}^{k-1}\left\|h^{k-m+1}\left(i-m\right)\cdot\mathcal{O}\left(x_{i-m+1} - x_{i-m+2}\right)\right\|\\ &\quad + \left\|g\left(i-k\right)\cdot\mathcal{O}\left(x_{i-k+1} - x_{i-k+2}\right)\right\| + \left\|g\left(i-k\right)\cdot\mathcal{O}\left(t_{i-k+1} - t_{i-k+2}\right)\right\|\\ &= \sum_{m=1}^{k}\mathcal{O}\left(t_{i-m+1} - t_{i-m+2}\right) + \mathcal{O}\left(x_{i-m+1} - x_{i-m+2}\right). \end{aligned} \tag{9}$$

The derivation also utilizes the property that $\|\epsilon_\theta(x_i, t_{i+1}) - \epsilon_\theta(x_{i+1}, t_{i+1})\|$ and $\|\epsilon_\theta(x_i, t_i) - \epsilon_\theta(x_i, t_{i+1})\|$ are upper-bounded by $\mathcal{O}(x_i - x_{i+1})$ and $\mathcal{O}(t_i - t_{i+1})$ respectively according to the Lipschitz continuity. Here $h^{k-m+1}\left(i-m\right)$ is defined as $h^{k-m+1}\left(i-m\right) := g\left(i-m\right)\prod_{j=1}^{k-m} f\left(i-m-j\right)$.

From the above derivation, it can be observed that the error of an arbitrary k-step skipping scheme is related to and upper-bounded by the accumulation of previous latent differences. Therefore, if the skipping step of noise prediction is large, the upper bound of the error will naturally increase, which is also empirically demonstrated by Fig. 5b.

A.2.4 Theoretical Relation between the 3rd-order Estimator and Optimal Skipping Strategy

To explore the theoretical relationship between the third-order estimator and the skipping strategy, we need to formulate the difference between the neighboring noise predictions. According to Eq. (1), we can get the following first-order differential equations regarding the latent x.

$$\begin{aligned} \Delta x_i = x_i - x_{i+1} = \left[1 - f\left(i\right)\right] x_{i+1} - g\left(i\right) \cdot \epsilon_\theta\left(x_{i+1}, t_{i+1}\right); \\ \Delta x_{i-1} = x_{i-1} - x_i = \left[1 - f\left(i-1\right)\right] x_i - g\left(i-1\right) \cdot \epsilon_\theta\left(x_i, t_i\right). \end{aligned} \tag{10}$$

Now, let $u\left(i\right) \coloneqq 1 - f\left(i-1\right)$, and we further derive the second-order differential equations based on the above equations.

$$\begin{aligned} &\Delta x_{i-1} - \Delta x_i \\ =& u\left(i\right) x_i - u\left(i+1\right) x_{i+1} + g\left(i\right) \cdot \epsilon_\theta\left(x_{i+1}, t_{i+1}\right) - g\left(i-1\right) \cdot \epsilon_\theta\left(x_i, t_i\right) \\ =& u\left(i\right)\left(x_i - x_{i-1}\right) + u\left(i\right) x_{i-1} - u\left(i+1\right)\left(x_{i+1} - x_i\right) - u\left(i+1\right) x_i + g\left(i\right) \cdot \epsilon_\theta\left(x_{i+1}, t_{i+1}\right) \\ &- g\left(i-1\right) \cdot \epsilon_\theta\left(x_i, t_i\right) \\ =& u\left(i\right) \Delta x_{i-1} - u\left(i+1\right) \Delta x_i + \Delta\left[u\left(i\right) x_{i-1}\right] + g\left(i\right) \cdot \epsilon_\theta\left(x_{i+1}, t_{i+1}\right) - g\left(i-1\right) \cdot \epsilon_\theta\left(x_i, t_i\right) \\ =& u\left(i\right) \Delta x_{i-1} - u\left(i+1\right) \Delta x_i + \Delta\left[u\left(i\right) x_{i-1}\right] + g\left(i\right)\left[\epsilon_\theta\left(x_{i+1}, t_{i+1}\right) - \epsilon_\theta\left(x_i, t_i\right)\right] \\ &+ \left[g\left(i\right) - g\left(i-1\right)\right] \epsilon_\theta\left(x_i, t_i\right) \\ =& u\left(i\right) \Delta x_{i-1} - u\left(i+1\right) \Delta x_i + \Delta\left[u\left(i\right) x_{i-1}\right] - g\left(i\right) \Delta\epsilon_\theta^i - \Delta g\left(i\right) \cdot \epsilon_\theta\left(x_i, t_i\right). \end{aligned} \tag{11}$$

After simplification of the above equation, we can get the following formulation:

$$f\left(i-1\right) \Delta x_{i-1} - f\left(i\right) \Delta x_i = \Delta\left[u\left(i\right) x_{i-1}\right] - g\left(i\right) \Delta\epsilon_\theta^i - \Delta g\left(i\right) \cdot \epsilon_\theta\left(x_i, t_i\right). \tag{12}$$

From the above equation, we can observe that the difference between noise predictions $\Delta\epsilon_\theta^i$ is related to the first- and second-order derivatives of x_i, as well as the noise prediction $\epsilon_\theta\left(x_i, t_i\right)$. Therefore, it would be difficult to estimate the difference without $\epsilon_\theta\left(x_i, t_i\right)$. Now we consider the third-order differential equation. From the above equation, we further obtain the following formulation.

$$\begin{aligned} & f\left(i\right) \Delta x_i - f\left(i+1\right) \Delta x_{i+1} = \Delta\left[u\left(i+1\right) x_i\right] - g\left(i+1\right) \Delta\epsilon_\theta^{i+1} - \Delta g\left(i+1\right) \cdot \epsilon_\theta\left(x_{i+1}, t_{i+1}\right); \\ \Rightarrow & \Delta\left[f\left(i-1\right) \Delta x_{i-1}\right] - \Delta\left[f\left(i\right) \Delta x_i\right] = \Delta^{(2)}\left[u\left(i\right) x_{i-1}\right] - \Delta\left[g\left(i\right) \Delta\epsilon_\theta^i\right] - \Delta\left[\Delta g\left(i\right) \cdot \epsilon_\theta\left(x_i, t_i\right)\right]; \\ \Rightarrow & \Delta\left[\Delta g\left(i\right) \cdot \epsilon_\theta\left(x_i, t_i\right)\right] = -\Delta^{(2)}\left[f\left(i-1\right) \Delta x_{i-1}\right] + \Delta^{(2)}\left[u\left(i\right) x_{i-1}\right] - \Delta\left[g\left(i\right) \Delta\epsilon_\theta^i\right]. \end{aligned} \tag{13}$$

From the above equation, it can be observed that the difference of the neighboring noise predictions is explicitly related to the third- and second-order derivatives of x_i, as well as the second-order derivative of ϵ_θ^i. Since $\lim_{i\to0} f\left(i\right) = 1, \lim_{i\to0} u\left(i\right) = 0, \lim_{i\to0} g\left(i\right) = 0$, we can finally get the conclusion that $\Delta\epsilon_\theta^i|_{i\to0} = \mathcal{O}\left(\Delta^{(3)} x_{i-1}\right)$.

A.2.5 Algorithms

Algorithm of greedy search for optimal skipping path. The specific details of the greedy search algorithm used as the optimal denoising path are shown in Alg. (1).

Algorithm 1 Greedy Search for the Optimal Skipping Path.

```
Input: Noise Prediction Model $\epsilon_\theta$, Sampling Scheduler $\phi$, Decoder $\mathcal{F}_d$, Target Skipping Step Number $N$, Sample Step $T$, Conditional embedding $c$;
Initialize Skipping Path $\mathcal{S}$ = [True] * $T$, Current Skipping Step Number $n_{skip}$ = 0.
Compute $x_0^{ori}$ by Eq. (1).
while $n_{skip} < N$ do
    Initialize $df$ = [];
    for $i$ in range($T - 1$) do
        if $\mathcal{S}[i]$==True then
            temp_path = $\mathcal{S}$.copy();
            temp_path[$i$]=False;
            Generate $x_0^{temp}$ by Eq. (2);
            Compute $\ell_1$ difference $\mathcal{L}_0 = \|x_0^{temp} - x_0^{ori}\|$;
            $df$.append($\mathcal{L}_0$);
        end if
    end for
    index = argmin($df$);
    Set $\mathcal{S}$[index] = False;
end while
return $\mathcal{S}$.
```

Algorithm of the overall skipping strategy in AdaptiveDiffusion. The designed skipping strategy used in our AdaptiveDiffusion are elaborated in Alg. (2).

Algorithm 2 AdaptiveDiffusion.

```
Input: Noise Prediction Model $\epsilon_\theta$, Sampling Scheduler $\phi$, Decoder $\mathcal{F}_d$, Sample Step $T$, Conditional embedding $c$, Maximum Skipping Step Number $C_{\text{max}}$, Threshold $\delta$;
Initialize Random Noise $x$, Skipping Path $\mathcal{S}$ = [], Previous Differential List $P_{\text{diff}}$ = [], Previous Latent List $P_{\text{latent}}$ = [].
for $i$ in range($T - 1$) do
    if $i \leq 2$ then
        $\text{O}_{pv}$ = O = $\epsilon(x, i)$;
        if $i \geq 1$ then
            $P_{\text{diff}}$.append($\|x - P_{\text{latent}}[-1]\|$);
        end if
    else
        if $\mathcal{S}[-1]$==True then
            $\text{O}_{pv}$ = O = $\epsilon(x, i)$;
        else
            O = $\text{O}_{pv}$;
        end if
    end if
    Compute $\phi(x)$ by Eq. (1);
    if $i \geq 3$ then
        $\mathcal{S}$.append($\longleftarrow \xi(x)$) in Eq. (4);
    end if
    $P_{\text{latent}}$.append($x$);
end for
return $\mathcal{F}_d(x)$.
```

A.3 Experimental Details

A.3.1 More Evaluation Details

Statistical Analysis of the Estimated Skipping Path. In Fig. 5, we aim to evaluate the correlation between the estimated skipping paths by our method and the searched optimal skipping paths under different target skipping numbers. This evaluation is crucial to validate the effectiveness of our approach in accurately predicting skipping paths. The paths are represented as sequences of binary choices (0 for a skipping step, 1 for a non-skipping step) at each timestep. We employ a χ^2 test of independence to assess the statistical correlation between the estimated and optimal paths. We first randomly select several prompts from the test set and use the greedy search algorithm to greedily search the optimal skipping paths under different skipping target steps. For the estimated skipping paths, we discard the setting of $C_{\max}$ to test the effectiveness of the third-order estimation without regularization and refine the threshold δ in a wide range to achieve various skipping paths with different skipping step numbers. Then, for each skipping step number, we construct a 2×2 contingency table from the searched optimal skipping paths and the estimated skipping paths, with the same skipping step numbers. Finally, using the contingency table, we compute the χ^2 statistics and the corresponding p-value to assess the independence between the estimated and the searched optimal skipping paths.

Evaluation Tools and Details. Following [16], we use TorchMetrics [*] to calculate PSNR and LPIPS, and use CleanFID [31][†] to calculate FID. Since our method is an adaptive method which means it can choose a suitable skipping path for different prompts, we set batch size to 1 during evaluation. The MACs reported in our experiments are the total MACs in the denoising process following [16]. Besides, the reported MACs, speedup ratio, and latency are the average over the whole dataset.

CodeBase. For a fair comparison, when compared with Adaptive DPM-Solver, we use the official codebase of DPM-Solver[‡]. Experiments for LDM-4-G are based on the official LDM codebase[§] and DeepCache codebase[¶]. Other experiments are based on Diffusers[‖].

Details of Reproduced DeepCache. In this paper, we mainly compare our AdaptiveDiffusion with DeepCache. To ensure the generation quality, we set `cache_interval=3` and `cache_branch_id=3` for SD-v1-5, SDXL, and ModelScope. Since the memory cost will be largely improved on video generation tasks especially high-resolution tasks (*e.g.*, I2VGen-XL), We set `cache_interval=3` and `cache_branch_id=0` to reduce the memory costs. For class-conditional image generation task, we follow the official setting and set `cache_interval=10`.

A.3.2 Prompts in AIGCBench

We randomly sample 100 prompts and their corresponding images as the test set for I2V task. Here, we give the list of the selected prompts in AIGCBench in Fig. 9 and Fig. 10. The selected test set includes different styles of images and prompts such as animation, realism, and oil painting.

A.3.3 Experiments on Unconditional Image Generation

We provide the acceleration performance of different methods on pure image generation for further comparisons. Specifically, following the experimental setting in Deepcache, we conduct unconditional image generation experiments on CIFAR10 [14] and LSUN [48] datasets. As shown in Table 6, our method achieves a larger speedup ratio and higher image quality than Deepcache on both benchmarks.

Table 6: Performance on Unconditional Image Generation.

Datasets	Methods	FID ↓	Speedup Ratio
CIFAR10	Deepcache	10.17	2.07x
	Ours	**7.97**	**2.09x**
LSUN	Deepcache	9.13	1.48x
	Ours	**7.96**	**2.35x**

A.3.4 Effectiveness on SDE Solver

We also explore the effectiveness of our work on SDE solver. Compared with the ODE solver, the SDE solver includes an additional noise item for the latent update, which is unpredictable by previous randomly generated noises. When the magnitude of random noise is not ignorable, the third-order derivative of the neighboring latents cannot accurately evaluate the difference between the neighboring noise predictions. Therefore, to apply our

[*] https://github.com/Lightning-AI/torchmetrics

[†] https://github.com/GaParmar/clean-fid

[‡] https://github.com/LuChengTHU/dpm-solver

[§] https://github.com/CompVis/latent-diffusion

[¶] https://github.com/horseee/DeepCache

[‖] https://github.com/huggingface/diffusers method to SDE solvers, we should design an additional indicator that decides whether the randomly generated noise is minor enough or relatively unchanged to trigger the third-order estimator. In this case, we design an additional third-order estimator for the scaled randomly generated noise. When the third-order derivatives of both the latent and the scaled randomly generated noises are under the respective threshold, the noise prediction can be skipped by reusing the cached noise.

To validate the effectiveness of our improved method, we conduct experiments for SDXL with the SDE-DPM solver on COCO2017. The results are shown in the following table. Compared with Deepcache, our method can achieve higher image quality with a comparable speedup ratio, indicating the effectiveness of AdaptiveDiffusion on SDE solvers.

Table 7: Performance on SDE-DPM Solver.

Method	PSNR	LPIPS	FID	Latency (s)	Speedup Ratio
Deepcache	16.44	0.346	8.15	9.2	1.63x
Ours	**18.80**	**0.232**	**6.03**	9.8	1.53x

A.4 More Generation Visualizations

Here, we further show more qualitative results tested on text-to-image task using LDM-4 on MS-COCO 2017 and text-to-video generation task using ModelScopeT2V on MSR-VIT, which are illustrated in Fig. 11 and Fig. 12 respectively.

1. *Behold a valiant knight in the throes of exploring a cave surrounded by the secret chambers of an underground laboratory, envisioned as vibrant pixel art*
2. *Discover a mischievous fairy, eagerly searching for hidden treasure amidst the neon-lit skyscrapers of a futuristic city, artfully rendered through the abstract lens of picasso*
3. *Behold a playful panda in the throes of carefully repairing a broken robot surrounded by the dusty streets of an old western town at high noon, envisioned with the soft touch of watercolor*
4. *Discover a mischievous fairy, running a marathon amidst the tranquil waters of a serene lake, artfully rendered with the soft touch of watercolor*
5. *Encounter a noble king as they are engaged in valiantly fighting a monstrous beast a mysterious crossroads in a mystical forest, all depicted with the soft touch of watercolor*
6. *Amidst the front lines of an ancient battlefield, a gentle ogre is building a snowman, captured like a lively cartoon*
7. *Behold a valiant knight in the throes of cleverly solving a puzzle surrounded by a bustling space station orbiting earth, envisioned as a stunning 3d render*
8. *Behold a wailing banshee in the throes of casting a spell surrounded by the crashing waves of the cerulean sea, envisioned in the impassioned strokes of van gogh*
9. *Behold a rogue robot in the throes of exploring a cave surrounded by a creaking pirate ship, envisioned with the soft touch of watercolor*
10. *Amidst the crashing waves of the cerulean sea, a wise old man is soaring through the sky, captured as vibrant pixel art*
11. *Amidst a sun-dappled forest, a valiant knight is casting a spell, captured with photorealistic precision*
12. *Within the realm of the neon-lit skyscrapers of a futuristic city, a rogue robot casting a spell, each moment immortalized in the style of an oil painting*
13. *Amidst the sandy shores of a deserted island, a noble king is casting a spell, captured with the stark realism of a photo*
14. *Discover a wailing banshee, hungrily eating a feast amidst the lush canopy of a deep jungle, artfully rendered with the soft touch of watercolor*
15. *Amidst the sandy shores of a deserted island, a brave girl is meticulously investigating a mysterious crime scene, captured as vibrant pixel art*
16. *Encounter a mischievous fairy as they are engaged in conducting an experiment the front lines of an ancient battlefield, all depicted as vibrant pixel art*
17. *Discover a forest-dwelling nymph, exploring a cave amidst the crashing waves of the cerulean sea, artfully rendered through the abstract lens of picasso*
18. *Amidst the tranquil waters of a serene lake, a mischievous fairy is hunting for ghosts, captured with photorealistic precision*
19. *Amidst the eerie halls of a haunted house, a playful panda is secretly whispering to animals, captured in the style of an oil painting*
20. *Amidst the crashing waves of the cerulean sea, a wise old man is performing magic tricks, captured like a lively cartoon*
21. *Discover a noble king, brewing a potion amidst the sandy shores of a deserted island, artfully rendered as a stunning 3d render*
22. *Amidst the sandy shores of a deserted island, a battle-scarred cyborg is building a snowman, captured with the stark realism of a photo*
23. *Amidst the sandy shores of a deserted island, a playful panda is casting a spell, captured in the style of an oil painting*
24. *Discover a fearsome dragon, secretly whispering to animals amidst the vibrant heart of a bustling market square, artfully rendered with the soft touch of watercolor*
25. *Behold a treasure-seeking pirate in the throes of exploring a cave surrounded by the echoing depths of a magical cave, envisioned through the abstract lens of picasso*
26. *Discover a mischievous fairy, running a marathon amidst a bustling space station orbiting earth, artfully rendered as vibrant pixel art*
27. *Behold a wise old man in the throes of painting a masterpiece surrounded by the neon-lit skyscrapers of a futuristic city, envisioned as vibrant pixel art*
28. *Within the realm of the echoing depths of a magical cave, a gentle ogre casting a spell, each moment immortalized with the soft touch of watercolor*
29. *Discover a playful panda, running a marathon amidst the tranquil waters of a serene lake, artfully rendered in the impassioned strokes of van gogh*
30. *Encounter a noble king as they are engaged in building a snowman a blooming enchanted garden, all depicted with photorealistic precision*
31. *Within the realm of the echoing depths of a magical cave, a battle-scarred cyborg masterfully riding a bike, each moment immortalized through the abstract lens of picasso*
32. *Amidst the narrow alleys of a medieval town, a valiant knight is secretly whispering to animals, captured in the style of an oil painting*
33. *Discover a forest-dwelling nymph, deciphering a map amidst a creaking pirate ship, artfully rendered like a lively cartoon*
34. *Behold a gentle ogre in the throes of masterfully riding a bike surrounded by the backdrop of an alien planet's red skies, envisioned in the impassioned strokes of van gogh*
35. *Encounter a mystical mermaid as they are engaged in cleverly solving a puzzle a mysterious crossroads in a mystical forest, all depicted with the stark realism of a photo*
36. *Behold a curious alien in the throes of playing the violin surrounded by a creaking pirate ship, envisioned as vibrant pixel art*
37. *Behold a curious alien in the throes of skillfully strumming the guitar surrounded by the narrow alleys of a medieval town, envisioned in the style of an oil painting*
38. *Behold a noble king in the throes of casting a spell surrounded by the neon-lit skyscrapers of a futuristic city, envisioned with photorealistic precision*
39. *Discover a brave girl, secretly whispering to animals amidst the dusty streets of an old western town at high noon, artfully rendered as a stunning 3d render*
40. *Encounter a playful panda as they are engaged in playing the violin the tranquil waters of a serene lake, all depicted through the abstract lens of picasso*
41. *Encounter a treasure-seeking pirate as they are engaged in deciphering a map the backdrop of an alien planet's red skies, all depicted with the stark realism of a photo*
42. *Encounter a brave girl as they are engaged in conducting an experiment the sandy shores of a deserted island, all depicted with the soft touch of watercolor*
43. *Within the realm of the eerie halls of a haunted house, a treasure-seeking pirate hunting for ghosts, each moment immortalized in the style of an oil painting*
44. *Encounter a fearsome dragon as they are engaged in exploring a cave the sandy shores of a deserted island, all depicted in the style of an oil painting*
45. *Discover a brave girl, skillfully strumming the guitar amidst a sun-dappled forest, artfully rendered like a lively cartoon*
46. *Discover a valiant knight, building a snowman amidst the backdrop of an alien planet's red skies, artfully rendered in the style of an oil painting*
47. *Encounter a playful panda as they are engaged in conducting an experiment the frosty peak of a snowy mountain, all depicted as a stunning 3d render*
48. *Encounter a noble king as they are engaged in hunting for ghosts the echoing depths of a magical cave, all depicted through the abstract lens of picasso*
49. *Encounter a mystical mermaid as they are engaged in hungrily eating a feast the ancient walls of a crumbling castle, all depicted in the impassioned strokes of van gogh*
50. *Amidst a sun-dappled forest, a mischievous fairy is carefully repairing a broken robot, captured in the style of an oil painting*
51. *Encounter a noble king as they are engaged in deciphering a map the frosty peak of a snowy mountain, all depicted in the style of an oil painting*
52. *Amidst a creaking pirate ship, a curious alien is conducting an experiment, captured in the impassioned strokes of van gogh*
53. *Discover a curious alien, carefully repairing a broken robot amidst the eerie halls of a haunted house, artfully rendered like a lively cartoon*
54. *Encounter a genius scientist as they are engaged in soaring through the sky the vibrant heart of a bustling market square, all depicted like a lively cartoon*
55. *Discover a forest-dwelling nymph, deciphering a map amidst the ancient walls of a crumbling castle, artfully rendered like a lively cartoon*
56. *Amidst the sandy shores of a deserted island, a mystical mermaid is painting a masterpiece, captured like a lively cartoon*
57. *Discover a time-traveling scholar, performing magic tricks amidst a sun-dappled forest, artfully rendered in the style of an oil painting*
58. *Amidst the vibrant heart of a bustling market square, a mystical mermaid is cleverly solving a puzzle, captured as vibrant pixel art*
59. *Behold a time-traveling scholar in the throes of painting a masterpiece surrounded by a sun-dappled forest, envisioned with the soft touch of watercolor*
60. *Encounter a gentle ogre as they are engaged in masterfully riding a bike a bustling space station orbiting earth, all depicted like a lively cartoon*

Figure 9: Prompts used in AIGCBench.

61. *Encounter an adventurous astronaut as they are engaged in brewing a potion a bustling space station orbiting earth, all depicted with photorealistic precision*
62. *Encounter a playful panda as they are engaged in cleverly solving a puzzle the narrow alleys of a medieval town, all depicted through the abstract lens of picasso*
63. *Discover a noble king, painting a masterpiece amidst the crashing waves of the cerulean sea, artfully rendered through the abstract lens of picasso*
64. *Amidst the ancient walls of a crumbling castle, a time-traveling scholar is painting a masterpiece, captured as vibrant pixel art*
65. *Discover a brave girl, secretly whispering to animals amidst a blooming enchanted garden, artfully rendered with photorealistic precision*
66. *Behold a gentle ogre in the throes of hungrily eating a feast surrounded by the eerie halls of a haunted house, envisioned with photorealistic precision*
67. *Behold a genius scientist in the throes of secretly whispering to animals surrounded by the sandy shores of a deserted island, envisioned like a lively cartoon*
68. *Encounter a brave girl as they are engaged in conducting an experiment a sun-dappled forest, all depicted with the stark realism of a photo*
69. *Within the realm of a blooming enchanted garden, a rogue robot deciphering a map, each moment immortalized like a lively cartoon*
70. *Within the realm of a sun-dappled forest, a curious alien valiantly fighting a monstrous beast, each moment immortalized through the abstract lens of picasso*
71. *Encounter a valiant knight as they are engaged in gracefully dancing under the moonlight the narrow alleys of a medieval town, all depicted with the soft touch of watercolor*
72. *Within the realm of a creaking pirate ship, a mystical mermaid cleverly solving a puzzle, each moment immortalized with the soft touch of watercolor*
73. *Amidst the sandy shores of a deserted island, a fearsome dragon is hunting for ghosts, captured with the soft touch of watercolor*
74. *Encounter a wailing banshee as they are engaged in conducting an experiment the ruins of an ancient temple, all depicted with the stark realism of a photo*
75. *Amidst a sun-dappled forest, a brave girl is meticulously investigating a mysterious crime scene, captured as a stunning 3d render*
76. *Amidst the frosty peak of a snowy mountain, a wise old man is performing magic tricks, captured like a lively cartoon*
77. *Discover a treasure-seeking pirate, exploring a cave amidst the vibrant heart of a bustling market square, artfully rendered in the impassioned strokes of van gogh*
78. *Amidst the narrow alleys of a medieval town, a gentle ogre is soaring through the sky, captured with the stark realism of a photo*
79. *Behold a treasure-seeking pirate in the throes of hungrily eating a feast surrounded by the ruins of an ancient temple, envisioned in the impassioned strokes of van gogh*
80. *Behold a wailing banshee in the throes of exploring a cave surrounded by the vibrant heart of a bustling market square, envisioned as a stunning 3d render*
81. *Amidst the ruins of an ancient temple, a playful panda is running a marathon, captured like a lively cartoon*
82. *Discover a wise old man, brewing a potion amidst the eerie halls of a haunted house, artfully rendered in the style of an oil painting*
83. *Within the realm of the dusty streets of an old western town at high noon, a fearsome dragon meticulously investigating a mysterious crime scene, each moment immortalized with photorealistic precision*
84. *Behold a genius scientist in the throes of playing the violin surrounded by the frosty peak of a snowy mountain, envisioned in the style of an oil painting*
85. *Discover a noble king, painting a masterpiece amidst the ruins of an ancient temple, artfully rendered with photorealistic precision*
86. *Discover an adventurous astronaut, conducting an experiment amidst the ruins of an ancient temple, artfully rendered like a lively cartoon*
87. *Discover a genius scientist, conducting an experiment amidst a bustling space station orbiting earth, artfully rendered as vibrant pixel art*
88. *Within the realm of the frosty peak of a snowy mountain, a fearsome dragon hunting for ghosts, each moment immortalized with photorealistic precision*
89. *Encounter a treasure-seeking pirate as they are engaged in gracefully dancing under the moonlight the dusty streets of an old western town at high noon, all depicted as a stunning 3d render*
90. *Within the realm of the front lines of an ancient battlefield, a mischievous fairy meticulously investigating a mysterious crime scene, each moment immortalized in the impassioned strokes of van gogh*
91. *Encounter a rogue robot as they are engaged in deciphering a map the sandy shores of a deserted island, all depicted with the soft touch of watercolor*
92. *Encounter a curious alien as they are engaged in deciphering a map the backdrop of an alien planet's red skies, all depicted like a lively cartoon*
93. *Discover a forest-dwelling nymph, brewing a potion amidst the lush canopy of a deep jungle, artfully rendered with photorealistic precision*
94. *Encounter a genius scientist as they are engaged in painting a masterpiece the sandy shores of a deserted island, all depicted like a lively cartoon*
95. *Within the realm of the vibrant heart of a bustling market square, a playful panda deciphering a map, each moment immortalized as vibrant pixel art*
96. *Within the realm of a sun-dappled forest, a mystical mermaid brewing a potion, each moment immortalized in the impassioned strokes of van gogh*
97. *Discover a curious alien, playing the violin amidst the secret chambers of an underground laboratory, artfully rendered through the abstract lens of picasso*
98. *Behold a brave girl in the throes of playing the violin surrounded by a mysterious crossroads in a mystical forest, envisioned through the abstract lens of picasso*
99. *Discover a mischievous fairy, painting a masterpiece amidst the secret chambers of an underground laboratory, artfully rendered with photorealistic precision*
100. *Within the realm of the frosty peak of a snowy mountain, a playful panda meticulously investigating a mysterious crime scene, each moment immortalized through the abstract lens of picasso*

Figure 10: Prompts used in AIGCBench.

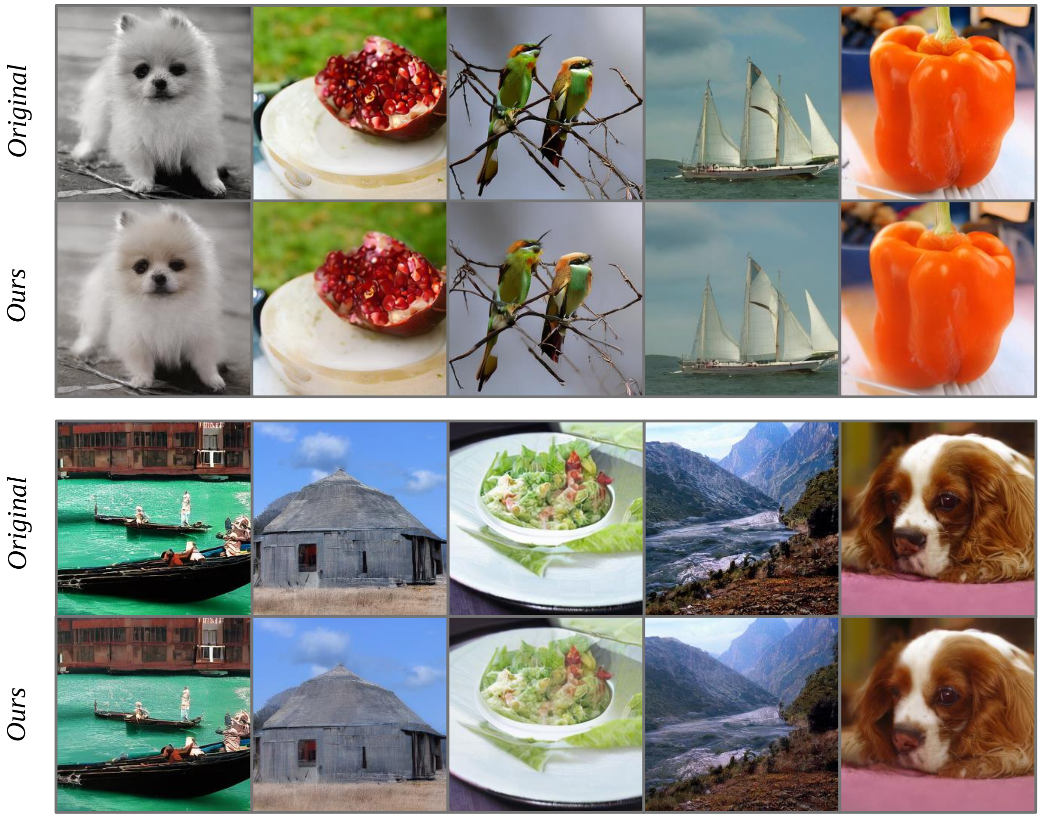


Figure 11: Qualitative results of text-to-image generation task using LDM-4 on ImageNet 256x256 benchmark.

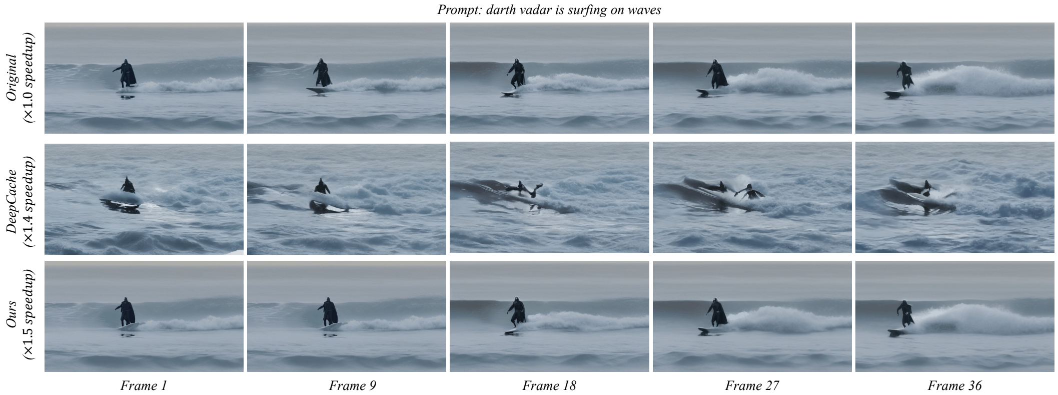


Figure 12: Qualitative results of text-to-video generation task using ModelScopeT2V on MSR-VIT benchmark.

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 motivations and contributions are well depicted and summarized in the abstract and introduction.

   Guidelines:

   - The answer NA 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 NA 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: Limitations are discussed in Appendix A.1.

   Guidelines:

   - The answer NA 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: [Yes]

   Justification: The theory assumptions and proofs are described in Section 3 and Appendix A.2.

   Guidelines:

   - The answer NA 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: The selected models and benchmarks are clearly and fully presented in Section 4.

   Guidelines:

   - The answer NA 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: The code is available at https://github.com/UniModal4Reasoning/AdaptiveDiffusion.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are carefully presented in Section 4.1.

   Guidelines:

   - The answer NA 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 χ^2 stats and p-value between the greedy searched paths and the proposed third-order estimated paths at different skipping targets.

   Guidelines:

   - The answer NA 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: The information of the employed compute resources is elaborated in the implementation details of Section 4.

   Guidelines:

   - The answer NA 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 make sure that the research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA 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: [NA]

    Justification: Our proposed method currently has no apparent societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA 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 have cited all papers that we used for experiments.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA 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: [NA]

Justification: This paper does not involve research with human subjects.

Guidelines:

- The answer NA 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.