

A LLM USAGE STATEMENT

The descriptive text for benchmarks and models presented in Appendix Sec. B was polished and refined using the LLM. The core factual information, including key metrics and technical specifications, was carefully curated and verified by the authors. The LLM was used solely as a tool to improve the clarity, coherence, and fluency of the writing. All content was thoroughly reviewed and approved by the authors to ensure accuracy.

B BENCHMARK AND MODEL

In this section, we provide an overview of the benchmarks, evaluation metrics, and the diffusion models in our main paper.

B.1 BENCHMARK

Pick-a-Pic. Pick-a-Pic (Kirstain et al., 2023) is an open dataset curated to capture user preference for T2I-synthesized images. Collected through an intuitive web application, it contains over 500,000 examples based on 35,000 unique prompts, providing a large-scale foundation for studying user preferences.

DrawBench. DrawBench (Saharia et al., 2022)¹ is a benchmark dataset introduced to enable comprehensive evaluation of T2I models. It consists of 200 meticulously designed prompts, categorized into 11 groups to assess model capabilities across various semantic dimensions. These dimensions include compositionality, numerical reasoning, spatial relationships, and the ability to interpret complex textual instructions. DrawBench is specifically designed to provide a multidimensional analysis of model performance, facilitating the identification of both strengths and weaknesses in T2I synthesis.

HPD v2. The human preference dataset v2 (HPD v2) (Wu et al., 2023) is an extensive dataset featuring clean and precise annotations. With 798,090 binary preference labels across 433,760 image pairs, it addresses the limitations of conventional evaluation metrics that fail to accurately reflect human preferences. Following the methodologies in (Wu et al., 2023; Shao et al., 2025), we employed four distinct subsets for our analysis: Animation, Concept-art, Painting, and Photo, each containing 800 prompts.

GenEval. GenEval (Ghosh et al., 2023) is an evaluation framework specifically designed to assess the compositional properties of synthesized images, such as object co-occurrence, spatial positioning, object count, and color. By leveraging state-of-the-art detection models, GenEval provides a robust evaluation of T2I generation tasks, ensuring strong alignment with human judgments. Additionally, the framework allows for the integration of other advanced vision models to validate specific attributes. The benchmark comprises 550 prompts, all of which are straightforward and easy to interpret.

T2I-Compbench. T2I-Compbench (Huang et al., 2023) is a comprehensive benchmark for evaluating open-world compositional T2I synthesis. It includes 6,000 compositional text prompts, systematically categorized into three primary groups: attribute binding, object relationships, and complex compositions. These groups are further divided into six subcategories: color binding, shape binding, texture binding, spatial relationships, non-spatial relationships, and intricate compositions.

ChronoMagic-Bench-150. Chronomagic-Bench-150, introduced in (Yuan et al., 2024) serves as a comprehensive benchmark for metamorphic evaluation of timelapse T2V synthesis. This benchmark includes 4 main categories of time-lapse videos: biological, human-created, meteorological, and physical, further divided into 75 subcategories. Each subcategory contains two challenging prompts, leading to a total of 150 prompts. We consider three distinct metrics in Chronomagic-Bench-150: UMT-FVD (\downarrow), UMTScore (\uparrow), GPT4o-MTScore (\uparrow) and MTScore (\uparrow).

¹<https://huggingface.co/datasets/shunk031/DrawBench>

FLUX-Kontext-Bench. FLUX-Kontext-Bench, introduced in (Labs et al., 2025), is a comprehensive benchmark for evaluating in-context image generation and editing models. It consists of 1026 image-prompt pairs derived from 108 base images. The benchmark spans five core task categories: local editing, global editing, text editing, style reference, and character reference. Designed to reflect real-world usage, FLUX-Kontext-Bench addresses limitations of prior synthetic or narrow-scope benchmarks and supports holistic evaluation of both single-turn quality and multi-turn consistency.

B.2 EVALUATION METRIC

PickScore. PickScore is a CLIP-based scoring model, developed using the Pick-a-Pic dataset, which captures user preferences for synthesized images. This metric demonstrates performance surpassing that of typical human benchmarks in predicting user preferences. By aligning effectively with human evaluations and leveraging the diverse range of prompts in the Pick-a-Pic dataset, PickScore offers a more relevant and insightful assessment of T2I models compared to traditional metrics like FID (Heusel et al., 2018) on datasets such as MS-COCO (Lin et al., 2015).

HPS v2. The human preference score version 2 (HPS v2) is an improved model to predict user preferences, created by fine-tuning the CLIP model (Radford et al., 2021) on the HPD v2. This refined metric is designed to align T2I generation outputs with human tastes by estimating the likelihood that a synthesized image will be preferred, thereby serving as a reliable benchmark for evaluating the performance of T2I models across diverse image distributions.

AES. The Aesthetic Score (AES) (Schuhmann) is a metric that evaluates the visual appeal of images. It is calculated using a model built on CLIP embeddings and enhanced with multilayer perceptron (MLP) layers. This metric provides a quantitative measure of the aesthetic quality of synthesized images, offering valuable insights into their alignment with human aesthetic standards.

ImageReward. ImageReward (Xu et al., 2023) is a specialized reward model designed to evaluate T2I synthesis based on human preferences. Trained on a large-scale dataset of human comparisons, the model effectively captures user inclinations by assessing multiple aspects of synthesized images, including their alignment with text prompts and their aesthetic quality. ImageReward has shown superior performance compared to traditional metrics such as the Inception Score (IS) (Salimans et al., 2016) and Fréchet Inception Distance (FID), establishing it as a highly promising tool for automated evaluation in T2I tasks.

B.3 FLOW MODELS

In the main paper, we totally use 3 flow-based T2I diffusion models, including FLUX-Dev (Labs, 2024), FLUX-Lite (Daniel Verdú, 2024), and StableDiffusion-3.5 (Esser et al., 2024), 1 flow-based T2V diffusion, Wan2.1-T2V-1.3B (Wan et al., 2025), and 1 flow-based T12I diffusion model, FLUX-Kontext (Labs et al., 2025).

FLUX-Dev. FLUX-Dev (Labs, 2024) is a family of T2I diffusion models built upon a transformer-based architecture, departing from the conventional U-Net design. Its core components include a dual text encoder system (CLIP and T5 (Chung et al., 2022)) for robust prompt understanding and a joint attention mechanism. This mechanism facilitates a bidirectional information flow between image and text representations within the transformer blocks, significantly enhancing prompt fidelity. The models are trained using a rectified flow formulation (Liu et al., 2022a), which enables high-quality image synthesis with fewer sampling steps compared to traditional diffusion models.

FLUX-Lite. FLUX-Lite is a lightweight and highly efficient version derived from the FLUX models, optimized for faster inference. This 8B parameter model achieving a 23% reduction in latency and a 7GB decrease in RAM usage. Its robustness is enhanced by a refined distillation process, trained on a diverse dataset and optimized for a broad range of guidance values (2.0-5.0) and step counts (20-32). The model’s efficiency stems from an architectural insight that its transformer blocks contribute heterogeneously. An analysis revealed that intermediate blocks possess a degree of redundancy, unlike the critical initial and final blocks. The property allows for effective distillation and optimization without significant degradation in generative performance.

Stable-Diffusion-3.5. StableDiffusion-3.5 marks a significant architectural shift in the StableDiffusion series to a Diffusion Transformer (DiT) (Peebles & Xie, 2023) model, aligning with the principles of rectified flow. As described by Esser et al. (2024), this model processes text and image modalities using separate transformer weights before fusing them with a joint attention mechanism. This approach enables a sophisticated, bidirectional interaction between the two modalities, leading to well performance in prompt adherence, typographic generation, and overall image coherence. Its design demonstrates predictable scaling, where improvements in training loss directly translate to superior synthesis quality.

Wan2.1. Wan2.1, introduced in (Wan et al., 2025), is an open-source video generation model developed by Alibaba, based on a Diffusion Transformer (DiT) architecture and flow matching framework. It supports multiple tasks including text-to-video (T2V) and image-to-video (I2V). The model is available in two versions: a 14B-parameter variant for high-quality 720p generation and a lightweight 1.3B variant suitable for consumer-grade GPUs. Due to the resource limits, in this paper, we utilize the Wan2.1-T2V-1.3B.

FLUX-Kontext. FLUX-Kontext, introduced in (Labs et al., 2025), is a unified flow matching model for in-context image generation and editing in latent space. It combines text and image conditioning through a simple sequence concatenation mechanism, enabling both local editing and generative tasks within a single architecture. The model excels in preserving character and object consistency across multiple iterative edits, supports high-resolution output at interactive speeds, and facilitates iterative workflows.

B.4 HYPERPARAMETER SETTINGS

For all the experiments in the main paper, the inference steps are default to 28, 28, 50, 50, and 50, corresponding to SD3.5, FLUX-Lite, FLUX-Dev, FLUX-Kontext and Wan-2.1-T2V-1.3B. For the standard guidance scales w are default to 4.5, 3.5, 3.5, 3.5, and 5.0, corresponding to SD3.5, FLUX-Lite, FLUX-Dev, FLUX-Kontext and Wan-2.1-T2V-1.3B.

For the hyperparameters, the interpolation weight $\beta_{high} = 0.7$, $\beta_{low} = 0.3$, and the merge ratio $\gamma = 0.5$ (for Wan-2.1-T2V-1.3B, $\gamma = 0.03$) across all the experiments. The amplifying weight $s_{high} = 3.5$, $s_{low} = 0$ for FLUX-Dev, and $s_{high} = 9$, $s_{low} = -1$ for FLUX-Lite, FLUX-Kontext, and SD3.5. The repeat time $\alpha = 1$ for SD3.5, FLUX-Dev, FLUX-Kontext, and Wan-2.1-T2V-1.3B, and $\alpha = 2$ for FLUX-Lite. For experiments in SD3.5, FLUX-Lite and FLUX-Dev, we execute RF-Sampling operations through all the inference steps, for FLUX-Kontext and Wan-2.1-T2V-1.3B, due to the time budgets, we only perform RF-Sampling operations in the first two steps.

C ADDITIONAL ANALYSIS

To further understand the effect of the parameter in our method, we conduct additional parameter analysis experiments as shown follows:

Experiments on Large-scale Dataset. To further validate the effectiveness of RF-Sampling, we conduct experiments on popular, large-scale benchmarks like GenEval (Ghosh et al., 2023) and T2I-CompBench (Huang et al., 2023) across 3 different flow models, shown in Tab. 6 and Tab. 7. The large-scale experiments demonstrate the generalizability and robustness of our proposed methods.

The form of Null prompt. Since the null-text representation can be either implemented as zero padding 0 or as an explicit null token \emptyset , we conduct experiments under different values of the parameter α . The results are reported in Tab. 4, which demonstrate that using a null-text representation provides our method with stronger unconditional information.

Effect of parameter α . We evaluated the impact of the parameter α on inference performance, with results shown in Tab. 4. Considering that α also affects inference speed, we set $\alpha = 2$ for FLUX Lite and $\alpha = 1$ for FLUX Dev in our final configuration.

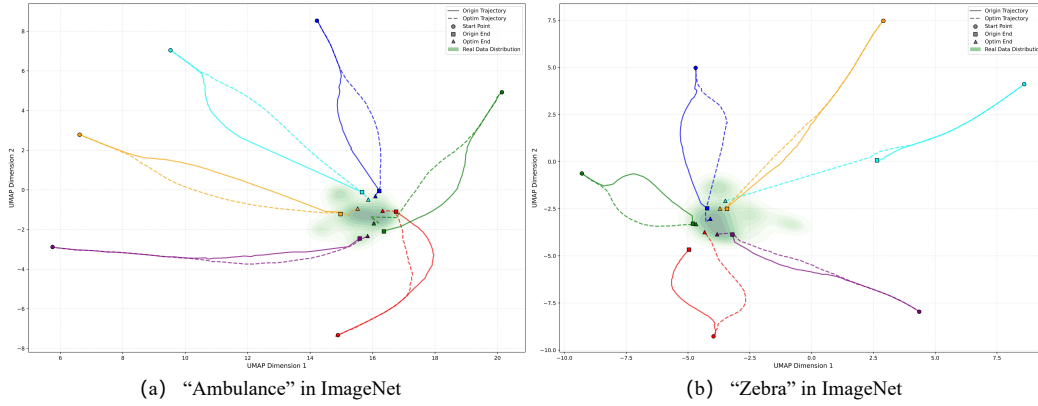


Figure 11: Visualizations of the sampling trajectories of RF-Sampling and the standard method. we randomly select two ImageNet classes (Russakovsky et al., 2015) (Ambulance and Zebra) and visualize their respective data distributions. For each class, we randomly sample 6 Gaussian noises and process them through both standard diffusion sampling and RF-Sampling methods using the prompt format "a photo of class in ImageNet." The results reveal that RF-Sampling trajectories consistently demonstrate stronger convergence towards the real data distribution compared to standard method trajectories.

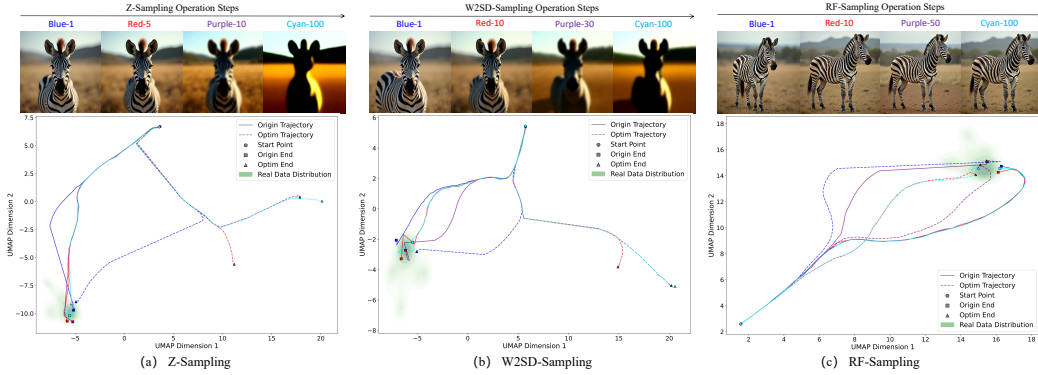


Figure 12: Visualizations of synthesized images of Z-Sampling and W2SD-Sampling under different numbers of sampling operations, and their corresponding sampling trajectories. The noises in each method are sampled from the fixed random seed. With the increase of sampling operations, Z-Sampling and W2SD-Sampling exhibit rapid degradation from clear zebra images to abstract, blurred silhouettes, while RF-Sampling preserves remarkable visual integrity and semantic coherence across all sampling steps. The trajectory analysis reveals that RF-Sampling produces highly optimized, stable latent space trajectories with tight clustering within the real data distribution, whereas the other methods show more divergent and less stable paths.

Effect of merge ratio γ . To determine the optimal merge ratio γ for our method, we conduct a quantitative study on the **Pick-a-Pic** dataset shown in Fig. 15. We systematically vary the value of γ and evaluate the quality of the synthesized images using four distinct metrics. Across all four metrics, the highest scores are usually achieved when γ was set to 0.5. This suggests that an equal balance in the merge operation is critical for producing the highest-quality images. Lower or higher values of γ led to a noticeable degradation in performance, indicating an imbalanced fusion.

Effect of Guidance Scale. Traditional T2I diffusion models can enhance the quality of synthesized images by increasing the inference steps and guidance scale. In RF-Sampling, we adopt **High-Weight Denoising** \rightarrow **Low-Weight Inversion**, which implicitly increases the inference steps and guidance scales. To validate our method, We further conduct an ablation study on the standard guidance scale w and inference steps as shown in Fig 16, Fig. 2, respectively. As w increases, we



Figure 13: Visualization of synthesized images with different s scales. $s_{high} > s_{low}$ serves as a necessary condition for achieving superior image synthesis quality. When $s_{high} - s_{low} < 0$, RF-Sampling shows minimal advantage over the standard method with poor text generation accuracy and blurred details. However, as $s_{high} - s_{low}$ increases to positive values, RF-Sampling exhibits dramatic improvements in text rendering precision, visual detail clarity, and overall image coherence, ultimately generating high-fidelity outputs that significantly outperform Standard approaches. The results indicate that the parameter relationship $s_{high} > s_{low}$ acts as a crucial control mechanism that enables RF-Sampling to leverage its full potential for complex text-to-image generation tasks.

observe a clear degradation in the quality of the synthesized images. In addition, with increasing inference steps, the performance gains of RF-Sampling. This finding confirms that the performance improvement of RF-Sampling does not originate from simply amplifying the guidance through a larger weight s , but rather from the reflective mechanism itself.

Distribution Analysis. To explore the RF-Sampling trajectories, we select two ImageNet classes (Russakovsky et al., 2015) (Ambulance and Zebra) and visualize their respective data distributions as green shaded regions in the UMAP space. For each class, we randomly sample 6 Gaussian noises and process them through both standard diffusion sampling and RF-Sampling methods on FLUX-Dev using the prompt format "a photo of class in ImageNet." The results shown in Fig. 11 reveal that RF-Sampling trajectories consistently demonstrate stronger convergence towards the real data distribution compared to standard method trajectories, as evidenced by the optimized endpoints (triangles) being more tightly clustered within or closer to the dense real data regions than their corresponding standard endpoints (squares). This convergence pattern indicates that RF-Sampling successfully refines the generation process by moving latent representations closer to the manifold of real images, thereby enhancing the fidelity and realism of generated samples while maintaining the semantic coherence of the target ImageNet classes.

Besides, we also visualize the synthesized images of Z-Sampling and W2SD-Sampling across different numbers of sampling operations and their corresponding sampling trajectories, shown in Fig. 12. The results demonstrate that RF-Sampling significantly outperforms both Z-Sampling and W2SD-Sampling in maintaining image fidelity and semantic consistency throughout the sampling process. While Z-Sampling and W2SD-Sampling exhibit rapid degradation from clear zebra images to abstract, blurred silhouettes, RF-Sampling preserves remarkable visual integrity and semantic coherence across all sampling steps. The trajectory analysis reveals that RF-Sampling produces highly optimized, stable latent space trajectories with tight clustering within the real data distribution, whereas the other methods show more divergent and less stable paths. These findings indicate that RF-Sampling offers superior guidance mechanisms for navigating the latent space towards the target data manifold, resulting in more realistic and semantically consistent image generations that maintain high fidelity throughout the denoising process.



Figure 14: Visualization of synthesized images with different β scales. By applying the interpolation weight β , the model can synthesize higher-quality, more detailed, and visually appealing images that better align with user expectations for complex prompts, especially when $\beta_{high} > \beta_{low}$.

Table 4: We conduct ablation studies of the value of α and the form of c_{uncond} on Pick-a-Pic dataset. It is noticed that as α increases, the quality of synthesized images improves, albeit at an inevitable cost to computational time. The form of c_{uncond} may be either the null prompt embedding \emptyset or the zero-padded prompt embedding $\mathbf{0}$. In relation to the form of c_{uncond} , the null prompt embedding \emptyset yields superior results, as it carries a greater amount of unconditional semantic information.

Model	Method	PickScore(\uparrow)	ImageReward(\uparrow)	AES(\uparrow)	HPSv2(\uparrow)	
FLUX-Lite (28 steps)	$\alpha = 1$	0	21.95	88.58	6.4346	29.90
		\emptyset	21.95	87.64	6.4576	30.05
	$\alpha = 2$	0	21.88	91.86	6.4667	29.61
		\emptyset	22.05	99.21	6.5379	31.16
FLUX-Dev (50 steps)	$\alpha = 1$	0	22.11	98.82	6.2566	30.00
		\emptyset	22.19	100.90	6.3113	31.06
	$\alpha = 2$	0	22.08	101.19	6.3168	30.74
		\emptyset	22.14	100.52	6.3342	30.88

Table 5: We extend our method to the video generation task. Due to the computational budget, we utilize Wan2.1-T2V-1.3B (Wan et al., 2025). The results on ChronoMagic-Bench-150 (Yuan et al., 2024) across 4 metrics show the promising scalability of our method to the video generation task.

Method	UMT-FVD (\downarrow)	UMTScore (\uparrow)	GPT4o-MTScore (\uparrow)	MTScore (\uparrow)
Standard	264.84	2.7053	3.4797	0.41497
RF-Sampling	229.49	2.9095	3.5302	0.43671

Table 6: We evaluate the effectiveness of RF-Sampling on T2I-CompBench (Huang et al., 2023) across 3 diffusion models. The results validate the effectiveness and generalizability of our method.

Model	Method	Attribute Binding			Object Relationship				Complex(↑)	Overall(↑)
		Color(↑)	Shape(↑)	Texture(↑)	2D-Spatial(↑)	3D-Spatial(↑)	Non-Spatial(↑)	numeracy(↑)		
SD3.5 (28 steps)	Standard	0.7511	0.5709	0.7119	0.2927	0.3751	0.3166	0.6078	0.3846	0.5013
	RF-Sampling	0.7817	0.5885	0.7241	0.2864	0.3974	0.3174	0.6121	0.3844	0.5119
FLUX-Lite (28 steps)	Standard	0.7030	0.4154	0.4887	0.2258	0.3710	0.3030	0.5564	0.3365	0.4249
	RF-Sampling	0.7613	0.4725	0.5970	0.2420	0.4042	0.3070	0.6090	0.3649	0.4698
FLUX-Dev (50 steps)	Standard	0.7535	0.5018	0.6167	0.2783	0.3866	0.3078	0.6052	0.3706	0.4775
	RF-Sampling	0.7761	0.5323	0.6422	0.2687	0.3943	0.3080	0.6082	0.3733	0.4887

Table 7: We evaluate the effectiveness of RF-Sampling on GenEval (Ghosh et al., 2023) across 3 diffusion models. The results show the superiority over the standard.

Model	Method	Single(↑)	Two(↑)	Counting(↑)	Colors(↑)	Positions(↑)	Color Attribution(↑)	Overall(↑)
SD3.5 (28 steps)	Standard	0.97	0.91	0.75	0.85	0.21	0.53	0.70
	RF-Sampling	0.99	0.91	0.72	0.89	0.19	0.54	0.71
FLUX-Lite (28 steps)	Standard	0.90	0.57	0.52	0.71	0.11	0.36	0.53
	RF-Sampling	0.93	0.62	0.59	0.73	0.18	0.42	0.58
FLUX-Dev (50 steps)	Standard	0.99	0.80	0.78	0.77	0.23	0.50	0.68
	RF-Sampling	0.99	0.82	0.76	0.80	0.25	0.50	0.69

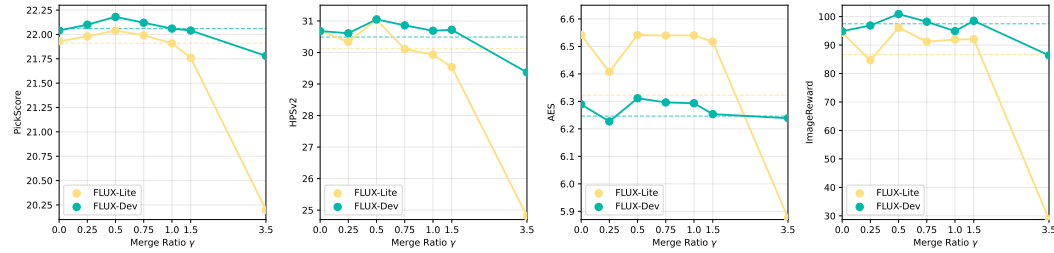


Figure 15: We explore the influence of merge ratio γ on Pick-a-Pic dataset. The results across 4 metrics reveal that $\gamma = 0.5$ is a better choice, where the synthesized images are the best. The dotted lines represents the performance of the standard method. This indicates that within a certain range of values, RF-Sampling perform better than the standard one, demonstrating the robustness of it.

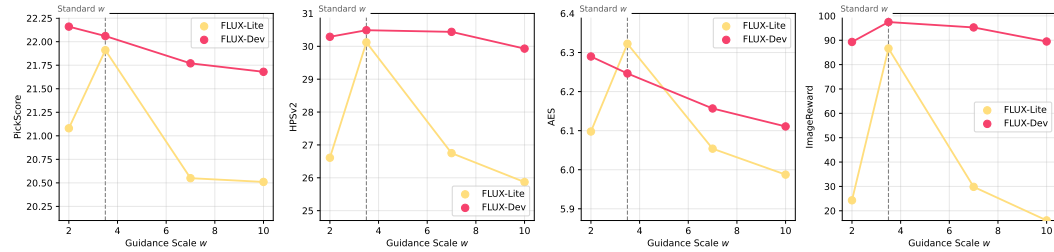


Figure 16: Ablation study of standard guidance scale w . With the increase of the standard guidance scale w , the quality of the synthesized images degrades a lot. The results guarantee that the performance gain of RF-Sampling is not introduced by the increase of the amplifying weight s .

D MORE VISUALIZATIONS

We provide more visualizations of synthesized images on FLUX-Dev and FLUX-Lite, across HPD v2, Pick-a-Pic, DrawBench and GenEval datasets as shown in Fig. 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, and 41.

Algorithm 1: Reflective Flow Sampling**Require:** Latent feature x_t , time steps $1, \dots, T$, merge ratio γ , and forward steps α .

```

1: for  $t = T, \dots, 1$  do
2:   // Stage 1: High-Weight Denoising ( $\alpha$  steps forward)
3:   Let  $\Delta t$  be the step size for each interval.
4:    $x_{fwd} \leftarrow x_t$ 
5:   for  $i = 1$  to  $\alpha$  do
6:      $c_{mix_{high}} = \beta_{high} \cdot c_{text} + (1 - \beta_{high}) \cdot c_{uncond}$ 
7:      $c' = c_{text} + s_{high} \cdot c_{mix_{high}}$ 
8:      $x_{fwd} \leftarrow x_{fwd} + v_{\theta}(x_{fwd}, t - (i - 1), c')\Delta t$ 
9:   end for
10:   $x_{t-\alpha} \leftarrow x_{fwd}$ 
11:  // Stage 2: Low-Weight Inversion ( $\alpha$  steps inversion)
12:   $x_{inv} \leftarrow x_{t-\alpha}$ 
13:  for  $i = 1$  to  $\alpha$  do
14:     $c_{mix_{low}} = \beta_{low} \cdot c_{text} + (1 - \beta_{low}) \cdot c_{uncond}$ 
15:     $c'' = c_{text} + s_{low} \cdot c_{mix_{low}}$ 
16:     $x_{inv} \leftarrow x_{inv} - v_{\theta}(x_{inv}, t - \alpha + (i - 1), c'')\Delta t$ 
17:  end for
18:   $x'_t \leftarrow x_{inv}$ 
19:  // Stage 3: Normal-Weight Denoising (1 step forward)
20:   $x''_t \leftarrow x_t + \gamma(x_t - x'_t)$ 
21:  // Standard Denoising
22:   $x''_{t-1} \leftarrow x''_t + v_{\theta}(x''_t, t, c)\Delta t$ 
23:   $x_{t-1} \leftarrow x''_{t-1}$ 
24: end for
25: return  $x_0$ 

```

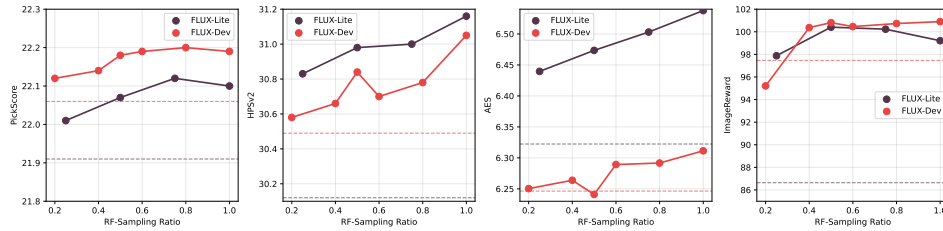


Figure 17: Robustness to the RF-Sampling steps. The horizontal axis shows the ratio of RF-Sampling operations during the whole inference steps. As the ratio increases, generation quality improves, indicating effective semantic information gain throughout the whole path. The dotted lines represents the performance of the standard method. This indicates that within a certain range of values, RF-Sampling perform better than the standard one, demonstrating the robustness of it.

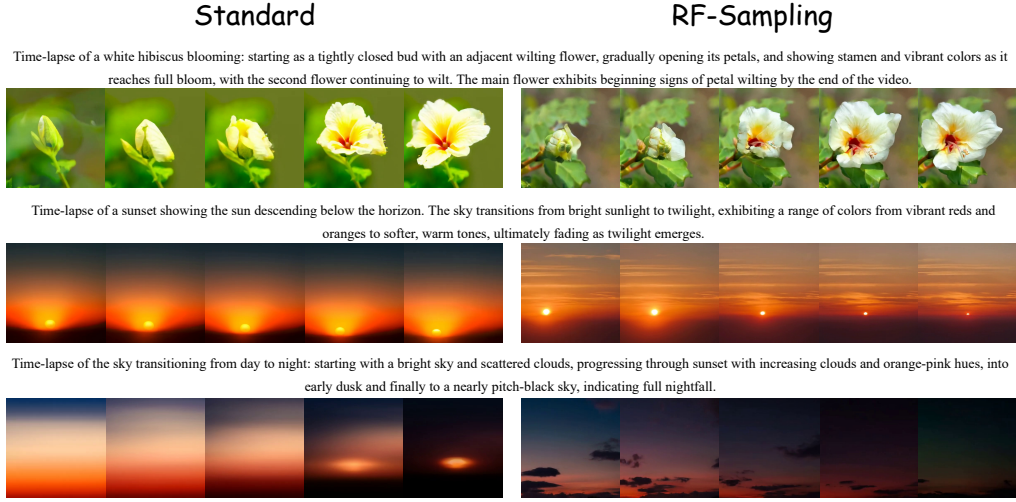


Figure 18: We directly extend our proposed method to video generation task on Wan2.1-T2V-1.3B. The visualizations show the superiority of our proposed method compared with standard sampling.

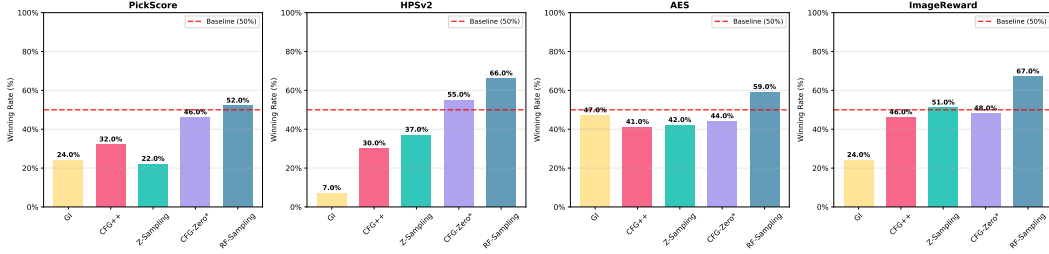


Figure 19: The winning rate of RF-Sampling over other methods on SD3.5 on Pick-a-Pic dataset. The standard sampling (baseline) winning rate defaults to 50%

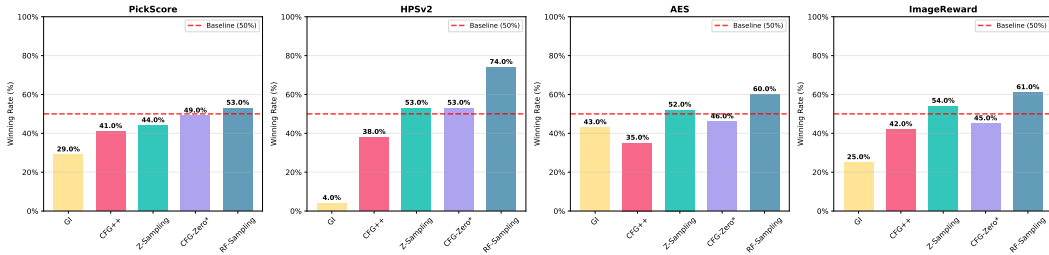


Figure 20: The winning rate of RF-Sampling over other methods on SD3.5 on DrawBench dataset. The standard sampling (baseline) winning rate defaults to 50%.

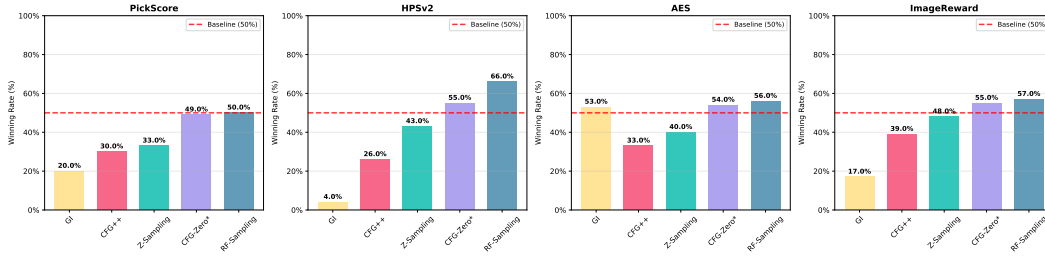


Figure 21: The winning rate of RF-Sampling over other methods on SD3.5 on the animation subset of HPD v2 dataset. The standard sampling (baseline) winning rate defaults to 50%.

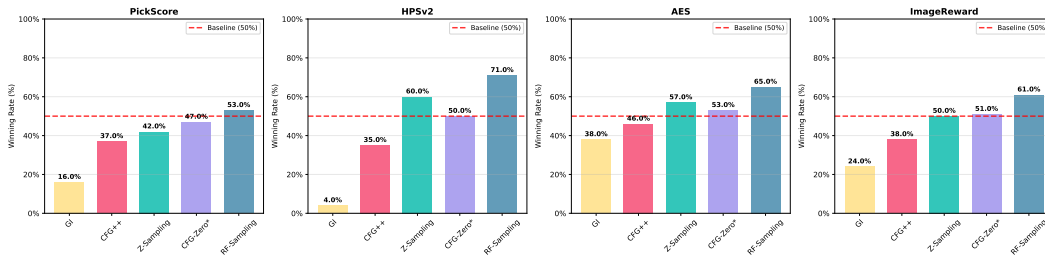


Figure 22: The winning rate of RF-Sampling over other methods on SD3.5 on the photo subset of HPD v2 dataset. The standard sampling (baseline) winning rate defaults to 50%.

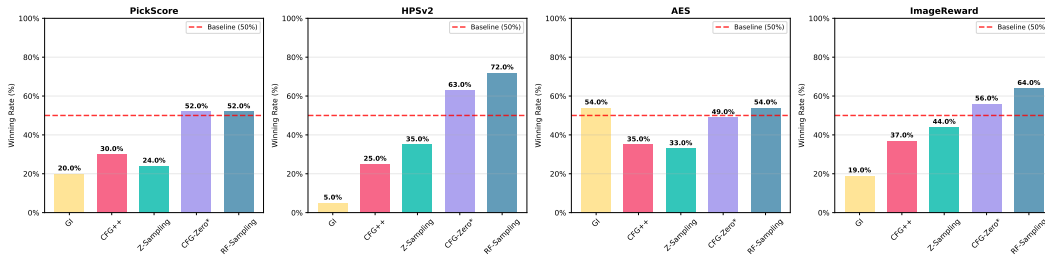


Figure 23: The winning rate of RF-Sampling over other methods on SD3.5 on the concept-art subset of HPD v2 dataset. The standard sampling (baseline) winning rate defaults to 50%.

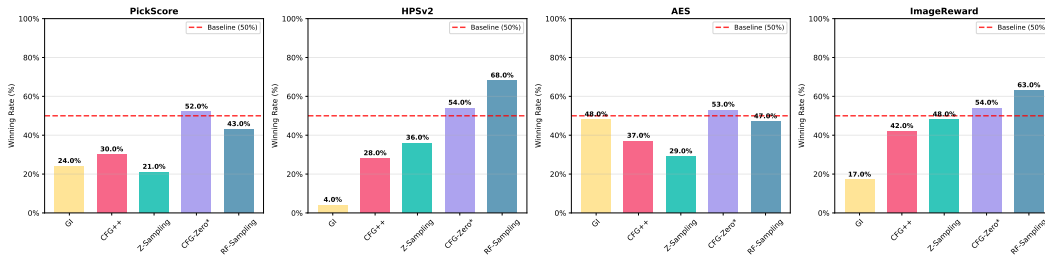


Figure 24: The winning rate of RF-Sampling over other methods on SD3.5 on the painting subset of HPD v2 dataset. The standard sampling (baseline) winning rate defaults to 50%..

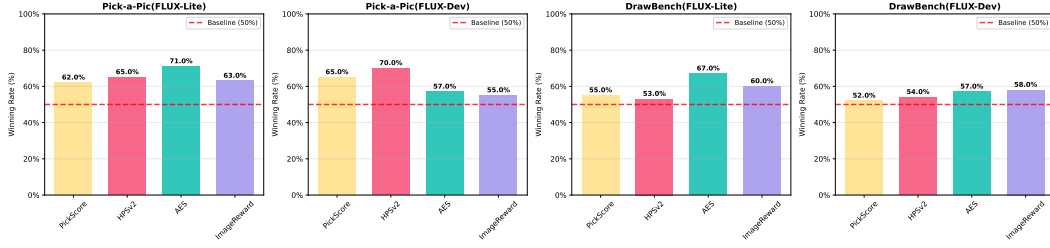


Figure 25: The winning rate of RF-Sampling over the standard one on FLUX-Lite and FLUX-Dev on Pick-a-Pic and DrawBench datasets. The standard sampling (baseline) winning rate defaults to 50%.

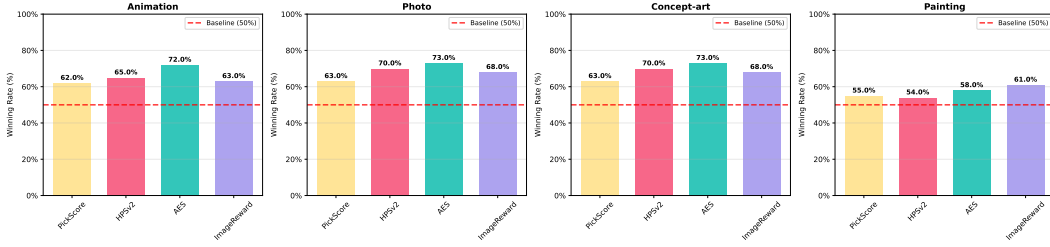


Figure 26: The winning rate of RF-Sampling over the standard one on FLUX-Lite on the 4 subsets of HPD v2 datasets. The standard sampling (baseline) winning rate defaults to 50%.

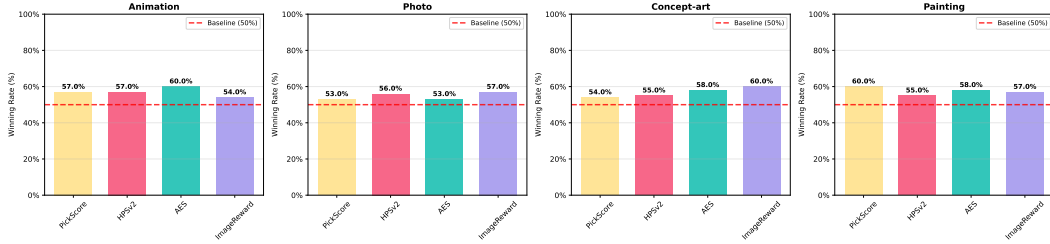


Figure 27: The winning rate of RF-Sampling over the standard one on FLUX-Dev on the 4 subsets of HPD v2 datasets. The standard sampling (baseline) winning rate defaults to 50%.

Table 8: To demonstrate the robustness of our method, we conducted repeated experiments on the Pick-a-Pic using FLUX-Lite with different random seeds. The results show that our approach consistently outperformed the standard method across varying random seeds, highlighting the robustness of RF-Sampling.

	Method	PickScore(↑)	HPSv2(↑)	AES(↑)	ImageReward(↑)
Round 1	Standard	21.91	30.12	6.3224	86.84
	RF-Sampling	22.05	31.16	6.5379	99.21
Round 2	Standard	21.95	30.33	6.3473	93.73
	RF-Sampling	22.04	30.82	6.5231	100.81
Round 3	Standard	21.94	30.20	6.3608	99.42
	RF-Sampling	21.99	30.63	6.5133	103.45
Round 4	Standard	21.96	30.23	6.3365	96.22
	RF-Sampling	22.02	30.83	6.5243	109.37
Average	Standard	21.94 ± 0.02	30.22 ± 0.08	6.3418 ± 0.0163	94.00 ± 5.43
	RF-Sampling	22.03 ± 0.03	30.86 ± 0.22	6.5247 ± 0.0101	103.21 ± 4.46

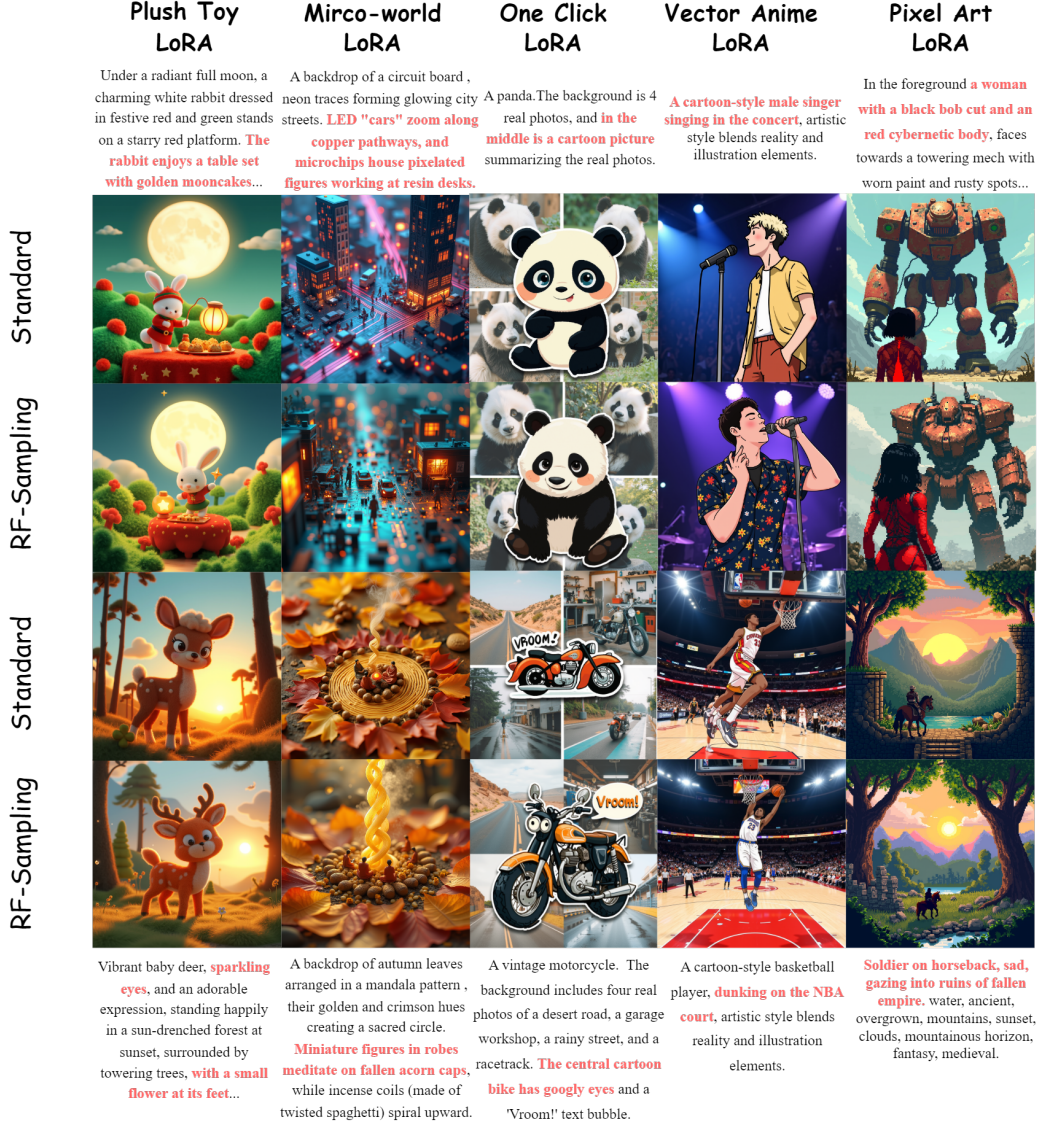


Figure 28: We combine our proposed methods with existing LoRAs in FLUX community. Our RF-Sampling can be directly applied to the corresponding downstream tasks, validating the generalizability of our method.

Table 9: Ablation on reflection component. We replace the full reflection step with a simple linear interpolation between embeddings (using Eqn. 2) under two distinct mixing weights. The results show that both linear variants fail to improve over the standard baseline, performing identically across all metrics. This demonstrates that the model-driven reflection is essential, as simpler heuristics cannot achieve the performance gains of our full RF-Sampling.

Method	PickScore(↑)	HPSv2(↑)	AES(↑)	ImageReward(↑)
Standard	21.99	29.32	5.9435	85.13
High Embedding Mix ($s = 9, \beta = 0.7$)	21.99	29.32	5.9435	85.13
Low Embedding Mix ($s = -1, \beta = 0.3$)	21.99	29.32	5.9435	85.13
RF-Sampling	21.99	29.90	5.9981	101.50

Table 10: Comparison of our RF-Sampling with Best-of-N method. RF-Sampling achieves a better trade-off between performance and efficiency: it outperforms standard sampling in all metrics and is competitive with Best-of-N methods. While Best-of-5 achieves the highest performance, it requires more than double the time per image compared to RF-Sampling. RF-Sampling outperforms Best-of-3 in PickScore, AES and ImageReward with approximately 1.5 times faster. These results demonstrate the effectiveness of our method in achieving high performance with reduced computational cost.

Method	PickScore(\uparrow)	HPSv2(\uparrow)	AES(\uparrow)	ImageReward(\uparrow)	s/img(\downarrow)
Standard (28 steps)	21.99	29.32	5.9435	85.13	29.93
Standard ($28 \times 3 = 84$ steps)	21.96	29.60	5.9109	89.87	67.06
Best-of-5	22.21	30.58	5.9849	106.69	154.17
Best-of-3	21.94	30.14	5.9642	100.40	97.63
RF-Sampling	21.99	29.90	5.9981	101.50	65.04

Table 11: Comparison under equivalent computational budget. To demonstrate the effectiveness of our method, we compare RF-Sampling against baselines using 84 steps (28×3), matching the total inference time. The results show that RF-Sampling almost outperforms all baseline methods across different metrics while maintaining comparable time per image, demonstrating its effectiveness under a fair computational setting.

Method	PickScore(\uparrow)	HPSv2(\uparrow)	AES(\uparrow)	ImageReward(\uparrow)	s/img (\downarrow)
Standard (28 steps)	21.99	29.32	5.9435	85.13	29.93
GI (28 steps)	21.19	24.63	5.9534	28.94	31.33
CFG++ (28 steps)	21.79	28.50	5.8821	85.17	32.46
CFG-Zero* (28 steps)	21.88	29.37	5.9536	86.78	28.91
Standard ($28 \times 3 = 84$ steps)	21.96	29.60	5.9109	89.87	67.06
GI ($28 \times 3 = 84$ steps)	21.25	25.27	5.9335	28.16	67.04
Z-Sampling ($28 \times 3 = 84$ steps)	21.73	28.84	5.9091	89.03	65.00
CFG++ ($28 \times 3 = 84$ steps)	20.98	27.02	5.6144	64.73	68.07
CFG-Zero* ($28 \times 3 = 84$ steps)	22.01	29.48	5.8949	97.22	65.47
RF-Sampling	21.99	29.90	5.9981	101.50	65.04

Table 12: Quantitative comparisons with Ma et al. (2025a) on DrawBench. RF-Sampling requires only 150 NFEs, far fewer than the baseline methods (2880 NFEs), yet still achieves the top results in both ImageReward and AES, demonstrating the dual advantages of our method in both efficiency and effectiveness.

Metric	Method				
	Standard	Aesthetic + Random	+ ZO-2	+ Path-2	RF-Sampling
NFEs	50	2880	2880	2880	$50 \times 3 = 150$
ImageReward	99.73	101.21	98.42	97.13	106.21

Metric	Method				
	Standard	CLIPScore + Random	+ ZO-2	+ Path-2	RF-Sampling
NFEs	50	2880	2880	2880	$50 \times 3 = 150$
AES	6.1459	6.0323	6.0512	6.0452	6.1866

Metric	Method				
	Standard	ImageReward + Random	+ ZO-2	+ Path-2	RF-Sampling
NFEs	50	2880	2880	2880	$50 \times 3 = 150$
AES	6.1459	6.1459	6.1265	6.0945	6.1966

Table 13: Quantitative comparisons with Ma et al. (2025a) on T2I-CompBench. RF-Sampling requires only 150 NFEs, far fewer than the baseline methods (1920 NFEs), yet almost achieves the top results across different dimensions, demonstrating the dual advantages of our method in both efficiency and effectiveness.

Method	Color(\uparrow)	Shape(\uparrow)	Texture(\uparrow)	Spatial(\uparrow)	Numeracy(\uparrow)	Complex(\uparrow)	Overall(\uparrow)
Standard	0.7535	0.5018	0.6167	0.2783	0.6052	0.3706	0.5210
Aesthetic + Random (1920 NFEs)	0.7518	0.5219	0.5926	0.2893	0.6059	0.3572	0.5199
RF-Sampling ($50 \times 3 = 150$ NFEs)	0.7761	0.5323	0.6422	0.2687	0.6082	0.3733	0.5335

Table 14: Detailed breakdown of Fig. 2, including step counts (NFEs) and wall time. As shown in the table below, RF-Sampling outperforms standard sampling with the same time consumption and significantly enhances the performance of FLUX-Lite and FLUX-Dev. With the increase of inference time, RF-Sampling consistently performs well, validating the scalability of our method.

Model	Method	NFEs	HPSv2(↑)	AES(↑)	s/img (↓)
FLUX-Lite	Standard	28	30.12	6.3224	34.63
		50	30.39	6.3045	46.60
		75	30.46	6.2864	60.61
	RF-Sampling ($\alpha = 2$)	$7 \times 5 + 21 = 56$	30.84	6.4397	49.63
		$14 \times 5 + 14 = 84$	30.98	6.4736	64.57
		$21 \times 5 + 7 = 112$	31.04	6.5032	76.84
$28 \times 5 = 140$		31.16	6.5379	95.26	
FLUX-Dev	Standard	50	30.49	6.2464	59.09
		75	30.54	6.2170	75.85
		100	30.60	6.1869	91.48
	RF-Sampling ($\alpha = 1$)	$10 \times 3 + 40 = 70$	30.58	6.2505	71.87
		$20 \times 3 + 30 = 90$	30.66	6.2639	86.07
		$30 \times 3 + 20 = 110$	30.70	6.2893	100.03
		$40 \times 3 + 10 = 130$	30.79	6.2917	114.30
$50 \times 3 = 150$	31.06	6.3113	127.95		

Table 15: Comparison of FID and IS between standard sampling and RF-Sampling on ImageNet-1K. We use FLUX-Lite with inference steps 28, combining the nunchaku (Li* et al., 2025) sampling acceleration framework, to generate 5,000 samples (5 images per class). RF-Sampling achieves a lower FID and a higher IS than the standard one, demonstrating its ability to better align with the real data distribution while maintaining high-quality and diverse image generation.

Method	FID (\downarrow)	IS (\uparrow)
Standard	35.08	150.07
RF-Sampling	33.12	155.21



Figure 30: Synthesized images of FLUX-Lite on anime subset of HPD v2.



Figure 31: Synthesized images of FLUX-Lite on photography subset of HPD v2.



Figure 32: Synthesized images of FLUX-Lite on painting subset of HPD v2.



Figure 33: Synthesized images of FLUX-Lite on concept-art subset of HPD v2.



Figure 34: Synthesized images of FLUX-Lite on GenEval.



Figure 35: Synthesized images of FLUX-Lite on Pick-a-Pic and DrawBench.

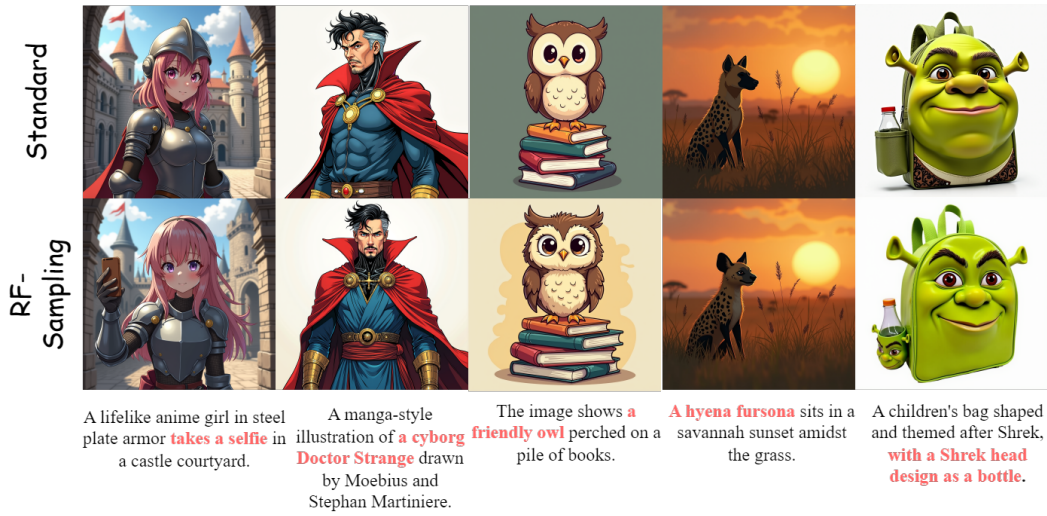


Figure 36: Synthesized images of FLUX-Dev on anime subset of HPD v2.



Figure 37: Synthesized images of FLUX-Dev on photography subset of HPD v2.



Figure 38: Synthesized images of FLUX-Dev on painting subset of HPD v2.



Figure 39: Synthesized images of FLUX-Dev on concept-art subset of HPD v2.



Figure 40: Synthesized images of FLUX-Dev on GenEval.

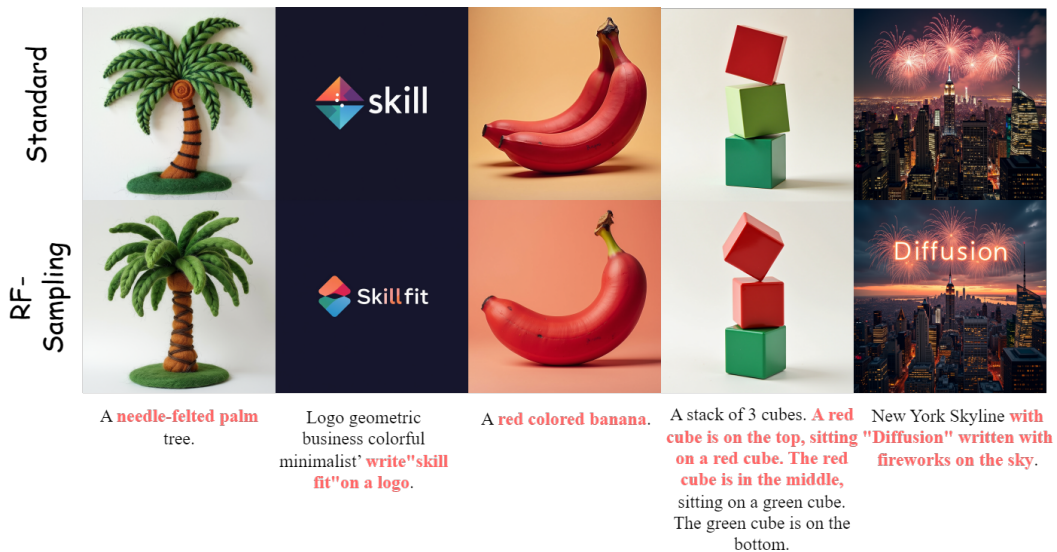


Figure 41: Synthesized images of FLUX-Dev on Pick-a-Pic and DrawBench.



Figure 42: Visual results of FLUX-Lite with guidance scale $w = 1$. The generated images remain semantically aligned with the input text prompts, demonstrating that the model’s output is still conditionally generated even at the minimum guidance scale. This empirically verifies that CFG-distilled models like FLUX do not possess a true unconditional generation mode, and setting $w = 1$ does not produce unconditional outputs.