

SKETCH FIRST, SCALE FAST: ON EFFICIENT INFERENCE-TIME SCALING FOR VI- SUAL GENERATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Diffusion models have exhibited exceptional capabilities in image generation tasks. However, due to their inherent stochasticity, the quality of the generated images varies across different settings and may not always be high-fidelity and accurately align with user requirements. To address this challenge, recent works begin to focus on enhancing human preference alignment and improving overall image fidelity during the inference process with inference time scaling, which involves generating multiple candidates through repeated sampling and selecting with predefined metrics. Although effective, it will introduce considerable computational extra costs with the redundant sampling steps. To overcome these limitations, we propose a novel inference time scaling framework SF2 that enables estimators to make decisions in the early step via Co-Fusion pipeline, significantly improving the whole process while maintaining the quality of the selected images. In addition, for the continue generation part, we propose vision reflection mechanism to further align and correct the images with user requirements. Numerous experiments demonstrate that our proposed method can achieve comparable performance while $2.2 \times$ and $2.0 \times$ accelerating the whole inference time scaling process on both Stable-Diffusion-3.5-Large and FLUX.1-dev models.

1 INTRODUCTION

In recent years, diffusion models have demonstrated impressive capabilities across various generative tasks and have been widely adopted in diverse modalities, *e.g.*, image Ho et al. (2020); Song et al. (2020); Rombach et al. (2022); Esser et al. (2024); Labs (2024); Podell et al. (2023), video Ho et al. (2022); Blattmann et al. (2023); Kong et al. (2024); Jin et al. (2024); Chen et al. (2024b); Zheng et al. (2024), 3D Luo & Hu (2021); Poole et al. (2022); Deng et al. (2023), and text Gong et al. (2022); Chen et al. (2023). However, diffusion models inherently rely on stochastic processes, specifically involving randomly sampled initial noise, resulting in variability and inconsistency in output fidelity. Recent studies Zhou et al. (2024); Ahn et al. (2024); Qi et al. (2024); Bai et al. (2024) further reveal that the effectiveness of these initial noise vectors varies markedly, often causing generated images to deviate from user requirements and do not consistently achieve high-quality outcomes.

In this case, a straightforward way to address this issue is to scale the training process Mei et al. (2024); Li et al. (2024); Liang et al. (2024); Podell et al. (2023); Chen et al. (2025) from aspects of training data volume, model size, and training duration, to enhance the generative capacity of models, and thereby reducing the probability of generating the undesirable output. However, training time scaling methods typically incur prohibitive training costs, and usually come with the issues including costly data acquisition and heavy computational resources demands. Consequently, recent research has turned to the inference time scaling method Ma et al. (2025); Xie et al. (2025); Kim et al. (2025), with the aim of obtaining at least a high-fidelity image with only a trivial increase in inference overhead compared to training costs. To retrieve outputs that align with user specifications, a simple, yet effective inference time scaling method performs a repeated sampling process to form a set of candidate images, from which the highest quality result is selected according to predefined evaluation metrics Ma et al. (2025); Xie et al. (2025). However, this method is inefficient and time-consuming, as it requires repeated complete passes through the entire image generation pipeline to obtain a sufficient large candidate set. Given a fixed limited sampling budget, a slower sampling

process restricts the number of candidate images produced, narrowing the selection range, and thus limiting the achievable image quality. Moreover, selection based on a one-time evaluation over a finite sample pool typically yields merely identifying the least flawed result rather than a truly high-fidelity image.

To address the above challenges, we propose a novel inference time scaling framework **SF2** aiming to efficiently generate a larger pool of candidate samples within the limited inference budget. Meanwhile, we further propose a reflection mechanism capable of rectifying flaws and refining details of the selected samples during the continue generation process. Specifically, we propose Co-Fusion pipeline that utilizes the blur and low-quality images at the early stage with a fast detail refiner to ensure the visual maintenance of the final output, enabling the estimators to make decisions at the early stage. Also, after performing the selection in the early stage, we leverage the powerful vision language model, *e.g.*, Qwen2.5-VL Bai et al. (2025), to provide the feedback, and inject the feedback through the continue image generation process to rectify defects and enhance fine details of the selected images. To validate the effectiveness of our proposed method, we conduct numerous experiments on current state-of-the-art diffusion models Stable-Diffusion-3.5-Large Esser et al. (2024) and FLUX.1-dev Labs (2024). The experimental results demonstrate that our proposed method can achieve the comparable performance with the same candidate size while accelerate the whole sampling process $2.2 \times$ for Stable-Diffusion-3.5-Large and $2.0 \times$ for FLUX.1-dev. Also, with limited extra inference time costs, our proposed method improves the total quality of Stable-Diffusion-3.5-Large and FLUX.1-dev by 10.84% and 9.90%.

In summery, our contributions can be concluded as follows:

- We propose a novel inference time scaling framework for fast sampling. It can efficiently generate a larger pool of candidate samples via Co-Fusion pipeline within the limited budget while maintaining the main content of the images compared with the original pipeline, providing a wider search space for samples and improve the overall performance;
- We propose a reflection mechanism enables the refinement of details and the rectify the anomalous regions of the selected samples. It injects the feedback of the vision language model into the continue generation process, suppressing undesired or unexpected elements in the generated results;
- Extensive experiments on Stable-Diffusion-3.5-Large and FLUX.1-dev demonstrate the effectiveness of our proposed method. Our proposed method can achieve the comparable performance with the same number of candidates while accelerating the whole sampling process. Also, with limited extra inference time costs, our proposed method improves the total quality of Stable-Diffusion-3.5-Large and FLUX.1-dev by 10.84% and 9.90%.

2 RELATED WORKS

2.1 INFERENCE TIME SCALING OF DIFFUSION MODELS

Diffusion models Ho et al. (2020); Song et al. (2020); Rombach et al. (2022); Esser et al. (2024); Labs (2024); Podell et al. (2023) excel in text-to-image generation but suffer from randomness and inconsistent quality Zhou et al. (2024); Ahn et al. (2024); Qi et al. (2024); Bai et al. (2024); Xu et al. (2023b). Prior works mainly enhance quality via training-time scaling, *e.g.*, larger models, datasets, or training duration, Mei et al. (2024); Li et al. (2024); Liang et al. (2024); Podell et al. (2023); Chen et al. (2025), which is computationally costly. Inspired by large language models Brown et al. (2024); Snell et al. (2024); Wu et al. (2024a); Gandhi et al. (2024); Balachandran et al. (2025); Zhang et al. (2025), inference time scaling has recently attracted interest in vision Ma et al. (2025); Xie et al. (2025); Kim et al. (2025). Inference-time methods fall into two categories: global trajectory search, finding better initial noise, *e.g.*, random search or Zero-Order Search Ma et al. (2025); and stepwise trajectory search, optimizing update directions during denoising, *e.g.*, Search over Paths Ma et al. (2025). The two are orthogonal, targeting the starting point and the generation process, and can be combined. Our method focuses on efficiency and can serve as a plug-and-play enhancement.

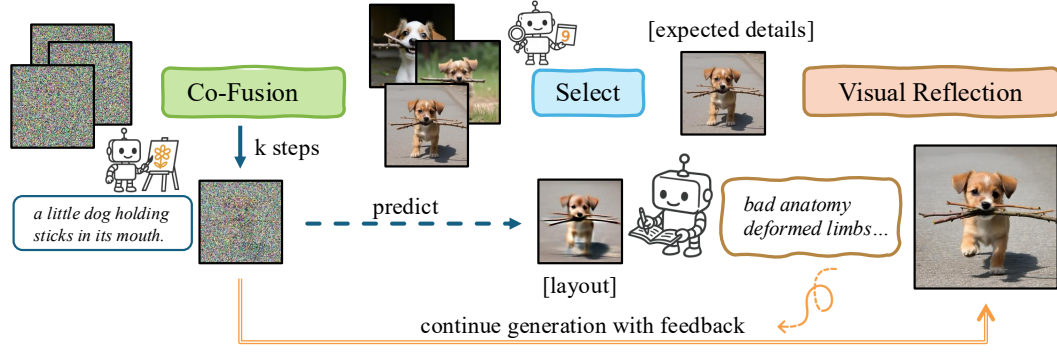


Figure 1: The overall framework of our proposed method. Our proposed method includes three parts. Firstly, we introduce Co-Fusion pipeline to fast sample larger set of candidates. Then, we adopt pre-defined evaluation metrics to select the highest fidelity candidate within the set. Lastly, we leverage vision language model to provide feedback of the potential flaws in the reference images, and inject the feedback to the continue generation process from early stage.

2.2 ALIGNMENT OF DIFFUSION MODELS

Alignment Liu et al. (2024); Lee et al. (2023); Xu et al. (2023a); Black et al. (2023); Fan et al. (2023); Clark et al. (2023); Wallace et al. (2024); Yang et al. (2024c;a); Liang (2024) adapts pretrained diffusion models for human-preferred outputs beyond training distribution. The two main training-based strategies are RLHF (learning a reward model then finetuning) Lee et al. (2023); Xu et al. (2023a); Black et al. (2023); Fan et al. (2023); Clark et al. (2023); Zhang et al. (2024); Uehara et al. (2024) and DPO (direct optimization on preference data without reward models) Wallace et al. (2024); Yang et al. (2024c;a); Liang (2024); Yuan et al. (2024); Zhu et al. (2025); Na et al. (2024). Other alignment methods avoid finetuning, such as prompt optimization Hao et al. (2023); Wang et al. (2023); Mañas et al. (2024); Mo et al. (2024), noise optimization Eyring et al. (2024); Tang et al. (2024); Bai et al. (2024); Zhou et al. (2024); Qi et al. (2024), and attention control Wu et al. (2024b); Yang et al. (2024b); Hong et al. (2023). Alignment refines the objective of models to favor human preferences, while inference time scaling searches within the output space to exploit model capacity. These two approaches are complementary, and our method can be integrated with both to further improve image quality.

3 METHOD

3.1 PRELIMINARY

Assume the time cost of full sampling process is τ , and the total time budget for inference time scaling is T , the number of candidate samples can be generated is $n = \lfloor \frac{T}{\tau} \rfloor$. For each $i = 1, \dots, n$, $x_0^{(i)} \sim \mathcal{N}(0, I)$, and $I^{(i)} = G(x_0^{(i)}, \theta)$. Here, $G(\cdot; \theta)$ refers the total denoising process mapping parameterized by θ . After repeated sampling process, the final images are selected from the candidate sets with the predefined metric $R(\cdot)$,

$$I^* = \arg \max_{I^{(i)} \in \{I^{(1)}, \dots, I^{(n)}\}} R(I^{(i)}). \quad (1)$$

Denote cumulative distribution function of each score $R^{(i)} = R(I^{(i)})$ as $F_R(\gamma) = P(R^{(i)} \leq \gamma)$, the expected maximum score over n samples is as follows:

$$Q(n) = \mathbb{E} \left[\max_{1 \leq i \leq n} R^{(i)} \right] = \int_{-\infty}^{\infty} P(\max_i R^{(i)} > \gamma) d\gamma = \int_{-\infty}^{\infty} [1 - F_R(\gamma)^n] d\gamma. \quad (2)$$

As n increases, the distribution shifts rightward, yielding higher expected scores, but marginal gains diminish accordingly.



Figure 2: The visualization and quantitative results for different choice of k and z . The results demonstrate that smaller Δ contributes to more similar production.

3.2 CO-FUSION PIPELINE

Current widely-adopted inference time scaling method requires repeated sampling through the whole pipeline, which is time-consuming. Also, as presented in Eq. 2, the increase in number of candidate samples benefits the overall expected score. In this case, we propose Co-Fusion pipeline. It accelerates the whole sampling process, while maintaining the visual of the outputs, so that the estimators can make decisions in the early steps by consensus.

Specifically, Co-Fusion pipeline involves two stages. First, we adopt the original model performing k denoising steps to capture coarse layout of the images. Then, the generation is switched to the distilled Turbo model to quickly and seamlessly refine the details. The visual results of Co-Fusion are maintained, as the generation process is constrained under the capability of the original model. Here, assume the original model with parameter θ requires T denoising steps for generation, and the timestep is denoted as t_i for step i ; the distilled Turbo model ψ only requires S denoising steps, and $S \ll T$, and the corresponding timestep is denoted as s_i for step i . Co-Fusion can be formulated as follows:

$$\begin{aligned} x_{\theta, t_{i+1}} &= x_{\theta, t_i} + \delta_{t_{i+1}} v_{\theta}(x_{\theta, t_i}, y; t_i), \text{ where } i = 0, 1, \dots, k-1, x_{\theta, t_0} \sim \mathcal{N}(0, I), \\ x_{\psi, s_{j+1}} &= x_{\psi, s_j} + \delta_{s_{j+1}} v_{\psi}(x_{\psi, s_j}, y; s_j), \text{ where } j = z, \dots, S-1, x_{\psi, s_z} = x_{\theta, t_k}, \end{aligned} \quad (3)$$

where $\delta_{t_{i+1}} = t_{i+1} - t_i$, and $\delta_{t_{j+1}} = t_{j+1} - t_j$. $x_{\psi, s_S} = x_N$ is the final output of the Co-Fusion pipeline, and the number of the total steps is $N = k + S - z < T$. y refer to the condition for image generation, e.g., text prompt. We also provide theoretical analysis of error bound of Co-Fusion pipeline to further validate its effectiveness. For more detailed theoretical analysis, please refer to the supplementary.

Proposition 1. *Error bound for Co-Fusion pipeline*

Proof. For the first k steps, Co-Fusion and the original pipeline share the same process. Thus, in this case, $e_{\theta, 0 \rightarrow k} = 0$, and $x_{\theta, t_k} = x_{\psi, s_z}$. For the rest original pipeline $T - k$ steps, we have:

$$\begin{aligned} e_{\theta, i} &= x_{\theta, t_i} - x_{\theta, t_i}^*, \quad \varepsilon_{\theta, i} = x_{\theta, t_i}^* - x_{\theta, t_{i-1}}^* - \delta_{t_i} v_{\theta}(x_{\theta, t_{i-1}}^*, y; t_{i-1}), \quad \|\varepsilon_{\theta, i}\| \leq C \delta_{t_i}^2, \\ e_{\theta, i} &= e_{\theta, i-1} + \delta_{t_i} [v_{\theta}(x_{\theta, t_{i-1}}, y; t_{i-1}) - v_{\theta}(x_{\theta, t_{i-1}}^*, y; t_{i-1})] - \varepsilon_{\theta, i}. \end{aligned} \quad (4)$$

Here, denote the Lipschitz constant for v_{θ} and x as L , we have:

$$\begin{aligned} \|v_{\theta}(x_{\theta, t_{i-1}}, y; t_{i-1}) - v_{\theta}(x_{\theta, t_{i-1}}^*, y; t_{i-1})\| &\leq L \|x_{\theta, t_{i-1}} - x_{\theta, t_{i-1}}^*\| = L \|e_{\theta, i-1}\|, \\ \|e_{\theta, i}\| &\leq \|e_{\theta, i-1}\| + \|\delta_{t_i} [v_{\theta}(x_{\theta, t_{i-1}}, y; t_{i-1}) - v_{\theta}(x_{\theta, t_{i-1}}^*, y; t_{i-1})]\| + \|\varepsilon_{\theta, i}\| \\ &\leq (1 + \delta_{t_i} L) \|e_{\theta, i-1}\| + C \delta_{t_i}^2. \end{aligned} \quad (5)$$

After iterative expansion, we have:

$$\begin{aligned} \|e_{\theta,T}\| &\leq \left[\prod_{m=k+1}^T (1 + \delta_{t_m} L) \right] \|e_{\theta,k}\| + C \sum_{m=k+1}^T \delta_{t_m}^2 \prod_{p=m+1}^T (1 + \delta_{t_p} L), \\ 1 + \delta L &\leq e^{\delta L}, \quad \|e_{\theta,T}\| \leq e^{(t_T - t_k)L} \|e_{\theta,k}\| + C \sum_{m=k+1}^T \delta_{t_m}^2 e^{(t_T - t_m)L}. \end{aligned} \quad (6)$$

Denote $\Delta_{ori} = t_T - t_k$, $\delta_{max_t} = \max_i \delta_{t_i}$, $\sum_{m=k+1}^T \delta_{t_m}^2 \leq \Delta_{ori} \delta_{max_t}$, $e^{\Delta_{ori} L} \Delta_{ori} \leq \frac{e^{\Delta_{ori} L} - 1}{L}$, then we have:

$$\|e_{\theta,T}\| \leq e^{\Delta_{ori} L} \|e_{\theta,k}\| + \frac{e^{\Delta_{ori} L} - 1}{L} C \delta_{max_t}. \quad (7)$$

For the rest Co-Fusion $S - z$ steps, we have:

$$\begin{aligned} x_{\psi,j} &= x_{\psi,j-1} + \delta_{s_j} v_{\psi}(x_{\psi,j-1}, y; s_{j-1}) = x_{\psi,j-1} + \delta_{s_j} [v_{\theta}(x_{\psi,j-1}, y; s_{j-1}) + \varsigma_{s_{j-1}}], \\ e_{\psi,j} &= x_{\psi,s_j} - x_{\theta,s_j}^*, \quad \varepsilon_{\theta,j} = x_{\theta,s_j}^* - x_{\theta,s_{j-1}}^* - \delta_{s_j} v_{\theta}(x_{\theta,s_{j-1}}^*, y; s_{j-1}), \quad \|\varepsilon_{\theta,j}\| \leq C \delta_{s_j}^2, \\ e_{\psi,j} &= e_{\psi,j-1} + \delta_{s_j} [v_{\theta}(x_{\psi,j-1}, y; s_{j-1}) - v_{\theta}(x_{\theta,s_{j-1}}^*, y; s_{j-1})] + \delta_{s_j} \varsigma_{s_{j-1}} - \varepsilon_{\theta,j}. \end{aligned} \quad (8)$$

Here, denote the distillation error bound as the constant ε_{dis} , $\|\varsigma\| \leq \varepsilon_{dis}$, and the Lipschitz constant for v_{θ} and x as L , we have:

$$\begin{aligned} \|e_{\psi,j}\| &\leq (1 + \delta_{s_j} L) \|e_{\psi,j-1}\| + C \delta_{s_j}^2 + \delta_{s_j} \|\varsigma_{s_{j-1}}\| \\ &\leq (1 + \delta_{s_j} L) \|e_{\psi,j-1}\| + C \delta_{s_j}^2 + \delta_{s_j} \varepsilon_{dis} \end{aligned} \quad (9)$$

Similarly, we apply iterative expansion to Eq. 9, and we have:

$$\begin{aligned} \|e_{\psi,S}\| &\leq \left[\prod_{m=z+1}^S (1 + \delta_{s_m} L) \right] \|e_{\psi,z}\| + \sum_{m=z+1}^S (C \delta_{s_m}^2 + \delta_{s_m} \varepsilon_{dis}) \prod_{p=m+1}^S (1 + \delta_{s_p} L) \\ &\leq e^{\Delta_{co} L} \|e_{\psi,z}\| + \frac{e^{\Delta_{co} L} - 1}{L} (C \delta_{max_s} + \varepsilon_{dis}), \end{aligned} \quad (10)$$

where $\Delta_{co} = s_S - s_z = \Delta_{ori}$, we denote it as Δ for clarification, and $\|e_{\psi,z}\| = \|e_{\theta,k}\| = 0$, $\delta_{max_s} = \max_j \delta_{s_j}$. The final error bound for Co-Fusion pipeline is as follows:

$$\|x_{\theta,t_T} - x_{\psi,s_S}\| \leq \frac{e^{\Delta \cdot L} - 1}{L} [C(\delta_{max_t} + \delta_{max_s}) + \varepsilon_{dis}]. \quad (11)$$

The error for the Co-Fusion pipeline is bounded with the length of the left steps and the distillation loss between the two models. $\square$

3.3 VISUAL REFLECTION

After selection part performed in the early stage, the selected trajectory will continue to generate the final output with the original model. However, as mentioned in Eq. 2, the performance of the final output is bounded by the size of the candidate set. Meanwhile, from the observation, the selected images are the least imperfect among the candidates and can not guarantee the high-fidelity. To further improve the quality of the selected images, we propose the visual reflection method during the continue generation process.

Specifically, we first project the noisy image at step k into the predicted final result:

$$\hat{x}_{\theta,t_T} = x_{\theta,t_k} - t_k v_{\theta}(x_{\theta,t_k}, y; t_k), \quad (12)$$

which presents the layout of the image. Then, we leverage the visual understanding capabilities of the vision language model $\mathcal{M}$, *e.g.*, Qwen2.5-VL, to point out the flaws of the reference images (the predicted image $\hat{x}_{\theta,t_T}$ and the corresponding candidate image x_N generated through Co-Fusion pipeline), such as distorted figures, low fidelity, *etc.* Here, the predicted image can reflect the overall

Algorithm 1: SF2 Framework

Input: Original model with parameter θ , turbo model with parameter ψ , vision language model $\mathcal{M}$, evaluation metric R , prompt y , the number of candidates n

Output: High-fidelity $x'_{\theta,T}$ for prompt y

Random initialize $x_{\theta,t_0} \sim \mathcal{N}(0, I)$

Candidates generation

for $idx = 1$ **to** n **do**

 # Co-Fusion pipeline

 Generate candidate $x_N^{(idx)}$ through Eq. 3 $\triangleright$ Store k step $x_{\theta,t_k}^{(idx)}$ for continue generation

$I_N^{(idx)} = \mathcal{D}_\psi(x_N^{(idx)})$

$idx^* = \arg \max_{1 \leq idx \leq n} R(I^{(idx)})$

 # Visual reflection

 Get predicted final output $\hat{x}_{\theta,t_T}^{(idx^*)}$ with $x_{\theta,t_k}^{(idx^*)}$ through Eq. 12

 Generate feedback of the flaws in the reference images through vision language model $\mathcal{M}$,

$y' = \mathcal{M}(\hat{I}_{\theta,t_T}^{(idx^*)}, I_N^{(idx^*)}, p)$, where $\hat{I}_{\theta,t_T}^{(idx^*)} = \mathcal{D}_\theta(\hat{x}_{\theta,t_T}^{(idx^*)})$

 # Continue generation

 Generate the final result $x'_{\theta,T}$ with $x_{\theta,t_k}^{(idx^*)}$ through Eq. 14.

return $x'_{\theta,T}$

layout of the final image and the candidate image can demonstrate the expected details of the final image. The process can be formulated as follows:

$$\begin{aligned} \hat{I}_{\theta,t_T} &= \mathcal{D}_\theta(\hat{x}_{\theta,t_T}), \quad I_N = \mathcal{D}_\psi(x_N) \\ y' &= \mathcal{M}(\hat{I}_{\theta,t_T}, I_N, p), \end{aligned} \quad (13)$$

where y' is the negative prompt identifying the potential defects of the selected images, $\mathcal{D}$ is decoder, p refer to the instruction prompt. After that, we inject the feedback y' into the continue generation process from the breakpoint k to get the final output $x'_{\theta,T}$:

$$x'_{\theta,t_{i+1}} = x_{\theta,t_i} + \delta_{t_{i+1}} v_\theta(x_{\theta,t_i}, y, y'; t_i), \text{ where } i = k, \dots, T-1. \quad (14)$$

Since the early step k is still in layout generation phase, it exhibits effective corrective efficacy for mitigating potential flaws, while preserving the initial favorable layout and the trajectory of fine-detail synthesis.

3.4 ALGORITHM SUMMARY

In this paper, we propose a novel inference time scaling framework for fast and broader sampling with early trajectory selection and reflection. The algorithm summary is shown in Algorithm 1. Our proposed framework includes three parts. Firstly, we fast build the candidate set through our proposed Co-Fusion pipeline, which enables the estimator to make decision in the early stage. After selection, we leverage the vision language model to generate feedback on the flaws of the reference images (the predicted image $\hat{I}_{\theta,t_T}^{(idx^*)}$ in the k step and the selected candidate image $I_N^{(idx^*)}$ generated through Co-Fusion pipeline). These reference images can demonstrate the layout and expected details information, respectively, offering a broader view for vision language model to provide correct and effective feedback. Then, we inject the generated feedback into the continue generation process from the breakpoint k of the original pipeline to get the final output $x'_{\theta,T}$. As k is in early stage of the whole pipeline, our method can smoothly correct the potential flaws while preserving the high-quality layout and preventing artifact generation.

Table 1: Quantitative results of baseline methods and our proposed method. Here, NFEs records the total budget each prompt requires to obtain the final image. Avg RI denotes the average relative improvement of the four evaluation metrics over the original.

Method/Settings	NFEs	Estimator	HPSv2.1	ImageReward	PickScore	Aesthetic Score	Avg RI (%)
Stable-Diffusion-3.5-Large							
Original	28	-	30.95	1.0441	23.03	5.435	-
SDS	280	-	30.31	1.1220	22.73	5.414	0.93%
Random ($n = 10$)	280	-	30.80	1.0408	23.02	5.435	-0.21%
BoN ($n = 10$)	280	HPSv2.1	32.78	1.2095	23.37	5.493	6.07%
Ours ($n = 10$)	150	HPSv2.1	32.92	1.2164	23.38	5.595	6.83%
Ours ($n = 20$)	280	HPSv2.1	33.22	1.2418	23.43	5.609	7.80%
BoN ($n = 10$)	280	ImageReward	31.64	1.4058	23.22	5.464	9.56%
Ours ($n = 10$)	150	ImageReward	31.98	1.3731	23.26	5.568	9.57%
Ours ($n = 20$)	280	ImageReward	32.10	1.4184	23.29	5.579	10.84%
BoN ($n = 10$)	280	PickScore	31.88	1.1751	23.67	5.490	4.84%
Ours ($n = 10$)	150	PickScore	32.02	1.1893	23.59	5.590	5.66%
Ours ($n = 20$)	280	PickScore	32.31	1.2133	23.70	5.604	6.65%
FLUX.1-dev							
Original	28	-	32.08	1.0835	23.32	5.719	-
SDS	280	-	31.69	1.0835	23.19	5.680	-0.61%
Random ($n = 10$)	280	-	32.00	1.0720	23.30	5.702	-0.42%
BoN ($n = 10$)	280	HPSv2.1	33.73	1.2334	23.66	5.744	5.22%
Ours ($n = 10$)	130	HPSv2.1	33.64	1.2392	23.64	5.805	5.53%
Ours ($n = 20$)	240	HPSv2.1	33.94	1.2586	23.70	5.817	6.33%
BoN ($n = 10$)	280	ImageReward	32.61	1.4441	23.50	5.715	8.91%
Ours ($n = 10$)	130	ImageReward	32.70	1.4163	23.50	5.788	8.66%
Ours ($n = 20$)	240	ImageReward	32.82	1.4627	23.54	5.796	9.90%
BoN ($n = 10$)	280	PickScore	32.91	1.2069	23.96	5.738	4.26%
Ours ($n = 10$)	130	PickScore	32.93	1.2180	23.88	5.804	4.74%
Ours ($n = 20$)	240	PickScore	33.07	1.2340	23.99	5.810	5.36%

Table 2: Total runtime for each prompt to obtain the final image via different inference time scaling method. Here, the runtime includes image generation and selection parts. For our proposed method, it includes candidate image generation through Co-Fusion pipeline, image selection, visual reflection and continue generation. Here, we report the total runtime with HPSv2.1, ImageReward, and PickScore estimators, and the results are separated by the slash.

Stable-Diffusion-3.5-Large					
Original	SDS	Random ($n = 10$)	BoN ($n = 10$)	Ours ($n = 10$)	Ours ($n = 20$)
19.82s	195.28s	197.12s	204.60s / 198.19s / 198.12s	98.55s / 92.15s / 92.08s	182.87s / 170.40s / 170.38s
FLUX.1-dev					
Original	SDS	Random ($n = 10$)	BoN ($n = 10$)	Ours ($n = 10$)	Ours ($n = 20$)
22.50s	223.04s	225.54s	233.02s / 226.61s / 226.54s	120.43s / 114.02s / 113.95s	221.77s / 209.36s / 209.34s

4 EXPERIMENTS

4.1 SETTINGS AND IMPLEMENTATION DETAILS

We conduct experiments on the open-source diffusion models Stable-Diffusion-3.5-Large and FLUX.1-dev to validate our proposed method SF2. For each prompt, we set 280 NFEs as the inference budget. The default number of generation steps is 28 (28 NFEs), for both two models across all experimental settings. For the original model and scale denoising steps (SDS) method, the seed is 42. For the rest methods (random select, best-of-n (BoN), and ours), the seeds are randomly sampled from a uniform distribution (the value is $0 \sim 10,000$), but remain same for the three methods. For our proposed method, we adopt the corresponding Turbo models (SD-3.5-Large-Turbo and

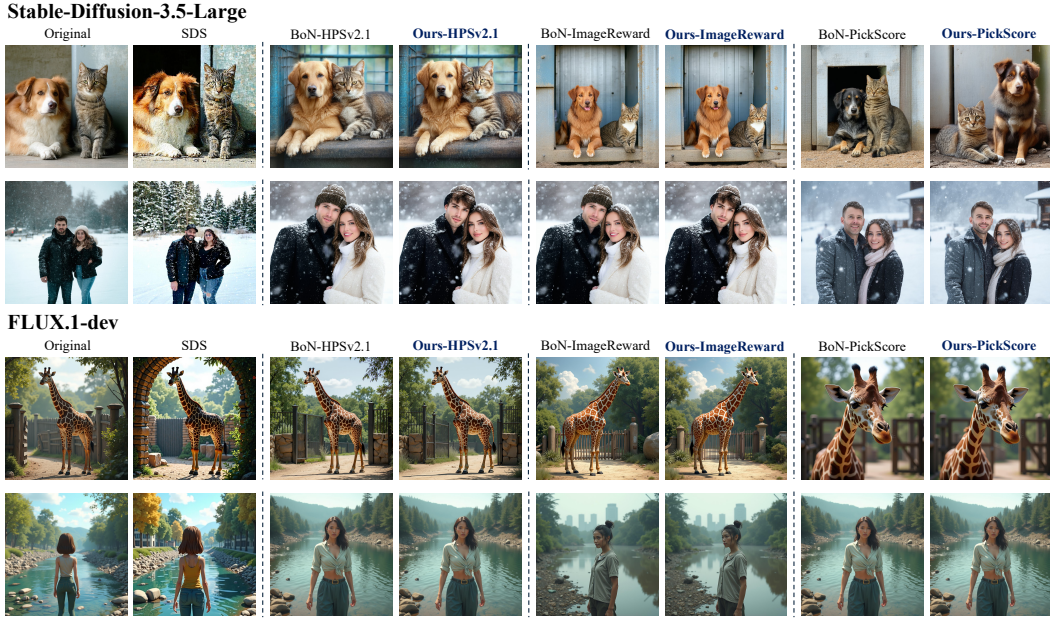


Figure 3: The visualization of our proposed method and BoN. The candidate images are generated with the prompts from COCO2014val dataset under the setting of $n = 10$.

FLUX.1-dev-Turbo) to collaboratively perform the Co-Fusion pipeline. The default number of steps for generation is 8 and 4 for SD-3.5-Large-Turbo and FLUX.1-dev respectively. Specifically, we set $k = 7$ and $z = 3$ for SD-3.5-Large and $k = 7$ and $z = 1$ for FLUX.1-dev. For visual reflection process, we adopt Qwen2.5-VL-7B model to process the selected images and generate feedback. For dataset, we adopt 5,000 prompts randomly selected from COCO2014val dataset Lin et al. (2014), and generate n candidates for each prompt. We adopt HPSv2.1 Wu et al. (2023), ImageReward Xu et al. (2023a), and PickScore Kirstain et al. (2023) as the estimator, and the performance are evaluated on the HPSv2.1, ImageReward, PickScore, and Aesthetic Score. We also report the required NFEs and runtime (s) to further demonstrate the efficiency of our proposed method. All experiments are conducted on NVIDIA RTX 6000 Ada. For more details and results, please refer to Appendix.

4.2 MAIN RESULTS

We evaluate our method on 5,000 prompts from COCO2014val dataset. For each prompt, the limited budget for generation and search is 280 NFEs. The results are shown in Table 1. We also evaluate the total runtime of the inference time scaling process for each prompt as shown in Table 2. The total runtime includes image generation and selection parts, and our method includes candidate image generation through Co-Fusion pipeline, image selection, visual reflection and continue generation. The visualization results are presented in Fig. 3. The results show that our proposed method can achieve comparable performance while only requires 53.57% and 46.43% NFEs budget of BoN. With similar NFEs budget, our proposed method substantially outperforms BoN method. For instance, our method improves FLUX.1-dev model by 9.90% with 240 NFEs compared with BoN method 8.91% with 280 NFEs. Moreover, the visualization results also show that our method can effectively restore the anomalous regions in the images, and enhance overall quality of images.

4.3 ABLATION STUDIES

We also conduct ablation studies to validate the effectiveness of each component proposed in our proposed method. The experimental settings are same with the baseline comparison. The results are demonstrated in Table 3. From the results, our proposed Co-Fusion pipeline can quickly and accurately restore the outputs of the original pipeline, enabling precise selection while significantly reducing required extra time. Moreover, our proposed visual reflection effectively corrects unnatural

Table 3: Ablation studies for each component of our proposed method. NFEs reports the total required budget for each prompt to obtain the final image. Avg RI records the average relative improvement of the four evaluation metrics over the original.

Method/Settings	NFEs	Estimator	HPSv2.1	ImageReward	PickScore	Aesthetic Score	Avg RI (%)
Stable-Diffusion-3.5-Large							
Original	28	-	30.95	1.0441	23.03	5.435	-
+ Co-Fusion ($n = 10$)	150	HPSv2.1	32.61	1.2014	23.34	5.492	5.71%
+ Visual-Reflection ($n = 10$)	150	HPSv2.1	32.92	1.2164	23.38	5.595	6.83%
+ Co-Fusion ($n = 20$)	280	HPSv2.1	32.92	1.2267	23.41	5.496	6.66%
+ Visual-Reflection ($n = 20$)	280	HPSv2.1	33.22	1.2418	23.43	5.609	7.80%
+ Co-Fusion ($n = 10$)	150	ImageReward	31.62	1.3662	23.22	5.463	8.59%
+ Visual-Reflection ($n = 10$)	150	ImageReward	31.98	1.3731	23.26	5.568	9.57%
+ Co-Fusion ($n = 20$)	280	ImageReward	31.74	1.4137	23.25	5.460	9.84%
+ Visual-Reflection ($n = 20$)	280	ImageReward	32.10	1.4184	23.29	5.579	10.84%
+ Co-Fusion ($n = 10$)	150	PickScore	31.78	1.1719	23.57	5.490	4.57%
+ Visual-Reflection ($n = 10$)	150	PickScore	32.02	1.1893	23.59	5.590	5.66%
+ Co-Fusion ($n = 20$)	280	PickScore	31.97	1.1946	23.68	5.494	5.40%
+ Visual-Reflection ($n = 20$)	280	PickScore	32.31	1.2133	23.70	5.604	6.65%
FLUX.1-dev							
Original	28	-	32.08	1.0835	23.32	5.719	-
+ Co-Fusion ($n = 10$)	130	HPSv2.1	33.59	1.2319	23.63	5.740	5.02%
+ Visual-Reflection ($n = 10$)	130	HPSv2.1	33.64	1.2392	23.64	5.805	5.53%
+ Co-Fusion ($n = 20$)	240	HPSv2.1	33.89	1.2534	23.69	5.749	5.86%
+ Visual-Reflection ($n = 20$)	240	HPSv2.1	33.94	1.2586	23.70	5.817	6.33%
+ Co-Fusion ($n = 10$)	130	ImageReward	32.62	1.4115	23.49	5.712	8.14%
+ Visual-Reflection ($n = 10$)	130	ImageReward	32.70	1.4163	23.50	5.788	8.66%
+ Co-Fusion ($n = 20$)	240	ImageReward	32.73	1.4595	23.52	5.721	9.41%
+ Visual-Reflection ($n = 20$)	240	ImageReward	32.82	1.4627	23.54	5.796	9.90%
+ Co-Fusion ($n = 10$)	130	PickScore	32.87	1.2076	23.88	5.736	4.15%
+ Visual-Reflection ($n = 10$)	130	PickScore	32.93	1.2180	23.88	5.804	4.74%
+ Co-Fusion ($n = 20$)	240	PickScore	33.01	1.2262	23.99	5.744	4.84%
+ Visual-Reflection ($n = 20$)	240	PickScore	33.07	1.2340	23.99	5.810	5.36%

artifacts in the source images and enhances the overall image quality, while preserving most of the original content, *e.g.*, layout, fine details. It can also be evidenced by the visualization results in Fig. 3. Also, the results demonstrate that increasing n can improve the quality of the final results.

4.4 DISCUSSION

The experimental results indicate that the final quality of the selected image is highly correlated with the adopted estimator. Although the estimators HPSv2.1, ImageReward, and PickScore are all trained on user preference datasets, which incorporate both aesthetics and prompt following, they diverge in their selection due to unavoidable biases in training data. Nevertheless, these models remain interrelated: using any estimator also boosts performance on the other metrics. Our proposed method effectively balances all four evaluation metrics by further incorporating visual reflection to restore the unnatural artifacts and improve the quality of the final image.

5 CONCLUSION

In this paper, we propose a novel efficient inference time scaling framework **SF2** for visual generation. First, we propose Co-Fusion pipeline for fast and broader sampling, while maintaining the similar image content for estimators to make decision. It can efficiently generate a larger candidate pool within the limited budget, providing a wider search space and improving the overall achievable performance. After selection, we propose visual reflection mechanism to rectify the unnatural regions of the selected images and further refine the overall details, while preserving the main contents. Extensive experiments demonstrate the effectiveness of our proposed method.

REFERENCES

- Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyounghwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024.
- Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In *The Thirteenth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=MKvQH1ekey>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuezhi Zhu, Mingkun Yang, Zhaozhai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Vidhisha Balachandran, Jingya Chen, Lingjiao Chen, Shivam Garg, Neel Joshi, Yash Lara, John Langford, Besmira Nushi, Vibhav Vineet, Yue Wu, et al. Inference-time scaling for complex tasks: Where we stand and what lies ahead. *arXiv preprint arXiv:2504.00294*, 2025.
- Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.
- 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.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.
- Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024a.
- Chen Chen, Rui Qian, Wenze Hu, Tsu-Jui Fu, Jialing Tong, Xinze Wang, Lezhi Li, Bowen Zhang, Alex Schwing, Wei Liu, et al. Dit-air: Revisiting the efficiency of diffusion model architecture design in text to image generation. *arXiv preprint arXiv:2503.10618*, 2025.
- Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024b.
- Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.
- Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.
- Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchun Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20637–20647, 2023.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

- Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024.
- Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.
- Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D Goodman. Stream of search (sos): Learning to search in language. *arXiv preprint arXiv:2404.03683*, 2024.
- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.
- Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.
- Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.
- Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.
- Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.
- Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, ZuoZhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.
- Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. On the scalability of diffusion-based text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9400–9409, 2024.
- Zhanhao Liang. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2(3), 2024.

- Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling laws for diffusion transformers. *arXiv preprint arXiv:2410.08184*, 2024.
- 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*, pp. 740–755. Springer, 2014.
- Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv preprint arXiv:2409.07253*, 2024.
- Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2837–2845, 2021.
- Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.
- Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024.
- Kangfu Mei, Zhengzhong Tu, Mauricio Delbracio, Hossein Talebi, Vishal M Patel, and Peyman Milanfar. Bigger is not always better: Scaling properties of latent diffusion models. *Transactions on Machine Learning Research*, 2024.
- Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt optimizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26627–26636, 2024.
- Sanghyeon Na, Yonggyu Kim, and Hyunjoon Lee. Boost your human image generation model via direct preference optimization. *arXiv preprint arXiv:2405.20216*, 2024.
- 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. *arXiv preprint arXiv:2307.01952*, 2023.
- Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. *arXiv preprint arXiv:2405.18881*, 2024.
- Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Sergey Levine, and Tommaso Biancalani. Feedback efficient online fine-tuning of diffusion models. *arXiv preprint arXiv:2402.16359*, 2024.

- Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.
- Yunlong Wang, Shuyuan Shen, and Brian Y Lim. Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–29, 2023.
- Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024a.
- Yihang Wu, Xiao Cao, Kaixin Li, Zitan Chen, Haonan Wang, Lei Meng, and Zhiyong Huang. Towards better text-to-image generation alignment via attention modulation. *arXiv preprint arXiv:2404.13899*, 2024b.
- Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023a.
- Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. *Advances in Neural Information Processing Systems*, 36:76806–76838, 2023b.
- Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiabin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024a.
- Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024b.
- Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024c.
- Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1191–1200, 2022.
- Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *arXiv preprint arXiv:2402.10210*, 2024.
- Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. What, how, where, and how well? a survey on test-time scaling in large language models. *arXiv preprint arXiv:2503.24235*, 2025.
- Ziyi Zhang, Sen Zhang, Yibing Zhan, Yong Luo, Yonggang Wen, and Dacheng Tao. Confronting reward overoptimization for diffusion models: A perspective of inductive and primacy biases. *arXiv preprint arXiv:2402.08552*, 2024.
- Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Zikai Zhou, Shitong Shao, Lichen Bai, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. *arXiv preprint arXiv:2411.09502*, 2024.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. Dspo: Direct score preference optimization for diffusion model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

A APPENDIX

A.1 THEORETICAL ANALYSIS FOR ERROR BOUND OF CO-FUSION PIPELINE

Co-Fusion pipeline has total inference steps $N = k + S - z < T$. Here, S refers to the number of total inference steps for Turbo model, and T refers to the number of total inference steps for the original model. For the first k steps, Co-Fusion and the original pipeline share the same process. Thus, in this case, the error for the first k steps $e_{\theta,0 \rightarrow k} = 0$, and $x_{\theta,t_k} = x_{\psi,s_z}$. For the rest original pipeline $T - k$ steps, we have:

$$\begin{aligned} e_{\theta,i} &= x_{\theta,t_i} - x_{\theta,t_i}^*, \quad \varepsilon_{\theta,i} = x_{\theta,t_i}^* - x_{\theta,t_{i-1}}^* - \delta_{t_i} v_{\theta}(x_{\theta,t_{i-1}}^*, y; t_{i-1}), \quad \|\varepsilon_{\theta,i}\| \leq C\delta_{t_i}^2, \\ e_{\theta,i} &= e_{\theta,i-1} + \delta_{t_i} [v_{\theta}(x_{\theta,t_{i-1}}, y; t_{i-1}) - v_{\theta}(x_{\theta,t_{i-1}}^*, y; t_{i-1})] - \varepsilon_{\theta,i}. \end{aligned} \quad (15)$$

Here, denote the Lipschitz constant for v_{θ} and x as L , we have:

$$\begin{aligned} \|v_{\theta}(x_{\theta,t_{i-1}}, y; t_{i-1}) - v_{\theta}(x_{\theta,t_{i-1}}^*, y; t_{i-1})\| &\leq L\|x_{\theta,t_{i-1}} - x_{\theta,t_{i-1}}^*\| = L\|e_{\theta,i-1}\|, \\ \|e_{\theta,i}\| &= \|e_{\theta,i-1} + \delta_{t_i} [v_{\theta}(x_{\theta,t_{i-1}}, y; t_{i-1}) - v_{\theta}(x_{\theta,t_{i-1}}^*, y; t_{i-1})] - \varepsilon_{\theta,i}\| \\ &\leq \|e_{\theta,i-1}\| + \|\delta_{t_i} [v_{\theta}(x_{\theta,t_{i-1}}, y; t_{i-1}) - v_{\theta}(x_{\theta,t_{i-1}}^*, y; t_{i-1})]\| + \|\varepsilon_{\theta,i}\| \\ &= \|e_{\theta,i-1}\| + \delta_{t_i} L \|e_{\theta,i-1}\| + \|\varepsilon_{\theta,i}\| \\ &\leq (1 + \delta_{t_i} L) \|e_{\theta,i-1}\| + C\delta_{t_i}^2. \end{aligned} \quad (16)$$

After iterative expansion, we have:

$$\begin{aligned} \|e_{\theta,T}\| &\leq \left[\prod_{m=k+1}^T (1 + \delta_{t_m} L) \right] \|e_{\theta,k}\| + C \sum_{m=k+1}^T \delta_{t_m}^2 \prod_{p=m+1}^T (1 + \delta_{t_p} L), \\ 1 + \delta L &\leq e^{\delta L}, \quad \|e_{\theta,T}\| \leq e^{(t_T - t_k)L} \|e_{\theta,k}\| + C \sum_{m=k+1}^T \delta_{t_m}^2 e^{(t_T - t_m)L}. \end{aligned} \quad (17)$$

Denote $\Delta_{ori} = t_T - t_k$, $\delta_{max_t} = \max_i \delta_{t_i}$, $\sum_{m=k+1}^T \delta_{t_m}^2 \leq \Delta_{ori} \delta_{max_t}$, $e^{\Delta_{ori} L} \Delta_{ori} \leq \frac{e^{\Delta_{ori} L} - 1}{L}$, then we have:

$$\|e_{\theta,T}\| \leq e^{\Delta_{ori} L} \|e_{\theta,k}\| + \frac{e^{\Delta_{ori} L} - 1}{L} C \delta_{max_t}. \quad (18)$$

For the rest Co-Fusion $S - z$ steps, we have:

$$\begin{aligned} x_{\psi,j} &= x_{\psi,j-1} + \delta_{s_j} v_{\psi}(x_{\psi,j-1}, y; s_{j-1}) = x_{\psi,j-1} + \delta_{s_j} [v_{\theta}(x_{\psi,j-1}, y; s_{j-1}) + \varsigma_{s_{j-1}}], \\ e_{\psi,j} &= x_{\psi,s_j} - x_{\theta,s_j}^*, \quad \varepsilon_{\theta,j} = x_{\theta,s_j}^* - x_{\theta,s_{j-1}}^* - \delta_{s_j} v_{\theta}(x_{\theta,s_{j-1}}^*, y; s_{j-1}), \quad \|\varepsilon_{\theta,j}\| \leq C\delta_{s_j}^2, \\ e_{\psi,j} &= e_{\psi,j-1} + \delta_{s_j} [v_{\theta}(x_{\psi,s_{j-1}}, y; s_{j-1}) - v_{\theta}(x_{\theta,s_{j-1}}^*, y; s_{j-1})] + \delta_{s_j} \varsigma_{s_{j-1}} - \varepsilon_{\theta,j}. \end{aligned} \quad (19)$$

Here, denote the distillation error bound as the constant ε_{dis} , $\|\varsigma\| \leq \varepsilon_{dis}$, and the Lipschitz constant for v_{θ} and x as L , we have:

$$\begin{aligned} \|e_{\psi,j}\| &\leq (1 + \delta_{s_j} L) \|e_{\psi,j-1}\| + C\delta_{s_j}^2 + \delta_{s_j} \|\varsigma_{s_{j-1}}\| \\ &\leq (1 + \delta_{s_j} L) \|e_{\psi,j-1}\| + C\delta_{s_j}^2 + \delta_{s_j} \varepsilon_{dis} \end{aligned} \quad (20)$$

Table 4: Hyper-parameter settings of Co-Fusion pipeline. Hyper-parameters not listed in table are set as default.

Models	Num Inference Steps	Guidance Scale	Max Sequence Length	k or z
SD-3.5-Large	28	3.5	256	$k = 7$
SD-3.5-large-TurboX	8	1.5	256	$z = 3$
FLUX.1-dev	28	3.5	512	$k = 7$
FLUX.1-Turbo-Alpha	4	3.5	512	$z = 1$

Similarly, we apply iterative expansion to Eq. 20, and we have:

$$\begin{aligned}
\|e_{\psi,S}\| &\leq \left[\prod_{m=z+1}^S (1 + \delta_{s_m} L) \right] \|e_{\psi,z}\| + \sum_{m=z+1}^S (C\delta_{s_m}^2 + \delta_{s_m} \varepsilon_{dis}) \prod_{p=m+1}^S (1 + \delta_{s_p} L) \\
&\leq e^{(s_S - s_z)} \|e_{\psi,z}\| + \sum_{m=z+1}^S (C\delta_{s_m}^2 + \delta_{s_m} \varepsilon_{dis}) e^{(s_S - s_z)} \\
&\leq e^{\Delta_{co} L} \|e_{\psi,z}\| + \frac{e^{\Delta_{co} L} - 1}{L} (C\delta_{max_s} + \varepsilon_{dis}),
\end{aligned} \tag{21}$$

where $\Delta_{co} = s_S - s_z = \Delta_{ori}$, we denote it as Δ for clarification, and $\|e_{\psi,z}\| = \|e_{\theta,k}\| = 0$, $\delta_{max_s} = \max_j \delta_{s_j}$. The final error bound for Co-Fusion pipeline is as follows:

$$\|x_{\theta,t_T} - x_{\psi,s_S}\| \leq \frac{e^{\Delta \cdot L} - 1}{L} [C(\delta_{max_t} + \delta_{max_s}) + \varepsilon_{dis}]. \tag{22}$$

The error for the Co-Fusion pipeline is bounded with the length of the left steps and the distillation loss between the two models.

A.2 MORE EXPERIMENTAL DETAILS

In the paper, experiments are conducted on the open-source diffusion models Stable-Diffusion-3.5-Large (*stabilityai/stable-diffusion-3.5-large*) and FLUX.1-dev (*black-forest-labs/FLUX.1-dev*) to validate the effectiveness of our proposed method SF2. For our proposed method, we adopt the corresponding Turbo models Stable-Diffusion-3.5-Large-TurboX (*tensorart/stable-diffusion-3.5-large-TurboX*) and FLUX.1-Turbo-Alpha (*alimama-creative/FLUX.1-Turbo-Alpha*) to collaboratively perform the Co-Fusion pipeline. The default total number of whole denoising process is 28 for Stable-Diffusion-3.5-Large and FLUX.1-dev, and 8 for Stable-Diffusion-3.5-Large-Turbo and 4 FLUX.1-Turbo-Alpha. Specifically, we set $k = 7$ and $z = 3$ for Stable-Diffusion-3.5-Large and $k = 7$ and $z = 1$ for FLUX.1-dev. For more information about Co-Fusion pipeline hyper-parameter settings, please refer to Table 4.

For dataset, we adopt 5,000 prompts randomly selected from COCO2014val dataset Lin et al. (2014), and generate n candidates for each prompt. We adopt HPSv2.1 Wu et al. (2023), ImageReward Xu et al. (2023a), and PickScore Kirstain et al. (2023) as the estimator, and the performance are evaluated on the HPSv2.1, ImageReward, PickScore, and Aesthetic Score. We also evaluate the performance on more evaluation metrics, MUSIQ Ke et al. (2021), MAN-IQA Yang et al. (2022), and TOPIQ Chen et al. (2024a), to show the quality of final images. Results are shown in Table 5. For each prompt, we set 280 NFEs as the inference budget. The default number of generation steps is 28, which is 28 NFEs for total denoising process, for both two models across all experimental settings. For the original model and scale denoising steps (SDS) method, the seed is 42. For the rest methods (random select, best-of-n (BoN), and ours), the seeds are randomly sampled from a uniform distribution (the value is $0 \sim 10,000$), but remain same for the three methods.

For visual reflection process, we adopt Qwen2.5-VL-7B (*Qwen/Qwen2.5-VL-7B-Instruct*) model to process the reference images (the selected image generated via Co-Fusion pipeline and the corresponding predicted final image from the early stage) and generate feedback. We also report the required NFEs and runtime (s) to further demonstrate the efficiency of our proposed method. All experiments are conducted on NVIDIA RTX 6000 Ada.

Table 5: Quantitative results of baseline methods and our proposed method. Here, NFEs records the total budget each prompt requires to obtain the final image. Avg RI denotes the average relative improvement of the four evaluation metrics over the original.

Method/Settings	NFEs	Estimator	MUSIQ $\uparrow$	MAN-IQA $\uparrow$	TOPIQ $\uparrow$	Avg RI (%)
Stable-Diffusion-3.5-Large						
Original	28	-	70.30	0.4446	0.5527	-
SDS	280	-	65.92	0.3314	0.4657	-15.81%
Random ($n = 10$)	280	-	70.47	0.4566	0.5596	1.40%
BoN ($n = 10$)	280	HPSv2.1	71.11	0.4529	0.5644	1.71%
Ours ($n = 10$)	150	HPSv2.1	72.14	0.4650	0.5837	4.27%
Ours ($n = 20$)	280	HPSv2.1	72.28	0.4654	0.5850	4.45%
BoN ($n = 10$)	280	ImageReward	70.68	0.4520	0.5603	1.19%
Ours ($n = 10$)	150	ImageReward	71.79	0.4637	0.5795	3.75%
Ours ($n = 20$)	280	ImageReward	71.77	0.4630	0.5799	3.72%
BoN ($n = 10$)	280	PickScore	70.71	0.4537	0.5615	1.41%
Ours ($n = 10$)	150	PickScore	71.88	0.4656	0.5815	4.06%
Ours ($n = 20$)	280	PickScore	71.82	0.4639	0.5806	3.85%
FLUX.1-dev						
Original	28	-	70.87	0.4642	0.6131	-
SDS	280	-	64.79	0.3717	0.4913	-16.12%
Random ($n = 10$)	280	-	71.27	0.4807	0.6191	1.70%
BoN ($n = 10$)	280	HPSv2.1	71.74	0.4817	0.6254	2.33%
Ours ($n = 10$)	130	HPSv2.1	72.22	0.4906	0.6379	3.88%
Ours ($n = 20$)	240	HPSv2.1	72.39	0.4923	0.6396	4.17%
BoN ($n = 10$)	280	ImageReward	71.40	0.4790	0.6205	1.71%
Ours ($n = 10$)	130	ImageReward	71.90	0.4871	0.6321	3.16%
Ours ($n = 20$)	240	ImageReward	71.91	0.4883	0.6322	3.26%
BoN ($n = 10$)	280	PickScore	71.23	0.4777	0.6205	1.54%
Ours ($n = 10$)	130	PickScore	71.61	0.4847	0.6302	2.75%
Ours ($n = 20$)	240	PickScore	71.63	0.4853	0.6304	2.81%

A.3 MORE EXPERIMENTAL RESULTS

A.3.1 QUANTITATIVE RESULTS ON MORE EVALUATION METRICS

Here, we report more quantitative results with other widely used evaluation metrics, MUSIQ Ke et al. (2021), MAN-IQA Yang et al. (2022), and TOPIQ Chen et al. (2024a), to show the effectiveness of our proposed method. The quantitative results are demonstrated in Table 5. The results demonstrate both the effectiveness and efficiency of our proposed method.

A.3.2 QUANTITATIVE RESULTS ON OTHER FAST SAMPLERS

Our proposed method supports any type of approximate fast samplers which preserve output similarity to the original pipeline, not necessarily limited to Turbo models. Here, we conduct experiments of applying our proposed framework to DBCache method on FLUX.1-dev model. The prompts are randomly selected from COCO2014val. Here, $k = 7$, and the rest part of Co-Fusion pipeline is carried out by DBCache, and $z = 8$. The estimator we adopt here is HPSv2.1. The results are as shown in the Table 6.

Table 6: Results on other fast sampler. The runtime for Ours+Turbo (in paper) to generate a candidate is 9.45s (compared with original pipeline 22.50s), and the runtime for Ours+DBCACHE is 14.15s. Evaluations are under 1024×1024 resolution.

Model	HPSv2.1 $\uparrow$	ImageReward $\uparrow$	PickScore $\uparrow$
FLUX.1-dev	32.02	1.0790	23.32
BoN	33.70	1.2282	23.64
Ours+Turbo (in paper)	33.91	1.2513	23.69
Ours+DBCACHE	34.01	1.2542	23.69

Table 7: Results on video generation. Here, the video generation model we adopted is CogVideoX-2B, and the fast sampler is TDM.

	Original	BoN	Ours w/o Visual Reflection	Ours w/ Visual Reflection
NFEs	50	300	294	294
Visual Quality $\uparrow$	0.2339	0.4915	0.4859	0.5264

A.3.3 QUANTITATIVE RESULTS ON VIDEO GENERATION

We also conduct additional experiments applying our reflection mechanism to video modality and high-resolution image generation tasks. The preliminary results demonstrated in Table 7 indicate that our method can effectively generalize to these scenarios, demonstrating promising scalability and adaptability. Here, for video modality, we conduct experiments on CogVideoX-2B model, and the fast generator we adopt is TDM. The total number of inference steps for the original CogVideoX pipeline is 50, and the original TDM pipeline is 4. Here, $k = 25$ and $z = 2$. We randomly select 100 prompts from VBench all_dimension_extended dataset, and for each prompts we generate 6 videos for BoN and 10 videos for our method. Our method can successfully adapt to video modality.

A.3.4 MORE VISUALIZATION RESULTS

We also present more visualization results in Fig. 4 to demonstrate the effectiveness of our proposed method. Here, all images are of 1024×1024 , and are generated with prompts from COCO2014val dataset.

A.4 LIMITATIONS

Our proposed method is constrained by the capabilities of the underlying model. In this paper, we focus on enhancing the efficiency and performance of the process of inference time scaling, which is primarily concerned with exploring and leveraging the potential of the model, rather than enhancing the inherent performance of the model. If the base model lacks the ability to generate images that meet the desired requirements, the overall effectiveness of our method will be limited.

A.5 BROADER IMPACTS

Our proposed method SF2 has positive social impacts. Specifically, we propose a novel inference time scaling method, which enhances the efficiency and performance of the inference time scaling process through the incorporation of Co-Fusion pipeline and visual reflection. Our proposed method offers practical advantages by maximizing the utility of existing pretrained models, thereby reducing the need for resource-intensive retraining or the development of new models. Also, our proposed method avoids redundant sampling process, which saves a lot of computational resources. By enabling more efficient and effective use of pretrained models, our proposed method supports the deployment of visual generation models in various real-world applications.

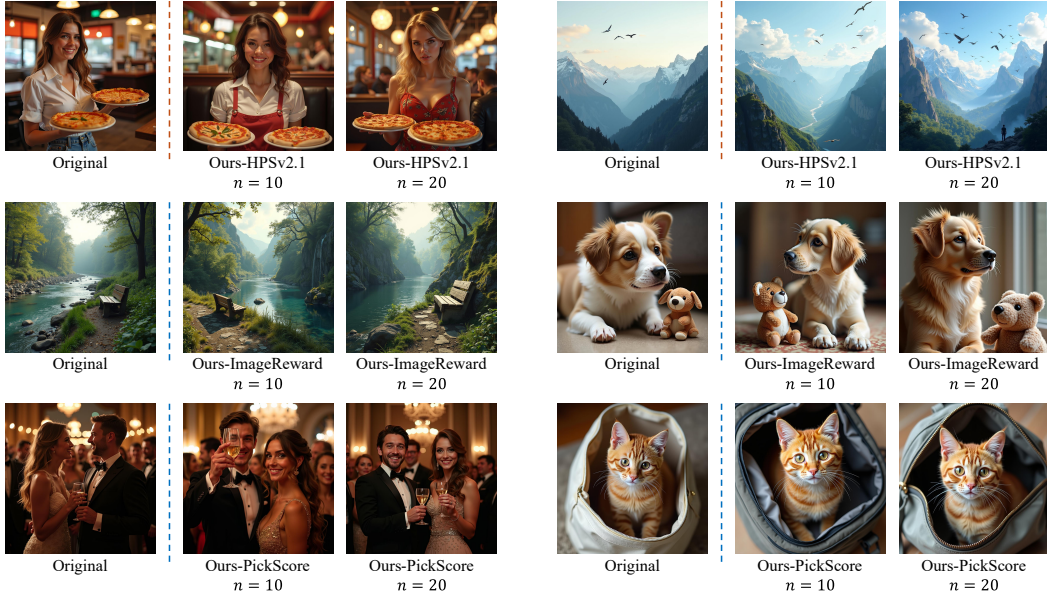
Stable-Diffusion-3.5-Large**FLUX.1-dev**

Figure 4: Visualization results of our proposed method. The images demonstrated here are of 1024×1024 , and generated with prompts from COCO2014val dataset.

A.6 ETHICS STATEMENT

This research follows the ethical guidelines established by the ICLR Code of Ethics. No experiments were conducted on human participants, and no private or sensitive data was collected or used. All datasets employed in our study are openly accessible and accompanied by appropriate licenses. While our work aims to advance the efficiency of diffusion model for inference-time scaling, we acknowledge that such techniques may be misapplied in harmful or malicious contexts. We therefore advocate for responsible use, in accordance with ethical norms and applicable legal frameworks.

A.7 REPRODUCIBILITY STATEMENT

To improve reproducibility, we provide thorough documentation of our methodology. The main content and Appendix detail the model architectures, hyperparameters, and experimental setups. All datasets are publicly released. Moreover, we provide pseudocode, inference scripts, and implementation details to simplify replication efforts. Our source code is also provided in the supplementary material to ensure faithful reproduction of results.

A.8 THE USE OF LARGE LANGUAGE MODELS

In this paper, large language models are only used to correct grammar and spelling errors.