Ultra-Resolution Adaptation with Ease

Ruonan Yu [*1] **Songhua Liu** [*1] **Zhenxiong Tan** [1] **Xinchao Wang** [1]

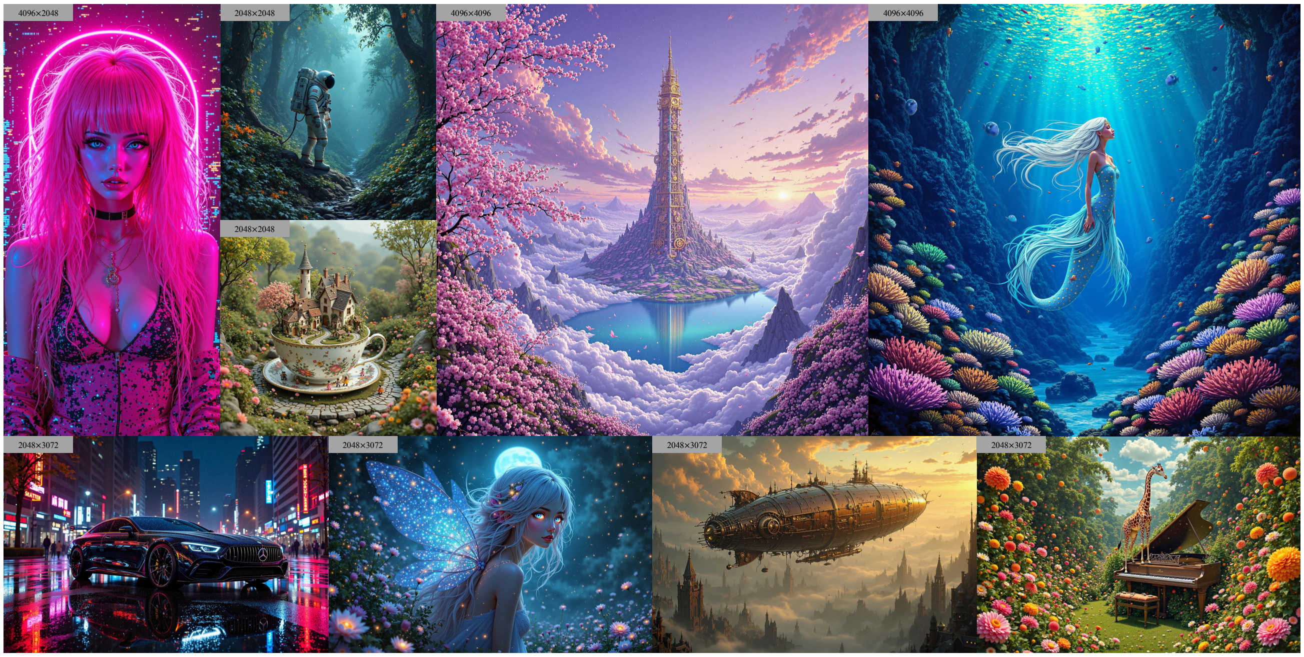

Figure 1: High-resolution results by our method.

Abstract

Text-to-image diffusion models have achieved remarkable progress in recent years. However, training models for high-resolution image generation remains challenging, particularly when training data and computational resources are limited. In this paper, we explore this practical problem from two key perspectives: data and parameter efficiency, and propose a set of key guidelines for ultra-resolution adaptation termed *URAE*. For data efficiency, we theoretically and empirically demonstrate that synthetic data generated by some teacher models can significantly promote training convergence. For parameter efficiency, we find that tuning minor components of the weight matrices outperforms widely-used low-rank adapters when synthetic data are unavailable, offering substantial performance gains while maintaining efficiency. Additionally, for models leveraging guidance distillation, such as FLUX, we show that disabling classifier-free guidance, *i.e.*, setting the guidance scale to 1 during adaptation, is crucial for satisfactory performance. Extensive experiments validate that URAE achieves comparable 2K-generation performance to state-of-the-art closed-source models like FLUX1.1 [Pro] Ultra with only 3K samples and 2K iterations, while setting new benchmarks for 4K-resolution generation. Codes are available here.

1. Introduction

Recent years have witnessed remarkable progress in text-to-image generation with diffusion models (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Rombach et al., 2022; Ho et al., 2020). From UNet-based architectures (Ronneberger

[*]Equal contribution [1]National University of Singapore, Singapore. Correspondence to: Xinchao Wang <xinchao@nus.edu.sg>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

et al., 2015; Rombach et al., 2022) to latest state-of-the-art Diffusion Transformers (DiTs) (Peebles & Xie, 2023; Bao et al., 2023; Chen et al., 2023; Esser et al., 2024; Li et al., 2024; Gao et al., 2024; Chen et al., 2024), these models leverage powerful backbones and multistep denoising schemes to generate high-quality and diverse images from textual prompts effectively, solidifying their leading position in this field (Croitoru et al., 2023; Yang et al., 2023a).

Nevertheless, extending current diffusion models to ultra-resolution generation, such as 4K, remains a significant challenge. The process typically demands massive amounts of high-quality data and substantial computational resources, making training at such resolutions daunting and accessible only to industry-scale efforts. Although recent attempts have been made to train 4K-resolution text-to-image models (Chen et al., 2024; Xie et al., 2024), they rely on internal datasets containing millions of high-resolution images to fine-tune base low-resolution models. In practice, collecting such large-scale datasets for training is highly cumbersome if not infeasible at all. Meanwhile, tuning the entire diffusion backbone introduces an intensive GPU memory footprint, especially for state-of-the-art models like FLUX (Black_Forest_Labs, 2023) and Stable Diffusion 3.5 (Esser et al., 2024).

Focusing on these drawbacks, we are curious about one practical question: *Can this ultra-resolution adaptation process be made easier?* In this paper, we answer the question positively by proposing *URAE*, a set of key guidelines, under which ultra-resolution adaptation is achievable with merely thousands of training samples and iterations.

Specifically, we initiate our exploration from two key aspects: data and parameter efficiency. On the one hand, we provide theoretical and empirical evidence that synthetic data produced by some teacher models can largely enhance training convergence. However, despite recent advancements in text-to-image generation, state-of-the-art models still face significant challenges in acquiring high-quality synthetic training data for ultra-resolution adaptation, such as 4K. We thus, on the other hand, investigate such scenarios where synthetic data are unavailable and identify that tuning minor components of the pre-trained weight matrices is more effective than commonly used parameter-efficient adaptation strategies like LoRA (Hu et al., 2022).

Furthermore, we delve into the principles of fine-tuning guidance-distilled models like FLUX and discover that disabling classifier-free guidance—by setting the guidance scale to 1—is essential, regardless of the availability of synthetic data. Backed up by the above guidelines, we conduct extensive experiments to demonstrate that URAE achieves performance comparable to state-of-the-art closed-source models like FLUX1.1 [Pro] Ultra with merely 3K training samples and 2K adaptation iterations. Meanwhile, it surpasses previous models in 4K generation performance and remains highly compatible with existing training-free high-resolution generation pipelines (Du et al., 2024b; Meng et al., 2021), enabling further performance improvements. In summary, the contributions of this paper are:

- We are the first to delve into the problem of ultra-resolution adaption to the best of our knowledge;
- We propose URAE, a set of key guidelines focusing on data efficiency, parameter efficiency, and classifier-free guidance, to facilitate the adaptation of existing text-to-image models to higher resolutions;
- We validate that URAE achieves comparable performance in 2K generation, superior capabilities in 4K generation, and strong compatibility with existing training-free high-resolution generation pipelines.

2. Related Works

2.1. Text-to-Image Diffusion Models

The diffusion model (Ho et al., 2020) has emerged as a powerful class of generative models. Unlike traditional approaches such as GANs (Goodfellow et al., 2014), diffusion models iteratively refine noisy maps with a UNet backbone (Ronneberger et al., 2015) to produce high-quality and detailed images (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021), which fuels significant advancements in large-scale text-to-image diffusion models (Rombach et al., 2022; Podell et al., 2023; Balaji et al., 2022; Ding et al., 2022; Nichol et al., 2021; Ramesh et al., 2022; Razzhigaev et al., 2023; Xu et al., 2023; Saharia et al., 2022). Leveraging billions of image-text pairs, they demonstrate remarkable semantic understanding and the ability to generate diverse and photorealistic images aligning with text prompts.

Most recently, Transformer (Vaswani et al., 2017) has been introduced as an alternative backbone to UNet (Peebles & Xie, 2022) in diffusion models, known as Diffusion Transformer (DiT). Then, text-to-image models based on it have progressively demonstrated dominant performance (Chen et al., 2024; Esser et al., 2024; Gao et al., 2024; Li et al., 2024; Zheng et al., 2024). We thus focus on DiT-based models and conduct experiments mainly on FLUX, which yields state-of-the-art text-to-image performance, in sake for superior ultra-resolution adaption results.

2.2. High-Resolution Generation

Training models at high resolutions demands substantial computational resources. To address this, a series of works propose training-free solutions, developing inference stage strategies that allow diffusion models trained at their native resolutions to operate effectively at higher scales (Bar-Tal et al., 2023; Meng et al., 2021; Du et al., 2024b; He et al., 2024; Du et al., 2024a; Huang et al., 2024; Wu et al., 2024; Zhang et al., 2023). While effective, without looking at any high-resolution images during training, in fact, they still fall short in accurately handling detailed structures and textures inherent in ultra-resolution images. By contrast, ultra-resolution adaptation focused in this paper dedicates on addressing this drawback through training, which is technically orthogonal to training-free approaches and can work as a plug-and-play component to enhance their performance.

There are indeed some works training for high-resolution generation like 4K (Chen et al., 2024; Xie et al., 2024; Zheng et al., 2024; Ren et al., 2024). However, millions of high-quality training data and industrial-scale computational resources are required to train the whole transformer backbone. In this paper, we focus on the challenges of data and parameter efficiency and demonstrate that comparable or even superior performance can be achieved with significantly less data and fewer trainable parameters.

Another line of research has concentrated on enhancing inference efficiency through the development of efficient and scalable diffusion backbones (Chen et al., 2024; Liu et al., 2024b;a). These designs and insights are orthogonal to our work, and it is promising to combine their strengths with our approach to achieve the best of both training and inference efficiency, which lies beyond the scope of this paper and is left for future exploration.

2.3. Parameter-Efficient Fine-Tuning

In many real-world scenarios, fine-tuning existing models for specific applications is often necessary. However, fine-tuning all parameters can lead to substantial computational overhead, particularly in terms of memory footprint. To address this limitation, a series of works propose parameter-efficient fine-tuning strategies (Hu et al., 2022; Hyeon-Woo et al., 2021; Meng et al., 2024; Yeh et al., 2023; Wang et al., 2024a). In this paper, we aim at an effective method specifically tailored for ultra-resolution adaptation.

3. Methodology

In this section, we delve into the motivations and technical details of URAE, our proposed strategy for ultra-resolution adaptation. We begin with some preliminary concepts, followed by three key components including training with synthetic data, parameter-efficient fine-tuning strategies, and classifier-free guidance.

3.1. Preliminary

State-of-the-art text-to-image diffusion models commonly adopt the flow matching training scheme (Esser et al., 2024; Lipman et al., 2022). Specifically, in each iteration, a batch of images x and their corresponding textual descriptions y is sampled. These images are then encoded into a latent map z_0 using a pre-trained VAE encoder, and a noise map ϵ is drawn from a Gaussian distribution. Let z_t denote the noisy version of z_0 after applying ϵ at the t-th diffusion timestep. The flow matching loss is then formulated as:

$$\mathcal{L}_{fm}(z_0, y, t, \epsilon) = \|(\epsilon - z_0) - \epsilon_\theta(z_t, t, y)\|_2^2, \quad (1)$$

where $\epsilon_\theta(\cdot)$ is the denoising backbone with parameters θ.

The inference process begins with a text prompt y and a random Gaussian noise ϵ, also denoted z_T, which is iteratively denoised using the trained backbone. After T steps, the resulting z_0 represents a clean sample in the latent space and is then decoded to a generated image x using a pre-trained VAE decoder. Such training and inference paradigms are also employed in our URAE framework.

3.2. Synthetic Data or Real Data?

Previous works train 4K-generation models using millions of high-quality training images (Chen et al., 2024; Xie et al., 2024), leading to significant challenges in collecting, transmitting, storing, and processing such large volumes of data. To alleviate these inconveniences, we target a data-efficient approach for ultra-resolution adaptation.

Building on recent advances in the distillation of diffusion models (Yang et al., 2023b; Kim et al., 2023; Liu et al., 2024b;a), we recognize that incorporating a teacher model for reference and a loss term for knowledge distillation (Hinton, 2015) can enhance training:

$$\mathcal{L}_{distill}(z_0, y, t, \epsilon) = \|\epsilon_\theta(z_t, t, y) - \epsilon_{\theta_{ref}}(z_t, t, y)\|_2^2, \quad (2)$$

where θ_{ref} represents the parameters of the teacher model. However, this approach relies on access to the diffusion backbone of the teacher model to compute the step-wise distillation loss, which is impractical for closed-weight models such as FLUX1.1 [Pro] Ultra. We therefore experiment with an alternative approach that optimizes the vanilla flow matching loss defined in Eq. 1 using data synthesized by the teacher model. We expect this to yield similar training benefits, as validated by the following theoretical analysis.

Before presenting our main results, we first set up some necessary assumptions. From the model perspective:

Assumption 3.1. Let u denote the input data pair (ϵ, y). The process from u to the output x can be characterized by a neural network $f(u; W)$ with infinite width, where W denotes the network's parameters.

It has been demonstrated that neural networks with a single hidden layer and sufficient width can approximate any complex functions (Cybenko, 1989; Hornik, 1991), and that infinite-width neural networks trained with gradient descent are equivalent to training linear models in the space of Neural Tangent Kernel (NTK) (Jacot et al., 2018). From the data perspective:

Assumption 3.2. The dataset contains both real and synthetic data, with the synthetic portion denoted as $p \in [0, 1]$. The total number of data is N. Given an input u, the corresponding target x in the real data distribution is characterized by $x_{real} = f(u; W^*) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$ is a scaler from a noise distribution and W^* denotes the optimal parameters, serving as an unknown oracle. For synthetic data, the target is given by $x_{syn} = f(u; W_{ref})$, where W_{ref} denotes the parameters of a pre-trained teacher model for reference. Without loss of generality, we restrict our discussion to the one-dimensional target case.

And from the training perspective:

Assumption 3.3. The model's parameters are initialized as W_0. The training is conducted through SGD using the square error as the loss function: $\mathcal{L} = \frac{1}{2N}\|f(u; W) - x\|_2^2$. The learning rate is η.

Our main results are summarized in Theorem 3.4:

Theorem 3.4. *Under the setting defined in Assumptions 3.1, 3.2, and 3.3, the error between W_T, the parameters after T training iterations, and the optimal W^* is bounded by:*

$$\begin{aligned}&\mathbb{E}[\|W_T - W^*\|_2^2] \leq \mathbb{E}[\|(I - \eta M)^T \Delta_0\|_2^2] + \\ &\eta^2(p(1-p)\mathbb{E}[\delta^2] + (1-p)\sigma^2) \sum_{i=1}^{N} \frac{(1 - (1 - \eta\lambda_i)^T)^2}{\lambda_i} \\ &+ p^2\|W_{ref} - W^*\|_2^2.\end{aligned} \tag{3}$$

where $\Delta_0 = W_0 - (pW_{ref} + (1-p)W^*)$, M is defined as $\nabla_W f(U; W_0)^\top \nabla_W f(U; W_0)$, $\delta = f(u; W_{ref}) - f(u; W^*)$, and λ_i is the i-th eigenvalue of M.

The proof can be found in Appendix A. Intuitively, Theorem 3.4 indicates that, when training data are drawn from a mixture of real and synthetic data points, the distance to the optimal parameters at convergence reflects a trade-off—governed by the synthetic portion p—between two factors: the error introduced by label noise in the real data distribution and the discrepancy between the reference model used for generating synthetic data and the optimal model, shown in the 2nd and 3rd terms of Eq. 3.4 separately. Fig. 2 provides an illustrative example to visualize this effect.

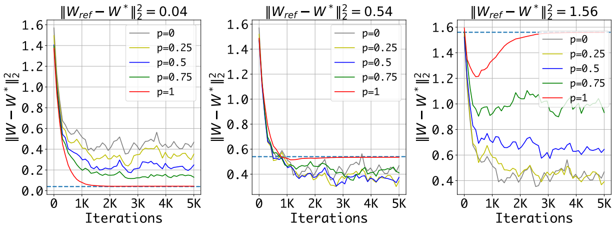

Figure 2: A toy linear regression case. There are real data with noisy labels and synthetic data generated by a reference model W_{ref}. The proportion of synthetic data is p.

Closely examining Eq.3.4, we can discover that, by diminishing label noise, synthetic data would be helpful if the reference model providing these data is accurate. Although some works reveal that synthetic data can result in model collapse (Dohmatob et al., 2024b;a) by amplifying the gap between real and synthetic distributions, we demonstrate that they are useful particularly for ultra-resolution adaption. On one hand, large-scale real datasets such as LAION-5B (Schuhmann et al., 2022) tend to be noisy, containing numerous low-quality images and mismatched text-image pairs. On the other hand, at 2K resolution, models like FLUX-1.1 [Pro] Ultra—although closed-weight—are available to produce high-quality synthetic data. Building on this analysis, we train our 2K-generation model using only synthetic data in this work, demonstrating superior performance across various scenarios.

3.3. Tune Major or Minor Components?

Parameter-efficient fine-tuning strategies enable adapting a pre-trained model from its original domain to a target domain by integrating lightweight adapters. For instance, in personalized text-to-image generation such as DreamBooth (Ruiz et al., 2023), attaching low-rank, e.g., rank $r = 4$, adapters, *i.e.*, LoRA, to the original model's weights can achieve satisfactory performance (Hu et al., 2022). Specifically, this is achieved by:

$$Y = XW + XAB, \tag{4}$$

where X, Y, and W are input, output, and original weight matrices respectively, $A \in \mathbb{R}^{c_{in} \times r}$ and $B \in \mathbb{R}^{r \times c_{out}}$ are low-rank matrices for adaptation, and c_{in} and c_{out} are input and output dimensions, output dimension separately. In practice, A is initialized using a normal distribution, whereas B is set to all zeros, which makes the adapter branch output zero initially and allows tuning to begin from the original parameters. After tuning, A and B can be merged into the original weight matrix via $W' = W + AB$, ensuring that the total number of model parameters remains unchanged. These low-rank adapters employ a small number of parameters that focus on the major components with the largest singular values, enabling efficient adaptation to the target domain (Meng et al., 2024).

However, different from DreamBooth modifying the styles and appearances of output images, ultra-resolution adaptation focuses on learning the arrangements of details and local textures, which may not correspond to the major components in weight matrices. Under this hypothesis, we introduce a method to tune the components associated with the smallest singular values instead.

Specifically, given a weight matrix $W \in \mathbb{R}^{c_{in} \times c_{out}}$ and $c = \min(c_{in}, c_{out})$, we first conduct Singular Value Decomposition (SVD) and derive $W = U\Sigma V$, where $U \in \mathbb{R}^{c_{in} \times c}$ and $V \in \mathbb{R}^{c \times c_{out}}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{c \times c}$ is a diagonal matrix with the singular values arranged from large to small. Then, r components with the smallest singular values and the rest $c - r$ ones are extracted via:

$$\begin{aligned} W^{small} &= U[:, -r:]\Sigma[-r:, -r:]V[-r:, :], \\ W^{res} &= U[:, :-r]\Sigma[:-r, :-r]V[:-r, :], \end{aligned} \tag{5}$$

where the indexing syntax in Numpy (Harris et al., 2020) and PyTorch (Paszke et al., 2019) are used to represent the operations for extracting multiple rows/columns. Analyzing from Eq. 5, W^{small} is a low-rank matrix. Therefore, for parameter efficiency, we formulate the training time behavior similar to Eq. 4:

$$\begin{aligned} Y &= XW^{res} + XAB, \\ A &= U[:, -r:]\sqrt{\Sigma[-r:, -r:]}, \\ B &= \sqrt{\Sigma[-r:, -r:]}V[-r:, :]. \end{aligned} \tag{6}$$

A and B are initialized using Eq. 6 and updated during fine-tuning. In terms of formulation, the approach is similar to PISSA (Meng et al., 2024); however, it fundamentally differs by tuning the components with the smallest singular values instead of the largest. Although (Wang et al., 2024a) introduce a similar approach in the field of large language model finetuning, they fail to analyze its applicability in various scenarios.

Empirically, we observe that this approach is particularly effective when no synthetic data are available to serve as a reliable reference, *e.g.*, in 4K generation. We speculate that the effectiveness stems from preserving the major components in the original weight matrices, thereby safeguarding the model's capacity to handle semantics, layouts, and appearances from label noise in the real data distribution.

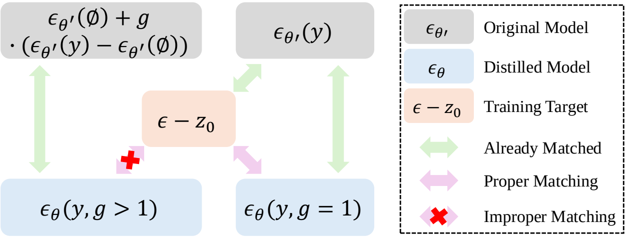
$\epsilon_{\theta'}(\emptyset) + g \cdot (\epsilon_{\theta'}(y) - \epsilon_{\theta'}(\emptyset))$
$\epsilon_{\theta'}(y)$
$\epsilon - z_0$
$\epsilon_\theta(y, g > 1)$
$\epsilon_\theta(y, g = 1)$
$\epsilon_{\theta'}$ Original Model
ϵ_θ Distilled Model
$\epsilon - z_0$ Training Target
Already Matched
Proper Matching
Improper Matching

Figure 3: For CFG-distilled models, classifier-free guidance should be disabled in the training time. z_t and t are omitted from the inputs of ϵ_θ and $\epsilon_{\theta'}$ here for simplicity.

3.4. Enable or Disable Classifier-Free Guidance?

Classifier-free guidance (CFG) (Ho & Salimans, 2022) aims to enhance the quality of the generated samples by introducing an additional "null-condition" branch. Specifically, at each denoising step in the inference time, the current latent map is processed by both the main branch and the null-condition branch, and the final output is then guided, with a certain strength, in the direction opposing the null-condition branch's prediction:

$$\epsilon_\theta(z_t, t, \emptyset) + g \cdot (\epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, \emptyset)), \tag{7}$$

where $\emptyset$ denotes the null condition, *e.g.*, an empty prompt, and g is a hyper-parameter controlling the strength.

Although effective, the additional null-condition branch doubles the inference cost. To address this issue, models like FLUX.1-dev use guidance distillation to train a distilled model that takes the CFG scale embedding as an additional input, encouraging its output aligns with the result in Eq. 7. Since g is typically larger than 1 in inference, in training, many works also set g to the same value used at inference time during fine-tuning (XLabs-AI, 2024; TencentARC, 2024). However, according to the experiments, it results in inferior performance, especially in the problem of ultra-resolution adaptation.

Specifically, during the distillation stage, the distilled model is trained with $g > 1$ as input Eq. 7, which involves the null condition. In contrast, during the adaptation stage, the target is $\epsilon - z_0$ defined in Eq. 1, which is irrelevant to the null condition. If g remains larger than 1, a mismatch arises between the training targets in these two stages, making the training process more challenging.

To address this issue, we note that incorporating the null-condition branch during adaptation is unnecessary; simply disabling CFG at training time by setting $g = 1$ works well and yields a consistent target across the two stages. Fig. 3 illustrates the mismatch triggered by $g > 1$ and how $g = 1$ addresses the problem.

During inference, CFG is still necessary by using $g > 1$. Although the model does not encounter $g > 1$ during adaptation, we find that it generalizes sufficiently well in practice.

Table 1: Quantitative results of the baseline methods and our proposed guidelines. The prompts are from HPD and DPG datasets. All images are at a resolution of 2048 × 2048. Here, FLUX.1-dev* is FLUX.1-dev with scaled RoPE, proportional attention, and removing dynamic shifting strategies.

Method/Setting	HPD Prompt						DPG Prompt
	FID (↓)	LPIPS (↓)	MAN-IQA (↑)	QualiCLIP (↑)	HPSv2.1 (↑)	PickScore (↑)	DPG Bench (↑)
FLUX1.1 [Pro] Ultra	-	-	0.4129	0.6424	29.61	22.99	84.76
Real-ESRGAN	36.25	0.6593	0.4653	0.6392	30.70	22.91	83.50
SinSR	35.09	0.6566	0.4194	0.5556	30.95	22.96	83.79
SDEdit	35.59	0.6456	0.3736	0.4480	30.92	22.86	83.56
w/ URAE	**34.07**	**0.6419**	**0.3872**	**0.5800**	**32.26**	**23.02**	**84.61**
I-Max	33.66	0.6394	0.3670	0.4797	31.12	23.02	83.92
w/ URAE	**32.24**	**0.6357**	**0.3833**	**0.5736**	**32.37**	**23.18**	**87.88**
PixArt-Sigma-XL	36.58	0.6801	0.2949	0.4438	30.66	22.92	80.60
Sana-1.6B	33.17	0.6792	0.3695	0.6718	30.92	22.83	85.14
FLUX.1-dev	43.78	0.6530	0.3821	0.3800	26.22	21.54	80.64
FLUX.1-dev*	34.86	0.6036	0.4110	0.5468	28.73	22.68	80.15
w/ URAE	**29.44**	**0.5965**	**0.4730**	**0.7191**	**31.15**	**23.15**	**83.83**

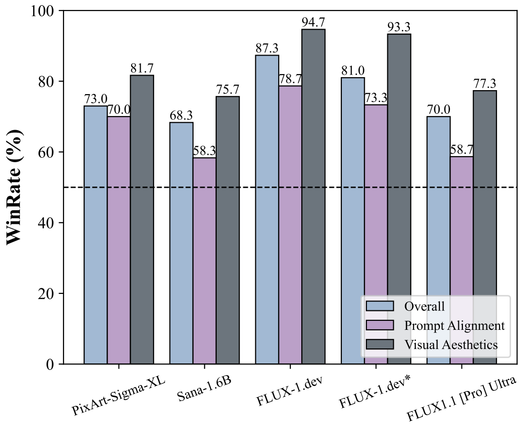

Figure 4: GPT-4o preferred evaluation against current SOTA T2I models. We request GPT-4o to select a better image regarding overall quality, prompt alignment, and visual aesthetics. Our proposed method are preferred against others.

4. Experiments

4.1. Settings and Implementation Details

In this paper, we adopt the open-source text-to-image FLUX.1-dev model (Black_Forest_Labs, 2023) as the base model to demonstrate the effectiveness of our proposed URAE guidelines, thanks to its superior performance. For our 2K-generation model, we collect 3K synthetic samples with various aspect ratios generated by the FLUX1.1 [Pro] Ultra model as the training dataset, and fine-tune the FLUX.1-dev on it for merely 2K iterations with a batch size of 8, which takes only ∼ 1 day on 2 H100 GPUs. For our 4K model, we utilize 30K images with at least 4K resolution from the LAION-5B dataset (Schuhmann et al., 2022) and fine-tune the base model FLUX.1-dev for 2K iterations on 8 H100 GPUs, which takes ∼ 1 days. In terms of training convergence, our method requires significantly fewer iterations compared with state-of-the-art methods, such as 10K for SANA (Xie et al., 2024).

For baseline, we apply URAE on the FLUX.1-dev model and compare the performance with PixArt-Sigma-XL (Chen et al., 2024), Sana-1.6B (Xie et al., 2024), and FLUX series models. In order to further demonstrate the effectiveness of URAE, we also apply URAE to the existing training-free high-resolution generation pipelines, *i.e.*, SDEdit (Meng et al., 2021) and I-Max (Du et al., 2024b). These pipelines require the base text-to-image model, *e.g.*, FLUX.1-dev, to generate low-resolution, *e.g.*, 1024 × 1024, images as the guidance, and upscale these images to higher resolutions through image-to-image pipelines. For comparison, we also include the super-resolution methods Real-ESRGAN (Wang et al., 2021) and SinSR (Wang et al., 2024b), based on GAN and diffusion, respectively. We conduct quantitative experiments on 2048 × 2048 samples generated with prompts from the HPD (Wu et al., 2023) and DPG (Hu et al., 2024) datasets. We evaluate FID and STLPIPS (Ghildyal & Liu, 2022) with the reference dataset generated by the FLUX1.1 [Pro] Ultra model. Furthermore, the quality of the generated images is assessed using metrics MAN-IQA (Yang et al., 2022) and QualiCLIP (Agnolucci et al., 2024). We adopt DPG Bench to measure the semantic consistency and coherence between the generated image and the corresponding prompt. Additionally, we use the HPSv2.1 (Wu et al., 2023) and PickScore (Kirstain et al., 2023) as human preference

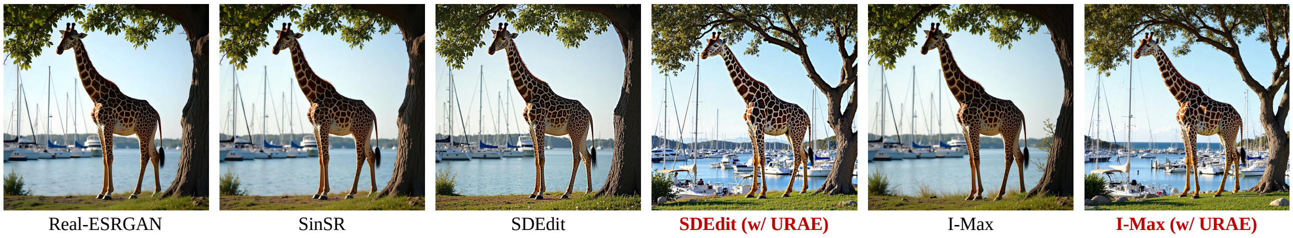

Figure 5: Visualizations of our proposed method apply to training-free high-resolution generation pipelines. The prompt is *A giraffe stands beneath a tree beside a marina.*

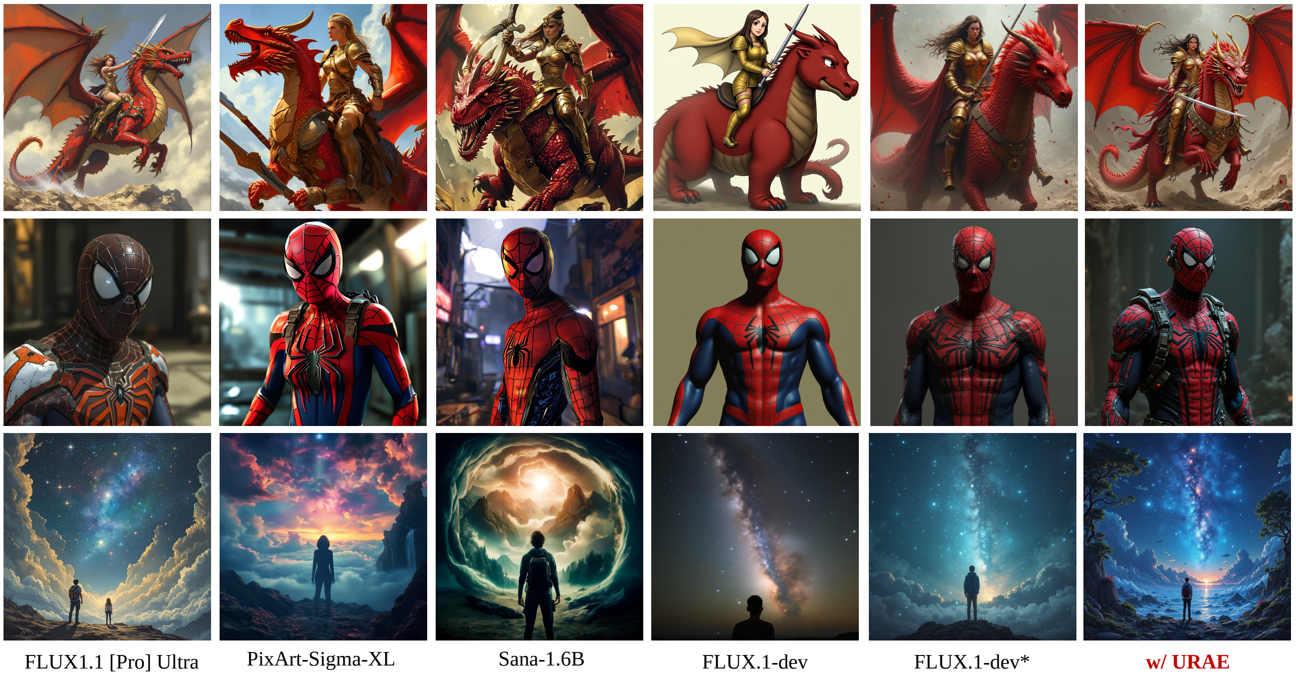

Figure 6: Qualitative comparisons with baseline methods. All the images are of 2048 × 2048 size.

metrics to further evaluate the quality and aesthetic appeal of the generated images. Following prior works like PixArt-Sigma and I-Max, we also utilize the GPT-4o to assess the generated images from prompt alignment, visual aesthetics, and overall quality key perspectives, at both 2K and 4K resolutions. These AI preference scores are derived from 300 randomly selected prompts in the COCO30K (Lin et al., 2014; Chen et al., 2024) dataset.

4.2. 2K Resolution

Here, we evaluate the performance of the proposed methods on 2K images generated with prompts from HPD and DPG datasets. The results are shown in Table 1. The quantitative results indicate that our proposed method is capable of significantly enhancing the ability of models to generate high-resolution images and demonstrates its versatility and adaptability across different methods. Our method surpasses the state-of-the-art model FLUX1.1 [Pro] Ultra in terms of image quality, demonstrating its superiority in generating visually refined images. Moreover, the remarkable improvements in image quality further underscore the strength of our method in achieving state-of-the-art visual results for all quality metrics, making it a highly effective solution for high-resolution image generation tasks. In addition, our method also achieves a substantial improvement in the performance of the base model in terms of prompt alignment, improving the original FLUX.1-dev by 3.19 in DPG Bench score, and 3.96 for I-Max pipeline.

For human preference study, we adopt HPSv2.1 and PickScore to benchmark the human preference score. The samples are generated with prompts from the HPD dataset and the resolution is 2048 × 2048. The results are shown in Table 1. The results show that our method improves the human preference score of the base model, indicating that

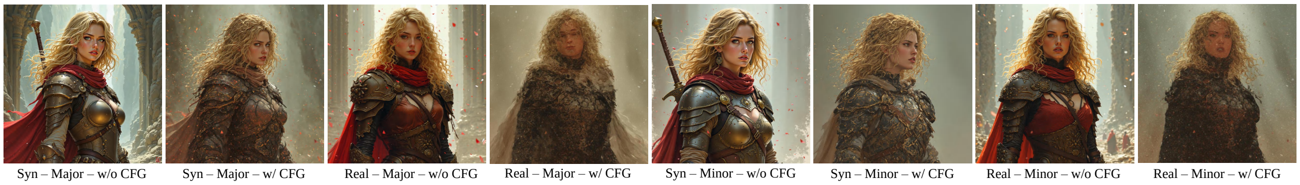

Figure 7: Visualization results of ablation studies. The prompt is *Imogen Poots portrayed as a D&D Paladin in a fantasy concept art by Tomer Hanuka*.

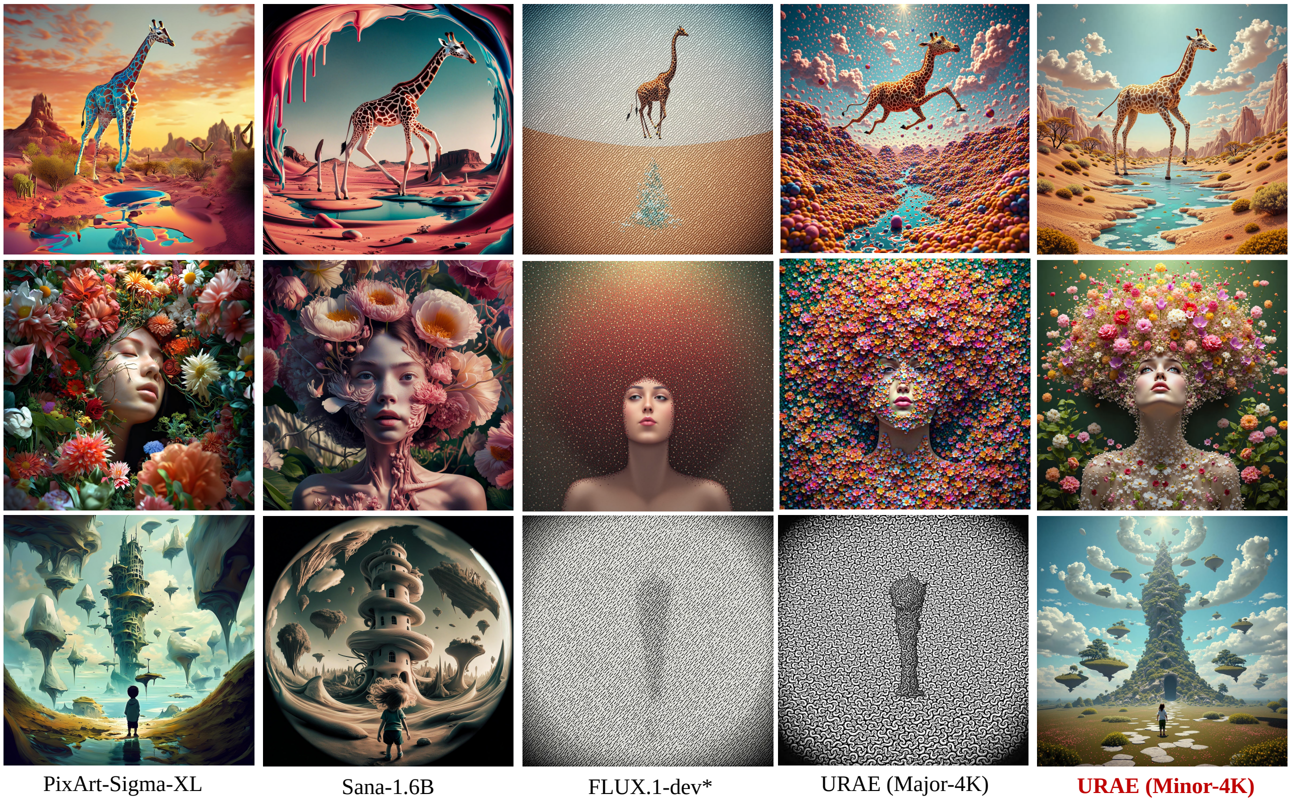

Figure 8: Visualization results for ultra-resolution image generation task. All the images are of 4096 × 4096 size.

our proposed guidelines are capable of generating images that better align with human preferences.

We also conduct AI preference studies with GPT-4o for pair comparison regarding overall quality, prompt alignment, and visual aesthetics aspects. The results are shown in Fig. 4. For the prompts for GPT-4o to assess the quality, prompt alignment, and visual aesthetics, please refer to Appendix B.1. The results demonstrate that our proposed method excels and is preferred in all three aspects. Please refer to Appendix C.1 for more quantitative results.

4.3. 4K Resolution

Here, we evaluate the performance of our proposed method in 4K-ultra-resolution image generation. The results are shown in Table 3. From the experimental results, fine-tuning minor components achieves outstanding performance when no synthetic data are available to serve as a reliable reference for the ultra-resolution image generation task, while the commonly adopted LoRA may fail on the overall semantics. Moreover, our proposed method also demonstrates exceptional performance compared with other methods, further validating its competitiveness to existing approaches.

Given that there is no well-defined benchmark specifically for 4K-ultra-resolution generation, we conduct a user study using randomly generated prompts listed in Appendix B.2. For each prompt, we present results generated by the candidates to users and let them select the best one from three aspects including overall quality, prompt alignment, and visual aesthetic aspects. We collect $1,020$ votes altogether. The results in Table 3 are coherent with the above analysis.

Table 2: Ablation studies on three key guidelines, using real (Real) or synthetic (Syn) data, whether to adopt CFG in training, and tuning of major or minor components. Evaluations are on the 2048 × 2048 images generated from HPD prompts.

	Method/Setting	FID (↓)	LPIPS (↓)	MAN-IQA (↑)	QualiCLIP (↑)	HPSv2.1 (↑)	PickScore (↑)
Major	Syn w/o CFG	29.44	0.5965	**0.4730**	**0.7191**	**31.15**	**23.15**
	Syn w/ CFG	76.07	0.6388	0.3992	0.5890	24.80	21.87
	Real w/o CFG	31.39	0.6076	0.4262	0.5953	29.33	22.86
	Real w/ CFG	133.68	0.5978	0.3254	0.3645	16.55	19.92
Minor	Syn w/o CFG	**27.90**	**0.5779**	0.4558	0.6616	30.40	22.87
	Syn w/ CFG	65.34	0.5858	0.3852	0.5312	23.77	21.64
	Real w/o CFG	32.09	0.6000	0.4485	0.6098	28.71	22.61
	Real w/ CFG	133.32	0.6026	0.3387	0.3672	16.36	19.68

Table 3: Evaluation on ultra resolution image generation task. The images are of 4096 × 4096, and generated with prompts randomly selected from COCO30K. For user study, the prompts are randomly generated as listed in Appendix B.2.

Method/Setting	**COCO30K Prompt**				**User Study**					
	MAN-IQA (↑)	Rank	QualiCLIP (↑)	Rank	Overall Quality	Rank	Prompt Alignment	Rank	Visual Aesthetic	Rank
PixArt-Sigma-XL	0.2935	5	0.2308	5	31.18%	2	30.88%	2	30.00%	2
Sana-1.6B	0.3288	3	0.4979	2	10.29%	3	10.00%	3	12.06%	3
FLUX.1-dev*	0.3673	2	0.2564	4	3.24%	4	3.24%	4	3.24%	4
w/ URAE (Major-4K)	0.3280	4	0.2700	3	2.06%	5	1.76%	5	1.76%	5
w/ URAE (Minor-4K)	**0.3999**	1	**0.5118**	1	**53.24%**	1	**54.12%**	1	**52.94%**	1

4.4. Ablation Study

To evaluate the effectiveness of our proposed method, we carry out ablation studies on three key guidelines that we proposed in URAE, the source of training data, tuning major or minor components, and the adoption of CFG in training. The experiments are conducted on the 2048 × 2048 images generated from HPD prompts. The results are shown in Table 2, and the visualization examples are shown in Fig. 7. For the source of training data, the results demonstrate that high-quality synthetic data can provide better performance than noisy real data. When the model is fine-tuned with real data, tuning minor components can bring more vivid details as shown in Fig. 7. As for CFG, although it is necessary in the inference stage, it can lead to significant performance degradation in the training stage. According to these results, we by default use synthetic data and tune major components, *i.e.* adopt LoRA, at 2K resolution, while use real data and tune minor components at 4K resolution. In both cases, CFG is disabled in training and enabled during inference.

5. Conclusions and Limitations

In this paper, we focus on the challenge of adapting text-to-image diffusion models from their native scales to ultra-resolution settings with limited training data and computational resources. Our proposed framework, URAE, tackles the problem from two complementary perspectives, *i.e.*, data and parameter efficiency, and provides a set of useful guidelines. First, by incorporating synthetic data generated by some teacher models, we demonstrate the potential to promote training convergence and achieve high-quality outcomes even under data-scarce conditions. Second, for the cases where synthetic data are unavailable, we introduce a parameter-efficient fine-tuning approach to tune the minor components of weight matrices, which outperforms standard low-rank adapters. Additionally, for models employing guidance distillation, e.g., FLUX, setting the guidance scale to 1 during adaptation proves crucial for achieving favorable results. Extensive experiments reveal that URAE matches the 2K-generation performance of leading closed-source solutions such as FLUX1.1 [Pro] Ultra using only 3K samples and 2K iterations, and further sets new milestones for 4K-resolution generation.

Nevertheless, the models presented in this work fall short of matching the inference-time efficiency exhibited by recent high-resolution text-to-image generation methods (Xie et al., 2024; Liu et al., 2024b;a; Chen et al., 2024), as we have not introduced architectural optimizations specifically targeting this aspect. In the future, we envision research aimed at streamlining the ultra-resolution generation process to balance quality and efficiency requirements in practice. It is also meaningful to integrate our methods into multi-modal large language models to unlock even broader and more versatile capabilities.

Acknowledgements

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE- T2EP20122-0006) and partially supported by NUS IT's Research Computing group using grant numbers NUSREC-HPC-00001.

Impact Statement

Our work on ultra-resolution text-to-image generation has broad implications for both research and real-world applications. By pushing resolution boundaries, we enable richer and more detailed visual content, benefiting domains such as digital art, virtual reality, advertising, and scientific visualization. At the same time, these advancements highlight important ethical considerations. High-fidelity images could be misused for creating deceptive content, and the computational demands of large-scale generation can have environmental impacts. We therefore emphasize responsible development practices, including efficient training strategies and transparent model documentation, to help ensure that the benefits of ultra-resolution text-to-image models are realized while mitigating potential risks.

References

Agnolucci, L., Galteri, L., and Bertini, M. Quality-aware image-text alignment for real-world image quality assessment. *arXiv preprint arXiv:2403.11176*, 5(6), 2024.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 2023.

Bar-Tal, O., Yariv, L., Lipman, Y., and Dekel, T. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*, pp. 1737–1752. PMLR, 2023.

Black_Forest_Labs. Flux. https://github.com/black-forest-labs/flux, 2023.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., and Li, Z. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Ding, M., Zheng, W., Hong, W., and Tang, J. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.

Dohmatob, E., Feng, Y., and Kempe, J. Model collapse demystified: The case of regression. *arXiv preprint arXiv:2402.07712*, 2024a.

Dohmatob, E., Feng, Y., Subramonian, A., and Kempe, J. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024b.

Du, R., Chang, D., Hospedales, T., Song, Y.-Z., and Ma, Z. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6159–6168, 2024a.

Du, R., Liu, D., Zhuo, L., Qi, Q., Li, H., Ma, Z., and Gao, P. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. *arXiv preprint arXiv:2410.07536*, 2024b.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Gao, P., Zhuo, L., Lin, Z., Liu, C., Chen, J., Du, R., Xie, E., Luo, X., Qiu, L., Zhang, Y., et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

Ghildyal, A. and Liu, F. Shift-tolerant perceptual similarity metric. In *European Conference on Computer Vision*, pp. 91–107. Springer, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., and Shan, Y. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Hinton, G. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., and Yu, G. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Huang, L., Fang, R., Zhang, A., Song, G., Liu, S., Liu, Y., and Li, H. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. *arXiv preprint arXiv:2403.12963*, 2024.

Hyeon-Woo, N., Ye-Bin, M., and Oh, T.-H. Fedpara: Low-rank hadamard product for communication-efficient federated learning. *arXiv preprint arXiv:2108.06098*, 2021.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Kim, B.-K., Song, H.-K., Castells, T., and Choi, S. Bk-sdm: Architecturally compressed stable diffusion for efficient text-to-image generation. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*, 2023.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al. Hunyuandit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. 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.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, S., Tan, Z., and Wang, X. Clear: Conv-like linearization revs pre-trained diffusion transformers up. *arXiv preprint arXiv:2412.16112*, 2024a.

Liu, S., Yu, W., Tan, Z., and Wang, X. Linfusion: 1 gpu, 1 minute, 16k image. *arXiv preprint arXiv:2409.02097*, 2024b.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Meng, F., Wang, Z., and Zhang, M. Pissa: Principal singular values and singular vectors adaptation of large language models. *arXiv preprint arXiv:2404.02948*, 2024.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., and Dimitrov, D. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

Ren, J., Li, W., Chen, H., Pei, R., Shao, B., Guo, Y., Peng, L., Song, F., and Zhu, L. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

TencentARC. Fluxkits. `https://github.com/TencentARC/FluxKits`, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, H., Li, Y., Wang, S., Chen, G., and Chen, Y. Milora: Harnessing minor singular components for parameter-efficient llm finetuning. *arXiv preprint arXiv:2406.09044*, 2024a.

Wang, X., Xie, L., Dong, C., and Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.-P., Liu, Z., Qiao, Y., Kot, A. C., and Wen, B. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024b.

Wu, H., Shen, S., Hu, Q., Zhang, X., Zhang, Y., and Wang, Y. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. *arXiv preprint arXiv:2408.11001*, 2024.

Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., and Li, H. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Xie, E., Chen, J., Chen, J., Cai, H., Lin, Y., Zhang, Z., Li, M., Lu, Y., and Han, S. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

XLabs-AI. x-flux. `https://github.com/XLabs-AI/x-flux`, 2024.

Xu, X., Wang, Z., Zhang, G., Wang, K., and Shi, H. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7754–7765, 2023.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023a.

Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., and Yang, Y. 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.

Yang, X., Zhou, D., Feng, J., and Wang, X. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 22552–22562, 2023b.

Yeh, S.-Y., Hsieh, Y.-G., Gao, Z., Yang, B. B., Oh, G., and Gong, Y. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In *The Twelfth International Conference on Learning Representations*, 2023.

Zhang, S., Chen, Z., Zhao, Z., Chen, Z., Tang, Y., Chen, Y., Cao, W., and Liang, J. Hidiffusion: Unlocking high-resolution creativity and efficiency in low-resolution trained diffusion models. *arXiv preprint arXiv:2311.17528*, 2023.

Zheng, W., Teng, J., Yang, Z., Wang, W., Chen, J., Gu, X., Dong, Y., Ding, M., and Tang, J. Cogview3: Finer and faster text-to-image generation via relay diffusion. *arXiv preprint arXiv:2403.05121*, 2024.

A. Theoretical Proof

We supplement the proof of our main theoretical results in Theorem 3.4 here.

Theorem A.1. *Under the setting defined in Assumptions 3.1, 3.2, and 3.3, the error between W_T, the parameters after T training iterations, and the optimal W^* is bounded by:*

$$\begin{aligned}\mathbb{E}[\|W_T - W^*\|_2^2] \leq & \mathbb{E}[\|(I-\eta M)^T \Delta_0\|_2^2] + \eta^2(p(1-p)\mathbb{E}[\delta^2] + (1-p)\sigma^2)\sum_{i=1}^{N}\frac{(1-(1-\eta\lambda_i)^T)^2}{\lambda_i} \\ & + p^2\|W_{ref} - W^*\|_2^2.\end{aligned} \tag{8}$$

where $\Delta_0 = W_0 - (pW_{ref} + (1-p)W^*)$, M is defined as $\nabla_W f(U;W_0)^\top \nabla_W f(U;W_0)$, $\delta = f(u;W_{ref}) - f(u;W^*)$, and λ_i is the i-th eigenvalue of M.

Proof. Under the assumption of infinite-width neural networks, the network output $f(u;W)$ can be viewed as the following linear form with respect to the parameter W:

$$f(u;W) \approx f(u;W_0) + \nabla_W f(u;W_0)(W - W_0). \tag{9}$$

We denote $\nabla_W f(W_0;u)$ as Φ for simplicity. According to the loss function $\mathcal{L} = \frac{1}{2N}\|f(u;W) - x\|_2^2$ and Eq. 9, the gradient of $\mathcal{L}$ with respect to W is:

$$\nabla_W \mathcal{L} = \frac{1}{N}\sum_N\{\Phi^\top(f(u;W) - x)\} = \frac{1}{N}\sum_N\{\Phi^\top[\Phi(W - W_0) + f(u;W_0) - x)]\}. \tag{10}$$

Training is conducted using SGD:

$$W_{t+1} = W_t - \eta\nabla_W\mathcal{L}_t = W_t - \eta\frac{1}{N}\sum_N\{\Phi^\top[\Phi(W_t - W_0) + f(u;W_0) - x)]\}. \tag{11}$$

Due to the linearity, the optimal parameter $W^{*'}$ when training on a mixture of real and synthetic data is given by:

$$W^{*'} = pW_{ref} + (1-p)W^*. \tag{12}$$

The target x can be viewed as the output of $f(u;W^{*'})$ with a noise term ξ':

$$x = f(u;W^{*'}) + \xi'. \tag{13}$$

Then we analyze the mean and variance of $\Sigma_{\xi'}$. According to Eq. 13, ξ' satisfies:

$$\xi' = \begin{cases} f(u;W_{ref}) - f(u;W^{*'}), & \text{with probability } p, \\ f(u;W^*) + \xi - f(u;W^{*'}), & \text{with probability } 1-p. \end{cases} \tag{14}$$

Given the linearity and Eq. 12,

$$f(u;W^{*'}) = pf(u;W_{ref}) + (1-p)f(u;W^*). \tag{15}$$

Denote $f(u;W_{ref}) - f(u;W^*)$ as δ. Then,

$$\xi' = \begin{cases} (1-p)\delta, & \text{with probability } p, \\ -p\delta + \xi, & \text{with probability } 1-p. \end{cases} \tag{16}$$

Since $\xi \sim \mathcal{N}(0,\sigma^2)$, the expectation of ξ' is:

$$\mathbb{E}[\xi'] = p(1-p)\delta + (1-p)(-p\delta) = 0. \tag{17}$$

And the variance, denoted as $\Sigma_{\xi'}$, is computed as:

$$\Sigma_{\xi'} = \mathbb{E}[\xi'^2] = p(1-p)^2\mathbb{E}[\delta^2] + (1-p)\mathbb{E}[(-p\delta+\xi)^2] = p(1-p)\mathbb{E}[\delta^2] + (1-p)\sigma^2. \tag{18}$$

Let M denote $\Phi^\top\Phi \in \mathbb{R}^{D\times D}$, where D is the number of parameters in the network, and Δ_t denote $W_t - W^{*'}$. Combining with Eqs. 11 and 13, we obtain:

$$W_{t+1} = (W^{*'} + \Delta_t) - \eta\frac{1}{N}\sum_N\{\Phi^\top[\Phi(\Delta_t + (W^* - W_0)) + f(u;W_0) - (f(u;W^{*'}) + \xi')]\}. \tag{19}$$

Note that:

$$f(u;W^{*'}) - f(u;W_0) \approx \Phi(W^{*'} - W_0). \tag{20}$$

We then have:

$$W_{t+1} = (W^{*'} + \Delta_t) - \eta\Phi^\top[\Phi\Delta_t + \xi'] = W^{*'} + [(I - \eta\Phi^\top\Phi)\Delta_t + \eta\Phi\xi']. \tag{21}$$

Thus,

$$\Delta_{t+1} = (I - \eta M)\Delta_t + \eta\Phi\xi'. \tag{22}$$

Starting from Δ_0, after T iterations, according to Eq. 22, we can obtain the explicit expression of Δ_T:

$$\Delta_T = \underbrace{(I-\eta M)^T\Delta_0}_{\text{Initial error decay}} + \underbrace{\eta\sum_{k=0}^{T-1}(I-\eta M)^k\Phi^\top\xi'}_{\text{Label noise accumulation}}. \tag{23}$$

Now, we are interested in $\mathbb{E}[\|\Delta_T\|_2^2]$, which contains quadratic terms of initial error decay and label noise accumulation as well as their cross term. Since label noise is independent of error caused by initializing model parameters and $\mathbb{E}[\xi'] = 0$, the cross term is 0. Thus,

$$\mathbb{E}[\|W_T - W^{*'}\|_2^2] = \mathbb{E}[\|(I-\eta M)^T\Delta_0\|_2^2] + \eta^2\mathbb{E}\left[\left\|\sum_{k=0}^{T-1}(I-\eta M)^k\Phi^\top\xi'\right\|_2^2\right]. \tag{24}$$

The first term is related to the initial error. For the noise term, let A_T be the cumulative sum of the updates:

$$A_T = \sum_{k=0}^{T-1}(I-\eta M)^k. \tag{25}$$

Then:

$$\sum_{k=0}^{T-1}\sum_{j=0}^{T-1}(I-\eta M)^k\Phi^\top\xi\xi'^\top\Phi(I-\eta M)^j = A_T\Phi^\top\Sigma_{\xi'}\Phi A_T^\top. \tag{26}$$

Since the noise covariance $\Sigma_{\xi'}$ is a scalar, using $M = \Phi^\top\Phi$, we get:

$$\mathbb{E}[\|A_T\Phi^\top\xi'\|_2^2] = \Sigma_{\xi'}\mathrm{Tr}\big(A_T M A_T^\top\big). \tag{27}$$

Using the matrix geometric series sum formula:

$$A_T = M^\dagger(I - (I-\eta M)^T), \tag{28}$$

where † denotes the Moore-Penrose pseudoinverse. Substituting,

$$A_T M A_T^\top = M^\dagger(I-(I-\eta M)^T)M(I-(I-\eta M)^T)M^\dagger. \tag{29}$$

Taking the trace:

$$\mathrm{Tr}\big(A_T M A_T^\top\big) = \mathrm{Tr}\big(M^\dagger(I-(I-\eta M)^T)^2\big). \tag{30}$$

Let M have eigenvalue decomposition:

$$M = V\Lambda V^\top, \quad \Lambda = \text{diag}(\lambda_1, ..., \lambda_D). \tag{31}$$

Note that for infinite-width network, $D \gg N$. Thus, $\lambda_i = 0$ for $N < i \leq D$. Then:

$$(I - \eta M)^{2T} = V(I - \eta\Lambda)^{2T} V^\top. \tag{32}$$

Therefore:

$$\text{Tr}\big[(I - \eta M)^{2T}\big] = \sum_{i=1}^{N} (1 - \eta\lambda_i)^{2T}. \tag{33}$$

Similarly,

$$\text{Tr}\big[M^\dagger(I - (I - \eta M)^T)^2\big] = \sum_{i=1}^{N} \frac{(1 - (1 - \eta\lambda_i)^T)^2}{\lambda_i}. \tag{34}$$

Thus, using Eq. 18, the result of $\mathbb{E}[\|\Delta_T\|_2^2]$ is:

$$\mathbb{E}[\|W_T - W^{*'}\|_2^2] = \mathbb{E}[\|(I - \eta M)^T \Delta_0\|_2^2] + \eta^2(p(1-p)\mathbb{E}[\delta^2] + (1-p)\sigma^2) \sum_{i=1}^{N} \frac{(1 - (1 - \eta\lambda_i)^T)^2}{\lambda_i}. \tag{35}$$

By triangle inequality, we have:

$$\begin{aligned}
\mathbb{E}[\|W_T - W^*\|_2^2] \leq & \mathbb{E}[\|W_T - W^{*'}\|_2^2] + \|W^{*'} - W^*\|_2^2 \\
= & \mathbb{E}[\|W_T - W^{*'}\|_2^2] + \|pW_{ref} + (1-p)W^* - W^*\|_2^2 \\
= & \mathbb{E}[\|(I - \eta M)^T \Delta_0\|_2^2] + \eta^2(p(1-p)\mathbb{E}[\delta^2] + (1-p)\sigma^2) \sum_{i=1}^{N} \frac{(1 - (1 - \eta\lambda_i)^T)^2}{\lambda_i} \\
& + p^2\|W_{ref} - W^*\|_2^2.
\end{aligned} \tag{36}$$

□

B. More Experimental Details

B.1. Prompts for AI Preference Study

To better compare the quality of generated images, we employ GPT-4o as the evaluator, assessing methods from three aspects: overall quality, visual aesthetics, and prompt alignment. The evaluation involved both pairwise comparisons and quantitative analysis. During the evaluation, for pairwise comparison, GPT-4o compares our method with the baseline methods, selecting the more preferred image. For quantitative analysis, GPT-4o assigns scores (0-100) to each image generated by each method. The prompts used in our testing are listed below, designed following the previous work PixArt-Sigma (Chen et al., 2024). For pairwise comparison, the designed prompt to evaluate the overall quality of images is as follows:

> As an AI visual assistant, you are an evaluator specialized in image quality analysis for high-resolution text-to-image generation models. Given a specific caption, please evaluate the overall quality of the image by considering both content alignment and technical excellence. For content alignment, assess the key information including object identities, properties, spatial relationships, object numbers and caption-specified style. For technical quality, evaluate the image's photorealism and aesthetics, focusing on clarity, richness of detail, artistic quality, and overall visual appeal. Please analyze how well the image performs in both aspects to determine its comprehensive quality, the prompt is *"your prompt"*. Please output [Image 1] if the first image is better, [Image 2] if the second image is better, and give me the reason.

The designed prompt to evaluate visual aesthetics of images is as follows:

> As an AI visual assistant, you are an evaluator specialized in image quality analysis for high-resolution text-to-image generation models. When presented with a specific caption, it is required to evaluate and determine which image exhibits greater photorealism and aesthetical, in terms of clarity, richness of detail, and overall quality. Please pay attention to the key factors, including image style, the artistic quality of the image, realism, etc., the prompt is *"your prompt"*. Please output [Image 1] if the first image is better, [Image 2] if the second image is better, and give me the reason.

The designed prompt to evaluate the prompt alignment of images is as follows:

> As an AI visual assistant, you are an evaluator specialized in image quality analysis for high-resolution text-to-image generation models. Given a specific caption, you need to judge which image aligns with the caption more closely. Please pay attention to the key information, including object identities, properties, spatial relationships, object numbers and image style, etc., the prompt is *"your prompt"*. Please output [Image 1] if the first image is better, [Image 2] if the second image is better, and give me the reason.

For quantitative analysis, the designed prompt to evaluate the overall quality is as follows:

> As an AI visual assistant, you specialize in evaluating image quality for high-resolution text-to-image generation models. Given a specific caption, please evaluate the overall quality of the image by considering both content alignment and technical excellence. For content alignment, assess the key information including object identities, properties, spatial relationships, object numbers and caption-specified style. For technical quality, evaluate the image's photorealism and aesthetics, focusing on clarity, richness of detail, artistic quality, and overall visual appeal. Please analyze how well the image performs in both aspects to determine its comprehensive quality. The prompt is: *"your prompt"*. Please output strictly the score from 0 to 100. Do not provide any explanation or additional text beyond this numeric score.

The designed prompt to evaluate the visual aesthetics of images is as follows:

> As an AI visual assistant, you specialize in evaluating image quality for high-resolution text-to-image generation models. When given a specific caption, you are required to assess the image and assign a 0-100 score, reflecting its photorealism, aesthetic appeal, clarity, richness of detail, and overall quality. Key factors to consider include image style, artistic quality, and realism. The prompt is: *"your prompt"*. Please output strictly the score from 0 to 100. Do not provide any explanation or additional text beyond this numeric score.

The designed prompt to evaluate the prompt alignment of images is as follows:

> As an AI visual assistant, you are an evaluator specialized in image quality analysis for high-resolution text-to-image generation models. Given a specific caption, you need to determine the score 0-100 that the image aligns with the caption. Please pay attention to the key information, including object identities, properties, spatial relationships, object numbers and image style, etc., the prompt is *"your prompt"*. Please output strictly the score from 0 to 100, reflecting how accurately the image aligns with the caption. Do not provide any explanation or additional text beyond this numeric score.

B.2. Prompts Used in User Study

1. *Craft an image depicting a surreal dreamscape with a majestic unicorn floating amidst tumultuous waves, viewed from an aerial perspective akin to observing through a bird's eyes soaring above the sea. This scene captures both serene beauty and chaotic turbulence in one fantastical landscape. Utilize vibrant, contrasting colors, featuring deep blues for the stormy sea and fiery oranges and purples for the swirling clouds overhead, creating an emotional gradient that evokes wonder, danger, and ethereal grace simultaneously.*
2. *A captivating Art Nouveau-inspired image showcases a celestial enchantress gracefully dancing amidst a swirling vortex*

of shimmering stardust, her ethereal gown intricately woven with delicate silver threads reminiscent of cosmic nebulae. Captured from an elevated perspective that accentuates the vastness and grandeur of the cosmos, this scene radiates a sense of wonder, enchantment, and serenity as it invites viewers to marvel at the luminous beauty of the muse against the backdrop of the infinite expanse.

3. *A surreal digital artwork depicting an enigmatic floating cityscape composed of inverted ziggurats suspended in midair. From an aerial perspective, the city appears to hover over an endless expanse of rippling water. As one draws closer to the water's surface, the reflection reveals a mirrored image of the city, but with its architecture twisted and distorted by the shifting tides. The vibrant colors and intricate details evoke a sense of wonder mixed with unease, inviting the viewer to contemplate the relationship between reality and illusion.*

4. *A surreal digital artwork depicting a bustling futuristic cityscape at night, with towering skyscrapers adorned in vibrant, abstract shapes. The scene transitions from sharp clarity to a soft, dreamlike atmosphere as it approaches the horizon, evoking both awe and uncertainty in the viewer. Neon lights pulse within the city, seemingly melting and warping in the air, creating mesmerizing patterns and reflections on the wet streets below. From a high-altitude perspective, this enigmatic metropolis is captured in an aerial view, inviting contemplation of the convergence of reality and dreams.*

5. *In the style of visionary art, depict a serene female figure draped in a luminous white gown with intricate mandala patterns in deep blue and vibrant teal hues. This ethereal portrait is viewed from an aerial perspective, showcasing the subject against a cosmic background that seamlessly blends into swirling galaxies and nebulae. The radiant colors and harmonious compositions evoke a profound sense of spiritual awakening and interconnectedness with the vast universe around us.*

6. *In the style of visionary art, depict a serene female figure draped in a luminous white gown with intricate mandala patterns in deep blue and vibrant teal hues. This ethereal portrait is viewed from an aerial perspective, showcasing the subject against a cosmic background that seamlessly blends into swirling galaxies and nebulae. The radiant colors and harmonious compositions evoke a profound sense of spiritual awakening and interconnectedness with the vast universe around us.*

7. *A dreamlike landscape emerges from a first-person viewpoint, immersing the observer in an alluring world where waterlilies of soft lavender and violet hues gracefully drift on the surface of an opalescent pond. Towering lotus blossoms stretch towards an indigo sky embellished with celestial bodies that gleam like stars, invoking both tranquility and awe. A regal swan presides over this fantastical garden, its iridescent feathers creating captivating ripples across the water that seem to distort time itself, crafting a harmonious melody of dreams and nature, encapsulating the spirit of beauty and whimsy in one stunning tableau.*

8. *Envision an otherworldly aquatic environment where a graceful mermaid adorned with iridescent scales reminiscent of deep-sea hues gracefully dances amidst vibrant coral formations. Her tranquil expression mirrors a state of contemplative introspection, as if she is enveloped in the enigmatic depths from an unconventional vantage point – that of a diminutive seashell. Delicate intricacies emerge, such as her cascading tresses mirroring tender seaweed and how sunlight weaves through the water to cast enchanting patterns upon her skin. This captivating scene instills a sense of awe and reflection while preserving an ethereal aura reminiscent of surrealistic art.*

9. *Craft an enchanting surrealist scene showcasing a 'bug's-eye view' perspective of a chess game occurring on a shifting landscape. In this scene, a majestic phoenix perches atop a black bishop, its vibrant wings casting intricate shadows across the checkerboard expanse below. The background alternates between lush tropical forests and barren deserts with each move made by the ethereal beings participating in this mysterious match, instilling feelings of intrigue and wonder as they traverse the unpredictable terrain under the eerie illumination of a full moon.*

10. *Imagine an enchanting digital artwork depicting a dragonfly's perspective above a tranquil pond, reminiscent of Monet's captivating water lilies. The scene showcases lush, vibrant vegetation surrounding the serene water surface, with an emotional gradient highlighting the beauty and enchantment as light gracefully dances across the composition. Merging elements of impressionism with high-resolution photorealistic textures, this piece evokes a sense of awe and wonder, creating an ethereal atmosphere that captures the dragonfly's mesmerizing flight amidst a blooming floral paradise.*

11. *Imagine a lively digital painting capturing the exhilarating spirit of an anime-inspired hoverboard race in a neon-drenched cyberpunk cityscape at twilight. The scene pulses with action as diverse characters navigate through a maze-like urban landscape of towering skyscrapers and radiant billboards, deftly maneuvering around airborne vehicles and zipping pedestrians while leaving trails of shimmering pixels behind them. As the sun dips below the horizon, its dramatic lighting casts elongated shadows that accentuate the futuristic architecture and high-tech trinkets sprinkled throughout, creating a rich tapestry of cool blues, purples, and pinks that intensify the emotional stakes and anticipation as the race nears its peak. This captivating image, with its dynamic composition and vibrant use of color and light, transports viewers into an enthralling world where technology and nature intertwine in a dazzling display of innovation and style.*

12. *An enigmatic digital artwork showcases an celestial ballerina gracefully spinning across the cosmos, her luminescent dress shimmering with iridescence against the inky black backdrop of space. The scene is observed from a unique 'worm's eye view', accentuating the grandeur and elegance of this otherworldly dancer as she whirls amidst nebulae and stars, creating an entrancing spectacle that captivates both the senses and the imagination while evoking a sense of wonder and awe for the hidden beauty within the universe.*

13. *A mesmerizing digital artwork portraying an enchanting celestial sorceress in shimmering robes adorned with starlight-inspired patterns. Seen from above, her captivating gaze reflects the light of distant galaxies as she skillfully shapes the cosmos with her mystical staff, creating swirling nebulae and brilliant constellations in a breathtaking display. The backdrop showcases a deep purple expanse of space filled with nebulous clouds, pulsating stars, and enigmatic planets, generating an emotional gradient that balances wonder and mystery within this celestial tableau.*

14. *Craft an image that embodies spiritual realism with celestial elements, depicting an astronaut in a reflective spacesuit adorned with intricate mandala patterns. This ethereal figure floats amidst the cosmic void, traversing through a luminous wormhole that connects our physical world to a higher plane of existence. The astronaut's journey is captured from an 'eye-of-the-needle' perspective, enveloped by vivid colors and profound symbolism that evoke wonder, transcendence, and spiritual awakening in the vast expanse of space.*

15. *A celestial ballet unfolds within an ethereal nebula as galaxies collide in a mesmerizing dance of cosmic forces. From an impossible vantage point, suspended above the event horizon, we gaze upon this breathtaking spectacle. The scene captures the chaotic beauty and sublime mystery of the universe, with swirling patterns of stars, nebulas, and cosmic dust illuminated by an otherworldly light source that casts long shadows across the celestial plane. This visually stunning composition evokes a sense of awe and wonder at the sheer scale and complexity of our cosmos, as well as the delicate balance between order and chaos in the grand design. Inspired by the visionary artistry of Escherian architecture and the cosmic explorations of space telescopes, this image invites viewers to ponder the infinite mysteries that lie beyond the stars.*

16. *Create a dreamlike still life scene blending impressionistic water lilies with surreal melting elements. Picture a tranquil garden pond adorned with softly ruffled blue, pink, and green flowers, as if captured in an ethereal dance of light and color. In the background, a tower melts into the distant horizon, its hands frozen in time, evoking a sense of timeless wonder. The reflection on the water's surface distorts the landscape into mesmerizing shapes and colors, blurring the line between reality and fantasy, enveloping the viewer in an enchanting atmosphere of surreal beauty and awe.*

17. *An awe-inspiring image reveals an enigmatic blend of swirling celestial patterns reminiscent of van Gogh's iconic brushstrokes and the surreal melting clocks synonymous with Dali's dreamscapes. Set within the opulent interior of a vast library, the scene captivates the viewer's gaze as they stand amidst towering shelves adorned with golden-hued spines. Above, an intricate celestial map unfolds across the ceiling, its constellations echoing van Gogh's dynamic style while clocks dissolve into nebulous forms that cast whimsical shadows over the polished marble floors. This surreal tableau masterfully combines wonder and introspection, inviting the observer to embark on a journey through time and space within its harmonious visual symphony.*

18. *A captivating digital artwork merges Victorian-era street market with an underwater world, creating a surreal fusion of reality and fantasy. The bustling activity of vendors selling exotic goods under a sky of swirling auroras transitions seamlessly into the vibrant depths of the ocean, inviting viewers to explore this mysterious realm. This unique blend of surface-level excitement and serene tranquility evokes curiosity and wonder, guiding the audience through a mesmerizing journey that defies traditional boundaries.*

19. *An enchanting digital artwork portrays an underwater haven where fantastical creatures interact harmoniously in a style reminiscent of James Gurney's Dinotopia. From an aerial perspective, this captivating scene reveals the gentle touch between a regal triceratops and a graceful mermaid, as they share a tender moment under the luminous sunlight filtering through the pristine waters. This mesmerizing encounter evokes wonder and tranquility, skillfully blending fantasy with prehistoric elements to create a visually stunning tableau that transcends time and imagination.*

20. *Craft an enchanting digital artwork that fuses surrealism with vibrant colors, depicting an underwater realm where fish gracefully perform ballet in harmony with the flowing currents. This unique perspective showcases a dreamlike dance between elegant sea creatures and floating bubbles, bathed in soft pastel hues evoking a sense of wonder and tranquility. Delicate brushstrokes and fantastical shapes intertwine to create a captivating visual symphony, inviting viewers into an otherworldly realm where reality and imagination seamlessly blend.*

C. More Experimental Results

C.1. More Results on AI Preference Study

We additionally use GPT-4o for a quantitative analysis, allowing for a more intuitive evaluation of performance of each method across different assessment dimensions. Our experimental setting and pairwise comparison remain consistent. We randomly select 300 prompts from the COCO30K dataset and generate images of 2048×2048. The results are shown in Table 4. The results indicate that our method achieves an exceptionally high level across all three dimensions and is comparable to SOTA model FLUX1.1 [Pro] Ultra.

Table 4: Results on AI preference study. Evaluation images are generated with COCO30K prompts with a resolution of 2048 × 2048.

Method/Setting	Overall Quality	Rank	Prompt Alignment	Rank	Visual Aesthetics	Rank
SDEdit	87.09	2	90.99	2	89.18	2
w/ URAE	88.23	1	92.49	1	90.09	1
I-Max	88.24	2	91.38	2	89.96	2
w/ URAE	89.12	1	92.58	1	90.86	1
FLUX1.1 [Pro] Ultra	90.42	1	93.53	2	90.42	2
PixArt-Sigma-XL	86.13	4	88.71	6	86.31	5
Sana-1.6B	86.46	3	90.25	3	87.80	3
FLUX.1-dev	84.23	6	89.05	5	84.88	6
FLUX.1-dev*	86.05	5	89.48	4	87.02	4
w/ URAE	89.71	2	93.64	1	91.47	1

C.2. More Results on FID and LPIPS against Real Images

In Table 1, we report the FID and LPIPS evaluation results of baseline methods against reference images generated by FLUX1.1 [Pro] Ultra model. In this section, we also provide FID and LPIPS results of baseline methods against real images. The results are shown in Table 5.

Table 5: Results on FID and LPIPS evaluated against real images. Evaluation images are generated with 2, 000 and 1, 000 prompts randomly selected from COCO2014val with a resolution of 2048 × 2048 and 4096 × 4096.

2048 × 2048	FID	LPIPS	4096 × 4096	FID	LPIPS
FLUX1.1 [Pro] Ultra	47.12	0.4518	FLUX1.1 [Pro] Ultra	-	-
PixArt-Sigma-XL	57.02	0.5075	PixArt-Sigma-XL	75.81	0.5066
Sana-1.6B	54.57	0.5122	Sana-1.6B	73.46	0.5108
Ours	**52.95**	**0.4669**	**Ours**	**70.44**	**0.4647**

C.3. More Results on Few-Step Diffusion Model

We also conduct experiments on FLUX.1-Schnell to demonstrate the generalization and the strong adaptation capabilities of our proposed method. Without any additional training, a trained adapter on FLUX.1-dev can be migrated onto FLUX.1-schnell, which can generate high-quality results with only 4 denoising steps and achieves 6× acceleration compared with FLUX.1-dev (25.8 v.s. 36.5 sec./image). The results are shown in Table 6.

Table 6: The quantitative results on FLUX.1-Schnell. All images are generated with the $2,000$ prompts randomly selected from COCO2014val, and are of 2048 × 2048 in size. Here, FID is evaluated against real images. Here, FLUX.1-Schnell* is FLUX.1-Schnell with scaled RoPE, proportional attention, and removing dynamic shifting strategies.

	FID (↓)	**HPSv2.1 (↑)**	**PickScore (↑)**
FLUX.1-schnell	42.42	27.97	22.07
FLUX.1-schnell*	42.20	28.17	22.38
w/ URAE	**38.66**	**29.63**	**22.74**

C.4. More Results on Ablation Studies

C.4.1. Choice of Singular Component Rank

We also conduct the ablation studies to analyze the sensitivity to the singular component rank r. The evaluation results are shown in Table 7. From the results, the performance remains stable when r is around 16.

Table 7: The quantitative results on the choice of singular component rank (r). All images are generated with the $2,000$ prompts randomly selected from COCO2014val, and are of 2048 × 2048 in size. Here, FID is evaluated with inference images generated by FLUX1.1 [Pro] Ultra.

	FID (↓)	**HPSv2.1 (↑)**	**PickScore (↑)**
$r = 1$	42.34	30.21	23.01
$r = 4$	38.08	30.99	23.04
$r = 16$	38.85	31.50	23.21
$r = 64$	38.97	30.14	22.91
$r = 256$	38.77	29.89	22.82

C.4.2. Training-Time GPU Memory Requirements for Different Ranks

For parameter efficiency, we conduct experiments on training-time GPU memory requirement (MB) with respect to various ranks of the adapters. The results are shown in Table 8. We observe that comparing with full-rank adaptation, the low-rank adapters save GPU memory over 50%.

Table 8: Training-time GPU memory requirements for different ranks for 2048 × 2048 and 4096 × 4096 image generation tasks. Here, we adopt $r = 16$ as the default setting in the paper.

	$r = 1$	$r = 4$	$r = 16$ (Default)	$r = 64$	$r = 256$	$r = 3072$ (Full)
2048 × 2048	35916	35958	36124	36816	39884	77880
4096 × 4096	62806	62850	63010	63704	66114	OOM