Diffusion-4K: Ultra-High-Resolution Image Synthesis with Latent Diffusion Models

Jinjin Zhang[1,2] Qiuyu Huang[3] Junjie Liu[3] Xiefan Guo[1,2] Di Huang[1,2*]
[1]State Key Laboratory of Complex and Critical Software Environment,
Beihang University, Beijing 100191, China
[2]School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[3]Meituan
{jinjin.zhang, xfguo, dhuang}@buaa.edu.cn {huangqiuyu, liujunjie10}@meituan.com

Figure 1. Example results synthesized by our **Diffusion-4K**, emphasizing exceptional fine details in generated 4K images.

Abstract

In this paper, we present Diffusion-4K, a novel framework for direct ultra-high-resolution image synthesis using text-to-image diffusion models. The core advancements include: (1) Aesthetic-4K Benchmark: addressing the absence of a publicly available 4K image synthesis dataset, we construct Aesthetic-4K, a comprehensive benchmark for ultra-high-resolution image generation. We curated a high-quality 4K dataset with carefully selected images and captions generated by GPT-4o. Additionally, we introduce GLCM Score and Compression Ratio metrics to evaluate fine details, combined with holistic measures such as FID, Aesthetics and CLIPScore for a comprehensive assessment of ultra-high-resolution images. (2) Wavelet-based Fine-tuning: we propose a wavelet-based fine-tuning approach for direct training with photorealistic 4K images, applicable to various latent diffusion models, demonstrating its effectiveness in synthesizing highly detailed 4K images. Consequently, Diffusion-4K achieves impressive performance in high-quality image synthesis and text prompt adherence, especially when powered by modern large-scale diffusion models (e.g., SD3-2B and Flux-12B). Extensive experimental results from our benchmark demonstrate the superiority of Diffusion-4K in ultra-high-resolution image synthesis. Code is available at

*Corresponding author.

https://github.com/zhang0jhon/diffusion-4k.

1. Introduction

Diffusion models have proven to be highly effective for modeling high-dimensional, perceptual data such as images [18, 21, 29, 31, 46]. In recent years, latent diffusion models have made significant advancements in text-to-image synthesis, demonstrating impressive generalization capabilities at high resolutions [9, 12, 23, 25, 34, 41]. Replacing convolutional U-Net with a transformer architecture shows promising improvements as model scalability increases, *e.g.*, Stable Diffusion 3 (SD3) at 8B [12], Flux at 12B [5], and Playground v3 at 24B [25]. Alternatively, flow-based formulations, using data or velocity prediction, have emerged as an viable choice due to their faster convergence and enhanced performance [1, 21, 24, 26, 28].

Despite substantial progress, most latent diffusion models focus on training and generating images at 1024×1024, leaving direct ultra-high-resolution image synthesis largely unexplored. Direct training and generating 4K images is valuable in realistic applications, but require considerable computational resources, especially as model parameters increase. Concurrent approaches such as PixArt-Σ [9] and Sana [49] address direct ultra-high-resolution image synthesis at 4K, showcasing the potentials of scalable latent diffusion transformer architectures. Both PixArt-Σ at 0.6B and Sana at 0.6/1.6B primarily focus on the efficiency of ultra-high-resolution image generation, however, the intrinsic advantages of 4K images, such as high-frequency details and rich textures, are neglected. Furthermore, due to the scarcity of photorealistic 4K images, there is currently no publicly available benchmark for 4K image synthesis, hindering further research on this valuable topic.

In this paper, we propose Diffusion-4K, a novel framework designed for direct ultra-high-resolution image synthesis using latent diffusion models. Specifically, we introduce Aesthetic-4K, including a high-quality dataset of curated ultra-high-resolution images, with corresponding captions generated by GPT-4o [19]. Additionally, most automated evaluation metrics, such as Fréchet Inception Distance (FID) [17], Aesthetics [45] and CLIPScore [16], offer only holistic measures at low resolutions, which are insufficient for the comprehensive benchmarking in high-resolution image synthesis, particularly for 4K images. To address the limitations, we introduce Gray Level Co-occurrence Matrix (GLCM) Score and Compression Ratio metrics, focusing on assessment of fine details in ultra-high-resolution images which has not yet been explored, with the goal of establishing a comprehensive benchmark for 4K image synthesis. Additionally, we propose a wavelet-based fine-tuning method that emphasizes high-frequency components while preserving low-frequent approximation in ultra-high-resolution image synthesis. Notably, our method is compatible with various latent diffusion models. We conduct experiments with open-sourced latent diffusion models including SD3-2B [12] and Flux-12B [5], capable of training and generating photorealistic images at 4096×4096 resolution. Consequently, our method demonstrates superior performance in 4K image synthesis, underscoring the effectiveness of the Diffusion-4K framework.

The main contributions are summarized as follows:

- We establish Aesthetic-4k, a comprehensive benchmark for 4K image synthesis, including a high-quality dataset of ultra-high-resolution images with corresponding precise captions, and elaborated evaluation metrics for ultra-high-resolution images generation.
- We propose a wavelet-based fine-tuning approach for latent diffusion model, focusing on generating ultra-high-resolution images with fine details.
- Extensive experiment results demonstrate the effectiveness and generalization of our proposed method in 4K image synthesis, particularly when powered by large-scale diffusion transformers, *e.g.* SD3 and Flux.

2. Related work

Latent Diffusion Models. Stable Diffusion (SD) [41] proposes latent diffusion models to operate diffusion process in compressed latent space using Variational Auto-Encoder (VAE). Widely used VAEs [7, 12, 31] employ a down-sampling factor of $F = 8$, compressing pixel space $\mathbb{R}^{H \times W \times 3}$ into latent space $\mathbb{R}^{\frac{H}{F} \times \frac{W}{F} \times C}$, where H and W represent height and width respectively, and C denotes the channel of latent space. In the field of latent diffusion models, Diffusion Transformer (DiT) [31] has made significant progress in the past year. Typically, the patch size of DiT is set to $P = 2$, resulting in $\frac{H}{FP} \times \frac{H}{FP}$ tokens. The transformer architecture demonstrates excellent scalability in latent diffusion models, as evidenced by PixArt [7, 9], SD3 [12], Flux [5], Playground [23, 25], *etc*.

In text-to-image synthesis, the text encoder plays a crucial role in prompt coherence. SD employs CLIP [35] as its text encoder, while subsequent diffusion models, such as Imagen [42] and PixArt [7, 9], utilize T5-XXL [36] for text feature extraction. Recent approaches, such as SD3 [12] and Flux [5], combine both CLIP and T5-XXL for enhanced text understanding. DALL-E 3 [4] demonstrates that training with descriptive image captions can significantly enhance the prompt coherence of text-to-image models. Sana [49] utilizes the latest decoder-only Large Language Model (LLM), Gemma 2 [48], as its text encoder to improve understanding and reasoning capabilities related to text prompts.

High-Resolution Image Synthesis. High-resolution image generation is valuable in various practical applications such as industry and entertainment [6, 11]. Currently, advanced

latent diffusion models typically train and synthesize images at 1024×1024 for high-resolution image generation due to computational complexity [34, 38, 39, 41, 43, 44]. However, increasing image resolution introduces quadratic computational costs, posing challenges particularly for 4K image synthesis. Several training-free fusion approaches for 4K image generation have been developed based on existing latent diffusion models [2, 10]. Additionally, Stable Cascade [32] employs multiple diffusion networks to increase resolution progressively. However, such ensemble methods can introduce cumulative errors.

PixArt-Σ [9] pioneers to direct image generation close to 4K (3840×2160) with efficient token compression for DiT, significantly improving efficiency and enabling ultra-high-resolution image generation. Sana [49], a pipeline designed to efficiently and cost-effectively train and synthesize 4K images, is capable of generating images at resolutions ranging from 1024×1024 to 4096×4096. Sana introduces a deep compression VAE [8] that compresses images with an aggressive down-sampling factor of $F = 32$, enabling content creation at low cost. Despite significant improvements in resolution, both PixArt-Σ-0.6B and Sana-0.6B/1.6B ignore the intrinsic high-frequency details and rich textures in 4K image synthesis. Furthermore, the potential for scalable DiT in 4K image generation remains unexplored.

3. Methods

In this section, we elaborate on the details of Diffusion-4K, illustrating how it facilitates photorealistic 4K image synthesis. In Sec. 3.1, we introduce Aesthetic-4K, outlining the design principles and the process of constructing our human-centric benchmark with ultra-high-resolution images. Subsequently, we present a wavelet-based fine-tuning approach for direct training with photorealistic 4K images, applicable to various latent diffusion models, in Sec. 3.2.

3.1. Aesthetic-4K Benchmark

Existing benchmarks for advanced latent diffusion models generally conduct experiments with images at 1024×1024 resolution [7, 12, 23, 34], however, direct training and assessing with 4K images has not been thoroughly addressed yet. In this section, we construct Aesthetic-4K, a comprehensive benchmark for ultra-high-resolution image synthesis. The details are outlined as follows.

Design Principles. *Emphasis on human-centric perceptual cognition:* To design a benchmark centered on human-level perceptual cognition, our primary goal is to explore the key factors of human perception on 4K images and formulate quantifiable indicators, while evaluating the general abilities of the generative model in these key factors. However, most automated evaluation metrics, such as FID [17], Aesthetics [45], and CLIPScore [16], provide only holistic measures at lower resolutions, which are insufficient for comprehensive assessment, particularly lacking the ability of evaluating fine details in 4K images.

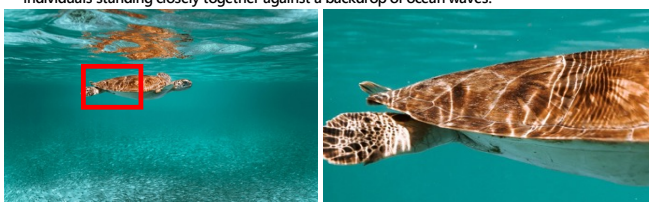
Figure 2. Analysis of GLCM Score↑ / Compression Ratio↓. Our indicators demonstrate strong alignment with human-centric perceptual cognition at the level of local patches.

Metric	GLCM Score	Compression Ratio	MUSIQ	MANIQA
SRCC	0.75	0.53	0.36	0.20
PLCC	0.77	0.56	0.41	0.26

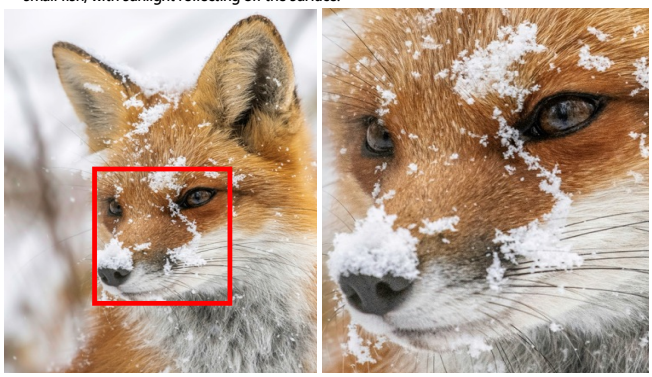
Table 1. Correlation with human evaluation.

Inspired by relevant literature in perceptual psychology, we found that highly structured textures play a decisive role in human perceptual cognition [3, 20, 47]. Therefore, we proposed additional indicators that accurately measure the richness of highly structured textures, including the commonly used quantitative indicators for texture analysis GLCM and the image compression ratio with discrete cosine transform (DCT). Considering human visual sensitivities, the GLCM, which captures the varying textural patterns, optical flow, and distortion through comprehensive spatial interactions of neighboring pixels, enables the creation of a truly representative characterization of human visual perception [13]. Simultaneously, the image compression ratio with DCT serves as an important reference for assessing the preservation of fine details in 4K images. To demonstrate alignment with human preferences, we conduct quantitative analysis of our indicators with adequate images, and present examples in Fig. 2. Moreover, as depicted in Tab. 1, we compute the Spearman Rank-order Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) with human evaluations using various patches sampled from generated images, demonstrating the superior alignment with human ratings compared to no-reference image quality assessment metrics such as MUSIQ [22] and MANIQA [50]. Our benchmark design ensures that the model performance assessment is directly relevant to human decision-making and cognitive abilities.

Prompt: A vast colony of king penguins densely populates a rocky shore, with numerous individuals standing closely together against a backdrop of ocean waves.



Prompt: A sea turtle glides gracefully through crystal-clear turquoise water above a school of small fish, with sunlight reflecting off the surface.

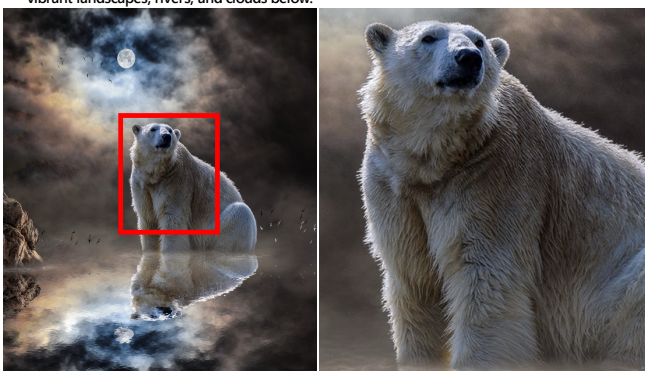


Prompt: A close-up of a fox's face, partially covered in snow, with sharp ears and bright, alert eyes. The snowy backdrop adds a serene atmosphere to the scene.

(a) Image-text samples in training set.

Prompt: A picturesque riverside scene featuring a medieval castle atop a hill, surrounded by vibrant autumn foliage, with colorful homes lining the waterfront and several boats docked along the river.

Prompt: A lone astronaut floats in space, gazing at a swirling black hole surrounded by vibrant landscapes, rivers, and clouds below.



Prompt: A majestic polar bear sits near a rocky shore, reflecting in calm waters under a full moon amid a dramatic, cloudy sky filled with flying birds. A small figure stands in the distance, observing the scene.

(b) Image-text samples in evaluation set.

Figure 3. Illustration of image-text samples in the Aesthetic-4K dataset, which includes high-quality images and precise text prompts generated by GPT-4o, distinguished by high aesthetics and fine details.

Aesthetic-4K Dataset. Addressing the absence of a publicly available 4K dataset, we propose the Aesthetic-4K dataset that includes a well-curated training and evaluation set for in-depth study. For the Aesthetic-4K training set, we collect high-quality images from the Internet, distinguished by exceptional aesthetics and fine details. Notably, the Aesthetic-4K training set consists of 12,015 high-quality images with a median height of 4128 and median width of 4640, representing a significant improvement in training data for ultra-high-resolution image synthesis. Additionally, we provide precise and detailed image captions generated by powerful GPT-4o [19], demonstrating strong alignment between vision and language. These curated images and corresponding captions comprise Aesthetic-Train, the training set of Aesthetic-4K.

For the evaluation set, Aesthetic-Eval, we select image-text pairs from the LAION-Aesthetics V2 6.5+ dataset, following the criterion of a short image side greater than 2048. The LAION-Aesthetics dataset includes 625,000 image-text pairs with predicted aesthetic scores of 6.5 or higher in LAION-5B [45]. Notably, the Aesthetic-4K evaluation set consists of 2,781 high-quality images with a median height of 2983 and median width of 3613. We do not sample the evaluation set from the collected images sourced from the Internet, to mitigate the risk of overfitting in comprehensive assessments. Within the evaluation set, 195 images have a short side greater than 4096, which we refer to as Aesthetic-Eval@4096. Instead of evaluating with images at 1024×1024 resolution [49], the proposed Aesthetic-4K evaluation set provides a suitable benchmark for ultra-high-resolution image synthesis.

In summary, the proposed dataset encompasses common categories in realistic scenarios, including nature, travel, fashion, animals, film, art, food, sports, street photography, *etc*. As illustrated in Fig. 3, we present several image-text pairs from both the training and evaluation sets of Aesthetic-4K, clearly demonstrating their exceptional quality. For more details about the Aesthetic-4K dataset, please refer to the supplementary materials.

Comprehensive Evaluation Metrics. To construct a comprehensive human-centric benchmark for 4K image synthesis, we provide conventional evaluation metrics in genera-

tive models, such as FID [17], Aesthetics [45], and CLIP-Score [16], for an intuitive understanding of image synthesis from a holistic perspective.

In addition, we report quantitative results for the proposed metrics assessing fine details in 4K images, including the GLCM Score and Compression Ratio, as significant complements to human-centric perceptual cognition in 4K images, which have not been addressed previously. The GLCM Score is formulated as follows: $s = -\frac{1}{P}\sum_{p=1}^{P} H(g_p)$, where H represents entropy, and g_p is the GLCM [15] derived from local patch p in the original image with 64 gray levels, defined by radius $\delta = [1, 2, 3, 4]$ and orientation $\theta = [0°, 45°, 90°, 135°]$. In practice, we divide the gray image into local patches of size 64, and compute the average GLCM Score based on the partitioned local patches. Regrading the Compression Ratio, it is calculated as the ratio of the original image size in memory to the compressed image size using the JPEG algorithm at a quality of 95. Qualitative and quantitative assessments constitute the comprehensive benchmark, Aesthetic-4K.

3.2. Wavelet-based Fine-tuning

In this section, we propose a **W**ave**L**et-based **F**ine-tuning (WLF) method suitable for various latent diffusion models, enabling direct training with photorealistic images at 4096×4096 resolution. The core improvements consist of two components:

Partitioned VAE. The most commonly used VAEs [5, 12] with $F = 8$ encounter out-of-memory (OOM) issues during direct training and inference at 4096×4096. To address this issue, we propose an efficient partitioned VAE, a simple yet effective method that compresses images with $F = 16$, significantly reducing memory consumption. Specifically, we use a dilation rate of 2 in the first convolutional layer of the VAE's encoder. In the last convolutional layer of the VAE's decoder, we partition the input feature map, up-sample by a factor of 2, apply the same convolution operator to each partitioned feature map, and then reorganize the final output. Notably, our method maintains consistency in the latent space of pre-trained latent diffusion models, eliminating the need for retraining or fine-tuning VAEs, which prevents distribution shifts in the latent space and ensures compatibility with various diffusion models.

Wavelet-based Latent Enhancement. Wavelet transform has achieved significant success in image processing and is widely used to decompose low-frequency approximations and high-frequency details from images or features [14, 33]. In this section, we propose wavelet-based latent enhancement, which emphasizes high-frequency components while preserving low-frequency information, thereby significantly enhancing fine details and rich textures in 4K image generation. Consider the diffusion process, *i.e.*

$$\boldsymbol{z}_t = \alpha_t \cdot \boldsymbol{x}_0 + \sigma_t \cdot \epsilon, \tag{1}$$

where $\boldsymbol{x}_0$ and ϵ represent data distribution and standard normal distribution, respectively, and α_t and σ_t are hyper-parameters in the diffusion formulation. The original training objective in latent diffusion models is noise prediction [18, 41], defined as:

$$\epsilon_\Theta(\boldsymbol{z}_t, t) = \epsilon_t, \tag{2}$$

where Θ denotes the neural network, *e.g.*, convolutional UNet or DiT. Recent approaches, such as SD3 [12] and Flux [5], adopt rectified flows to predict velocity $\boldsymbol{v}$ parameterized by Θ, with the objective as follows:

$$\boldsymbol{v}_\Theta(\boldsymbol{z}_t, t) = \epsilon - \boldsymbol{x}_0. \tag{3}$$

Our method decomposes the low-frequency approximation and high-frequency details of latent features using wavelet transform, defining the training objective as follows:

$$\mathcal{L}_{WLF} = \mathbb{E}\left[w_t \| f(\boldsymbol{v}_\Theta(\boldsymbol{z}_t, t) - \epsilon) - f(\boldsymbol{x}_0)\|^2\right], \tag{4}$$

where w_t is the loss weight, and $f(\cdot)$ denotes discrete wavelet transform (DWT). Notably, we use the Haar wavelet, widely adopted in real-world applications due to its efficiency. Specifically, $L = \frac{1}{\sqrt{2}}[1, 1]$ and $H = \frac{1}{\sqrt{2}}[-1, 1]$ denote the low-pass and high-pass filters, used to construct four kernels in DWT with stride 2, namely LL^T, LH^T, HL^T, HH^T. The DWT kernels are then used to decompose the input features into four sub-bands, the low-frequency approximation $\boldsymbol{x}_{ll}$ and high-frequency components $\boldsymbol{x}_{lh}, \boldsymbol{x}_{hl}, \boldsymbol{x}_{hh}$. As a result, both low-frequency information and high-frequency details are incorporated in Eq. (4), contributing to a comprehensive optimization in 4K image synthesis. In addition, our method supports various diffusion models by simply substituting the reconstruction objective, enabling seamless integration with conventional noise prediction approaches.

4. Experiments

To demonstrate the effectiveness of our method, we conduct experiments with the state-of-the-art latent diffusion models at various scales, including open-source SD3-2B [12] and Flux-12B [5]. Specifically, mainstream evaluation metrics, such as FID [17], Aesthetics [45] and CLIPScore [16], as well as the proposed GLCM Score and Compression Ratio metrics, are reported for comprehensive assessments. Additionally, we present quantitative and qualitative results that showcase the high-resolution image reconstruction and generation capabilities of partitioned VAE and WLF, respectively.

4.1. Experimental Settings

Implementation Details. Regrading the VAEs in both SD3 [12] and Flux [5], we encapsulate them using partitioned VAE with $F = 16$, ensuring the direct training at

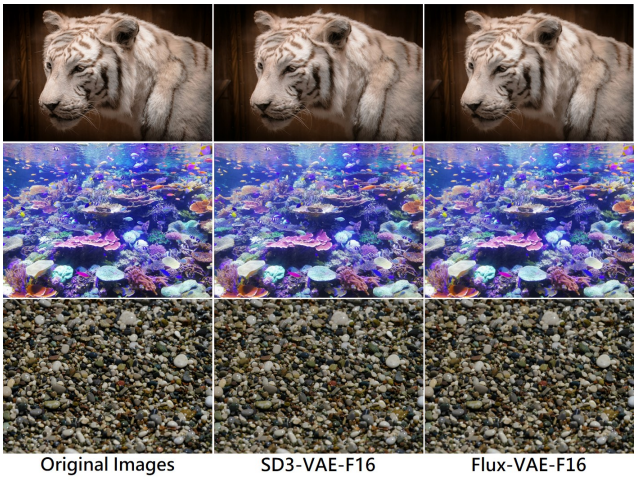


Figure 4. Reconstruction results of 4K images with partitioned VAEs of $F = 16$.

4096×4096 resolution without OOM issues. Notably, WLF is used for fine-tuning diffusion models, while the partitioned VAE and text encoder in pre-trained models are kept frozen, significantly improving the training efficiency. In practice, we use the AdamW [27] optimizer with a constant learning rate of 1e-6 and weight decay of 1e-4. We employ mixed-precision training with a batch size of 32 and use ZeRO Stage 2 with CPU offload techniques [37, 40]. The fine-tuning of SD3-2B and Flux-12B is conducted on 2 A800-80G GPUs and 8 A100-80G GPUs, respectively.

4.2. Experiment Results

Analysis of partitioned VAEs. In Tab. 2, we report the evaluation results, including rFID, Normalized Mean Square Error (NMSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and LPIPS [51], to assess the reconstruction ability of partitioned VAEs on Aesthetic-4K at 4096×4096 resolution. We provide detailed results for VAEs in SD3 and Flux, enhanced by our partition techniques with $F = 16$. Notably, our partitioned VAE resolves the OOM issue encountered by the original VAE without requiring retraining or fine-tuning, thereby avoiding potential distribution shifts in the latent space. In addition, we present visualizations of original and reconstruction images in Fig. 4. The qualitative results demonstrate the effectiveness of the partitioned VAEs in 4K image reconstruction, ensuring consistency in the latent space for subsequent fine-tuning with 4K images.

Image Quality Assessment. Regarding image quality assessment, we conduct comprehensive comparisons using mainstream evaluation metrics, such as FID [17], Aesthetics [45] and CLIPScore [16], to provide an intuitive understanding of image quality and text prompt adherence. In Tab. 3, we report experimental results with SD3-2B and Flux-12B on Aesthetic-Eval@2048 in Tab. 3, demonstrating the effectiveness of our method.

Model	rFID	NMSE	PSNR	SSIM	LPIPS
SD3-VAE-F16	1.40	0.09	28.82	0.76	0.15
Flux-VAE-F16	1.69	0.08	29.22	0.79	0.16

Table 2. Quantitative reconstruction results of VAEs with downsampling factor of $F = 16$ on our Aesthetic-4K training set at 4096×4096.

Model	FID ↓	CLIPScore ↑	Aesthetics ↑
SD3-F16@2048	43.82	31.50	5.91
SD3-F16-WLF@2048	**40.18**	**34.04**	**5.96**
Flux-F16@2048	50.57	30.41	6.36
Flux-F16-WLF@2048	**39.49**	**34.41**	**6.37**

Table 3. Quantitative benchmarks of latent diffusion models on Aesthetic-Eval@2048 at 2048×2048 resolution.

Model	GLCM Score ↑	Compression Ratio ↓
SD3-F16@2048	0.75	11.23
SD3-F16-WLF@2048	**0.79**	**10.51**
Flux-F16@2048	0.58	14.80
Flux-F16-WLF@2048	**0.61**	**13.60**

Table 4. GLCM Score and Compression Ratio of latent diffusion models on Aesthetic-Eval@2048 at 2048×2048 resolution.

As aforementioned, these conventional evaluation metrics are insufficient for comprehensive assessment in ultra-high-resolution image synthesis, particularly in evaluating the fine details of 4K images. In addition to commonly used benchmarks, we also compare the GLCM Score, which aims to assess the texture richness in ultra-high-resolution images. Simultaneously, we also report the Compression Ratio using the JPEG algorithm at a quality setting of 95, which can serve as an important reference for assessing image quality with fine details. As depicted in Tab. 4, experimental results highlight the advantages of WLF in these metrics, distinctly demonstrating our method's ability to generate 4K images with fine details.

Qualitative 4K Image Synthesis. As illustrated in Fig. 5, we present qualitative high-resolution images synthesized with Diffusion-4K, powered by the state-of-the-art latent diffusion model, Flux-12B. Although WLF fine-tunes at 4096×4096 resolution, our method can synthesize ultra-high-resolution images at various aspect ratios, and even supports direct photorealistic image generation at higher resolutions. In addition, we report the inference details in Tab. 5, highlighting the time and memory consumption of our method for directly generating 4K images. For sampling, images are generated by discretizing the ordinary

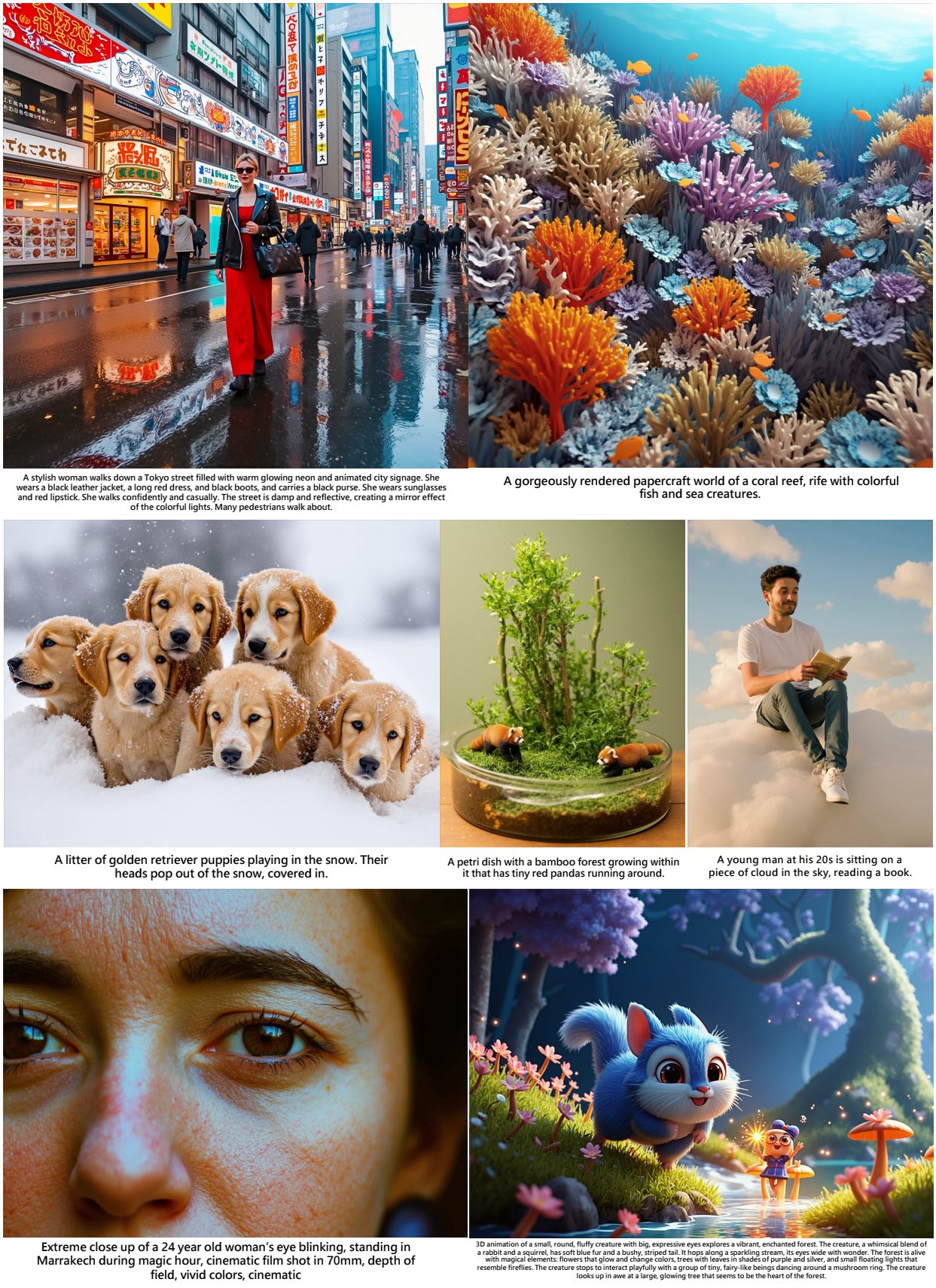


Figure 5. Qualitative 4K images synthesis of Diffusion-4K. Prompts are from Sora [30].

Model	Memory	Time (s/step)
SD3-F8@4096	OOM	-
SD3-F16-WLF@4096	31.3GB	1.16
SD3-F16-WLF@4096 (CPU offload)	16.1GB	1.22
Flux-F8@4096	OOM	-
Flux-F16-WLF@4096	50.4 GB	2.42
Flux-F16-WLF@4096 (CPU offload)	26.9 GB	3.16

Table 5. Memory consumption and inference speed of direct image synthesis at 4096×4096. The result is tested on one A100 GPU with BF16 Precision.

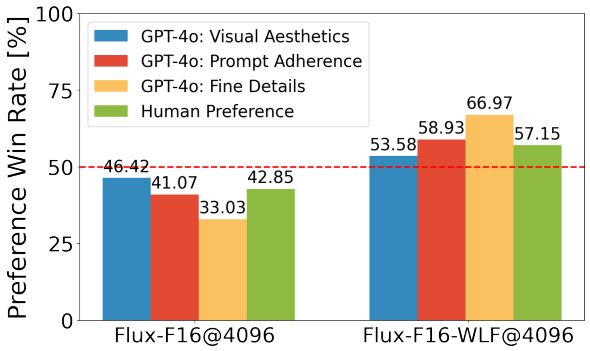


Figure 6. Human and GPT-4o preference evaluation.

differential equation (ODE) process using an Euler solver. Please refer to the supplementary materials for more comparisons and details.

Preference Study. Human preference is assessed by rating the pairwise outputs from two models, *i.e*. Flux with and without WLF. Additionally, we employ the advanced multi-modal model, GPT-4o [19], as the evaluator for the preference study. The following questions are asked:

Visual Aesthetics: *Given the prompt, which image is of higher-quality and aesthetically more pleasing?*

Prompt Adherence: *Which image looks more representative to the text shown above and faithfully follows it?*

Fine Details: *Which image more accurately represents the fine visual details? Focus on clarity, sharpness, and texture. Assess the fidelity of fine elements such as edges, patterns, and nuances in color. The more precise representation of these details is preferred! Ignore other aspects.*

As shown in Fig. 6, our method demonstrates a superior win rate in both human and AI preference, highlighting the effectiveness of our WLF method in various aspects, including visual aesthetics, prompt adherence and fine details performance in 4K images.

4.3. Ablation Studies

Ablation on Quality of Image Captions. We compare 4K image synthesis using original captions from LAION-5B [45] and captions generated by GPT-4o. As shown in Tab. 6, both SD3 and Flux exhibit enhanced results in Aesthetic-Eval@4096 with captions from GPT-4o. Quantitative results indicate that prompts generated by GPT-4o improve image synthesis and prompt coherence, demonstrating the significance of high-quality prompts in 4K image generation.

Captions	Model	CLIPScore ↑	Aesthetics ↑
LAION-5B	SD3-F16@4096	29.37	5.90
GPT-4o	SD3-F16@4096	**33.12**	**5.97**
LAION-5B	Flux-F16@4096	29.12	6.02
GPT-4o	Flux-F16@4096	**33.67**	**6.11**

Table 6. Ablation study on quality of image captions on Aesthetic-Eval@4096 at 4096×4096.

Model	CLIPScore ↑	Aesthetics ↑
SD3-F16@4096	33.12	5.97
SD3-F16-finetune@4096	34.14	5.99
SD3-F16-WLF@4096	**34.40**	**6.07**

Table 7. Ablation on CLIPScore and Aesthetics of SD3 on Aesthetic-Eval@4096 at 4096×4096. SD3-F16-finetune@4096 represents fine-tuning without WLF.

Model	GLCM Score ↑	Compression Ratio ↓
SD3-F16@4096	0.73	11.97
SD3-F16-finetune@4096	0.74	11.41
SD3-F16-WLF@4096	**0.77**	**10.50**

Table 8. Ablation on GLCM Score and Compression Ratio of SD3 on Aesthetic-Eval@4096 at 4096×4096. SD3-F16-finetune@4096 denotes fine-tuning without WLF.

Ablation on WLF. To demonstrate the effectiveness of WLF, we conduct ablation studies with SD3, comparing fine-tuning diffusion models with and without WLF, and report the experiment results of Aesthetic-Eval@4096 in Tab. 7 and Tab. 8. Compared to fine-tuning without WLF, our WLF method demonstrates superior performance in CLIPScore, Aesthetics, GLCM Score and Compression Ratio, significantly showcasing its effectiveness in visual aesthetics, prompt adherence, and high-frequency details.

5. Conclusion

In this paper, we present Diffusion-4K, a novel framework for ultra-high-resolution text-to-image generation. We introduce Aesthetic-4K benchmark and wavelet-based fine-tuning, capable of training with the state-of-the-art latent diffusion models, such as SD3 and Flux. Qualitative and quantitative results demonstrate the effectiveness of our approach in training and generating photorealistic 4K images.

6. Acknowledgement

This work is partly supported by the National Natural Science Foundation of China (82441024), the Research Program of State Key Laboratory of Critical Software Environment, and the Fundamental Research Funds for the Central Universities.

References

[1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 2

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 3

[3] James R Bergen and Edward H Adelson. Early vision and texture perception. *Nature*, 333(6171):363–364, 1988. 3

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2

[5] black-forest labs. Flux, 2024. https://github.com/black-forest-labs/flux. 2, 5

[6] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *European conference on computer vision*, pages 170–188. Springer, 2022. 2

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2, 3

[8] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024. 3

[9] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 2, 3

[10] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6159–6168, 2024. 3

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[12] 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. 2, 3, 5

[13] Dhanashree Gadkari. Image quality analysis using glcm. 2004. 3

[14] Florentin Guth, Simon Coste, Valentin De Bortoli, and Stephane Mallat. Wavelet score-based generative modeling. *Advances in neural information processing systems*, 35:478–491, 2022. 5

[15] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973. 5

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2, 3, 5, 6

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2, 3, 5, 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 5

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 4, 8

[20] Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. 3

[21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 2

[22] 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*, pages 5148–5157, 2021. 3

[23] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 2, 3

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[25] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024. 2

[26] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2

[27] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[28] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024. 2

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[30] OpenAI. Sora, 2024. https://openai.com/index/video-generation-models-as-world-simulators. 7

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[32] Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023. 3

[33] Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10199–10208, 2023. 5

[34] 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. 2, 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2

[37] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 6

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 3

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[40] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. {Zero-offload}: Democratizing {billion-scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 551–564, 2021. 6

[41] 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*, pages 10684–10695, 2022. 2, 3, 5

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[43] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 3

[44] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024. 3

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, 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. 2, 3, 4, 5, 6, 8

[46] 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. 2

[47] Peter Stockwell. *Texture-a cognitive aesthetics of reading*. Edinburgh University Press, 2020. 3

[48] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 2

[49] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Yujun Lin, Zhekai Zhang, Muyang Li, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 2, 3, 4

[50] 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*, pages 1191–1200, 2022. 3

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6