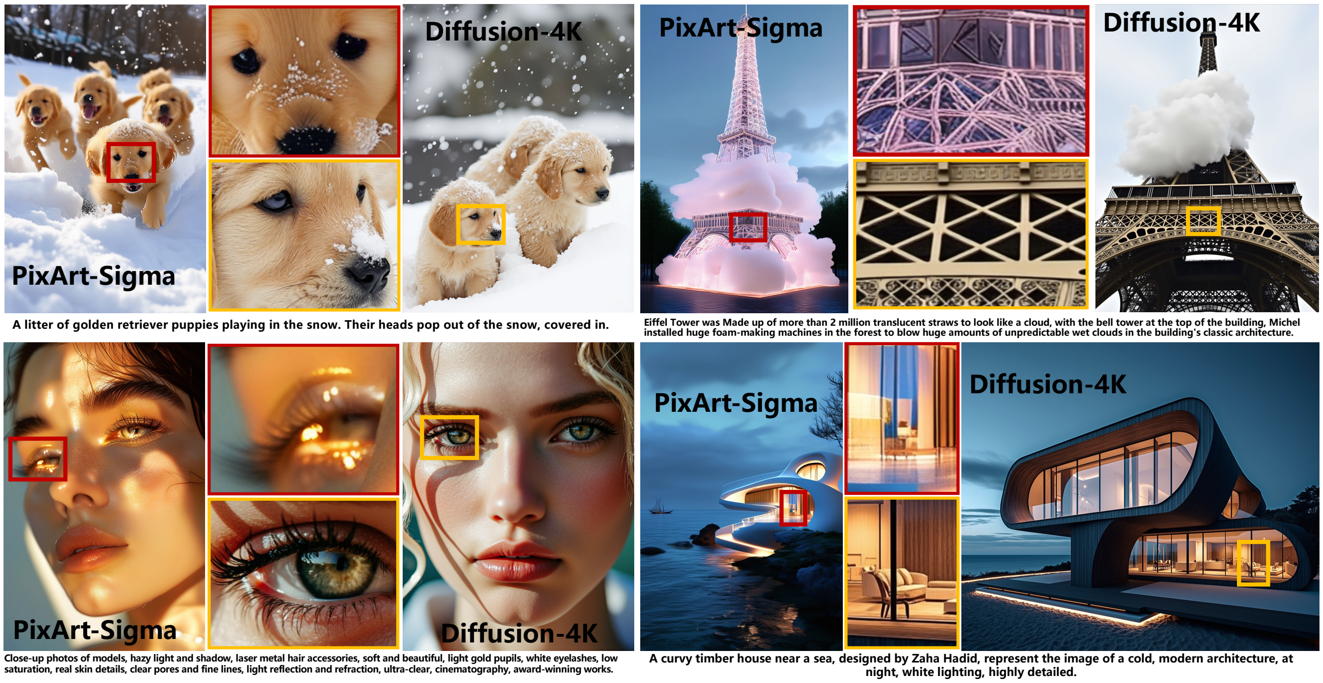
Diffusion-4K: Ultra-High-Resolution Image Synthesis with Latent Diffusion Models

Supplementary Material



A litter of golden retriever puppies playing in the snow. Their heads pop out of the snow, covered in.

Eiffel Tower was Made up of more than 2 million translucent straws to look like a cloud, with the bell tower at the top of the building, Michel installed huge foam-making machines in the forest to blow huge amounts of unpredictable wet clouds in the building's classic architecture.

Close-up photos of models, hazy light and shadow, laser metal hair accessories, soft and beautiful, light gold pupils, white eyelashes, low saturation, real skin details, clear pores and fine lines, light reflection and refraction, ultra-clear, cinematography, award-winning works.

A curvy timber house near a sea, designed by Zaha Hadid, represent the image of a cold, modern architecture, at night, white lighting, highly detailed.

Figure 7. We present comparisons with PixArt-Σ [9] using identical prompts, with images from PixArt-Σ displayed on the left and those synthesized by our Diffusion-4K shown on the right. Our approach demonstrates significant superiority over PixArt-Σ in fine details, as evidenced by the yellow patches *vs*. the red patches.

In the supplementary material, we provide additional details omitted from the main paper due to space constraints.

7. Comparisons

Performance Comparisons. We present detailed comparisons with other direct ultra-high-resolution image synthesis methods, including PixArt-Σ [9] and Sana [49]. As shown in Fig. 7, our Diffusion-4K consistently produces impressive fine details in generated images under identical prompts, highlighting its superiority over PixArt-Σ [9] in 4K image generation. In addition, we compare the generated images of our Diffusion-4K with Sana [49] in Fig. 8. Notably, PixArt-Σ generates ultra-high-resolution images exclusively at a resolution of 3840×2160, and the images of PixArt-Σ and Sana are sourced directly from their official websites.

Qualitative Evaluation of WLF. As shown in Eq. (4), WLF decomposes the latent into high- and low-frequency components, enabling the model to refine details (high-frequency) while preserving the overall structure (low-frequency). This decomposition not only enhances the model's capability to generate fine details but also ensures that the changes don't disrupt the underlying patterns, making the fine-tuning process both efficient and precise. To comprehensively assess WLF, we additionally provide qualitative comparisons of latent fine-tuning with and without WLF to demonstrate its effectiveness. As illustrated in Fig. 9, images generated with WLF exhibit richer details compared to those generated without WLF.

More Results of 4K Image Synthesis. In Fig. 10, we present the generated 4K images with different text prompts, demonstrating the effectiveness of our Diffusion-4K method in terms of visual aesthetics, adherence to text prompts and fine details. Additionally, as displayed in Fig. 11 and Fig. 12, we provide the prompts and corresponding synthesized 4K images with varying aspect ratios, random seeds and spelled texts, illustrating diversity within the generated images. Qualitative results significantly demonstrate the impressive performance of our approach in 4K image generation, with particular attention to fine details.

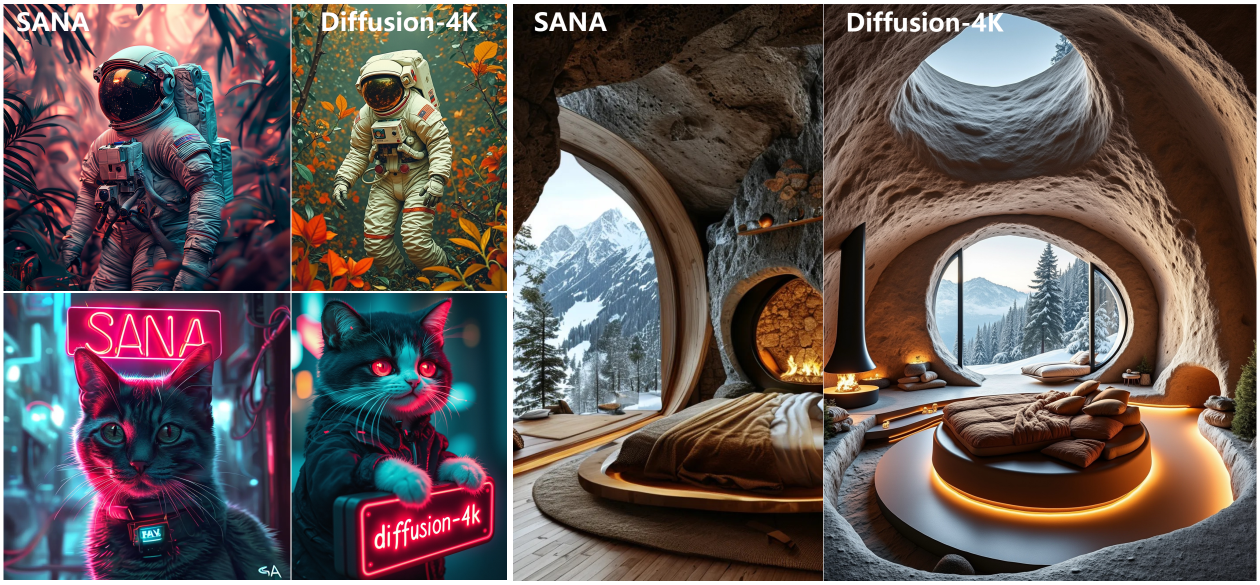
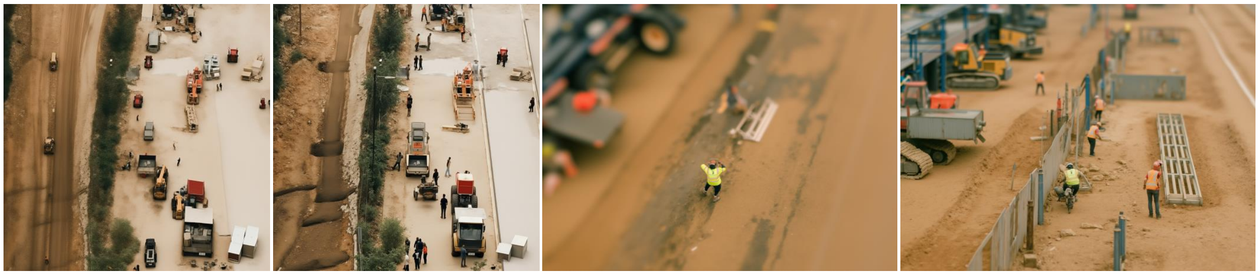


Figure 8. Comparisons with Sana [49].



fine-tuning w/o WLF | fine-tuning w/ WLF | fine-tuning w/o WLF | fine-tuning w/ WLF

Figure 9. Ablation study on WLF.

8. Aesthetic-4K Dataset

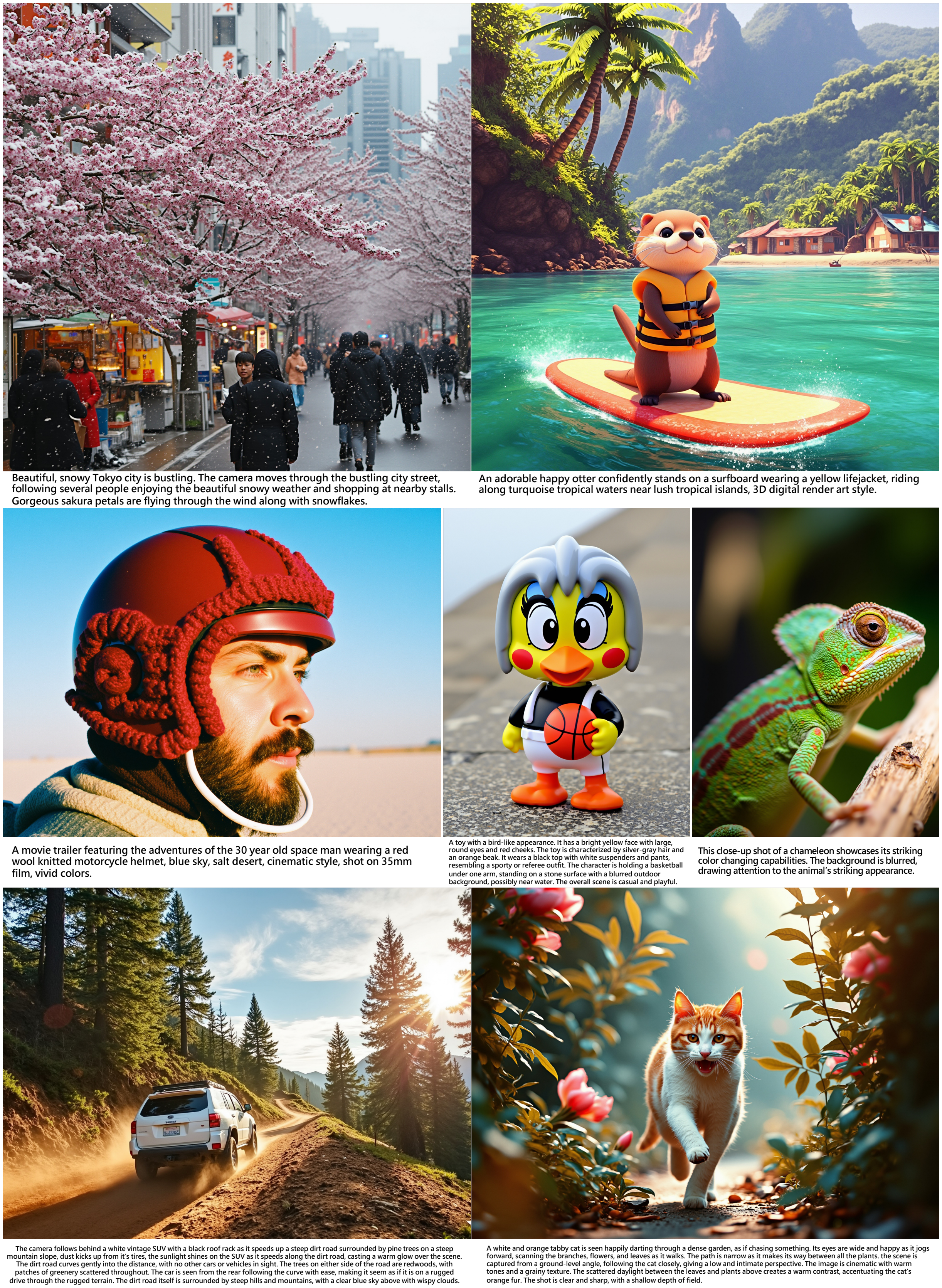
Details of Aesthetic-4K Dataset. The LAION-Aesthetics dataset contains approximately 0.03% 4K images, demonstrating the scarcity of 4K images in open-source datasets. In the open-source PixArt-30k dataset [9], the median height and width are 1615 and 1801 pixels, respectively. In comparison, the Aesthetic-Train dataset has a median height and width of 4128 and 4640 pixels, respectively, marking a substantial improvement, as shown in Tab. 9. Additionally, we have meticulously filtered out low-quality images through manual inspection, excluding those with motion blur, focus issues, and mismatched text prompts, *etc*. We believe that constructing a high-quality training dataset at 4K resolution is one of the critical factors for 4K image generation.

We provide the statistical histograms of image height and width for the Aesthetic-4K dataset in Fig. 13, highlighting the significant improvement in image resolution. In addition, we present a word cloud of image captions from the Aesthetic-4K dataset in Fig. 14, providing a visual representation of textual distribution.

Dataset	Median height	Median width	Average height	Average width
PixArt-30k	1615	1801	2531	2656
Aesthetic-Train	4128	4640	4578	4838
Aesthetic-Eval@2048	2983	3613	3143	3746
Aesthetic-Eval@4096	4912	6449	5269	6420

Table 9. Statistical comparisons of Aesthetic-4K and PixArt-30k.

Qualitative Samples in Aesthetic-4K Dataset. As depicted in Fig. 15, we provide more image-text samples from the training set of our Aesthetic-4K dataset to illustrate its diversity and richness. As previously noted, the Aesthetic-4K dataset stands out due to its exceptional quality, presenting ultra-high-resolution images paired with precisely generated captions by GPT-4o [19].



Beautiful, snowy Tokyo city is bustling. The camera moves through the bustling city street, following several people enjoying the beautiful snowy weather and shopping at nearby stalls. Gorgeous sakura petals are flying through the wind along with snowflakes.

An adorable happy otter confidently stands on a surfboard wearing a yellow lifejacket, riding along turquoise tropical waters near lush tropical islands, 3D digital render art style.

A movie trailer featuring the adventures of the 30 year old space man wearing a red wool knitted motorcycle helmet, blue sky, salt desert, cinematic style, shot on 35mm film, vivid colors.

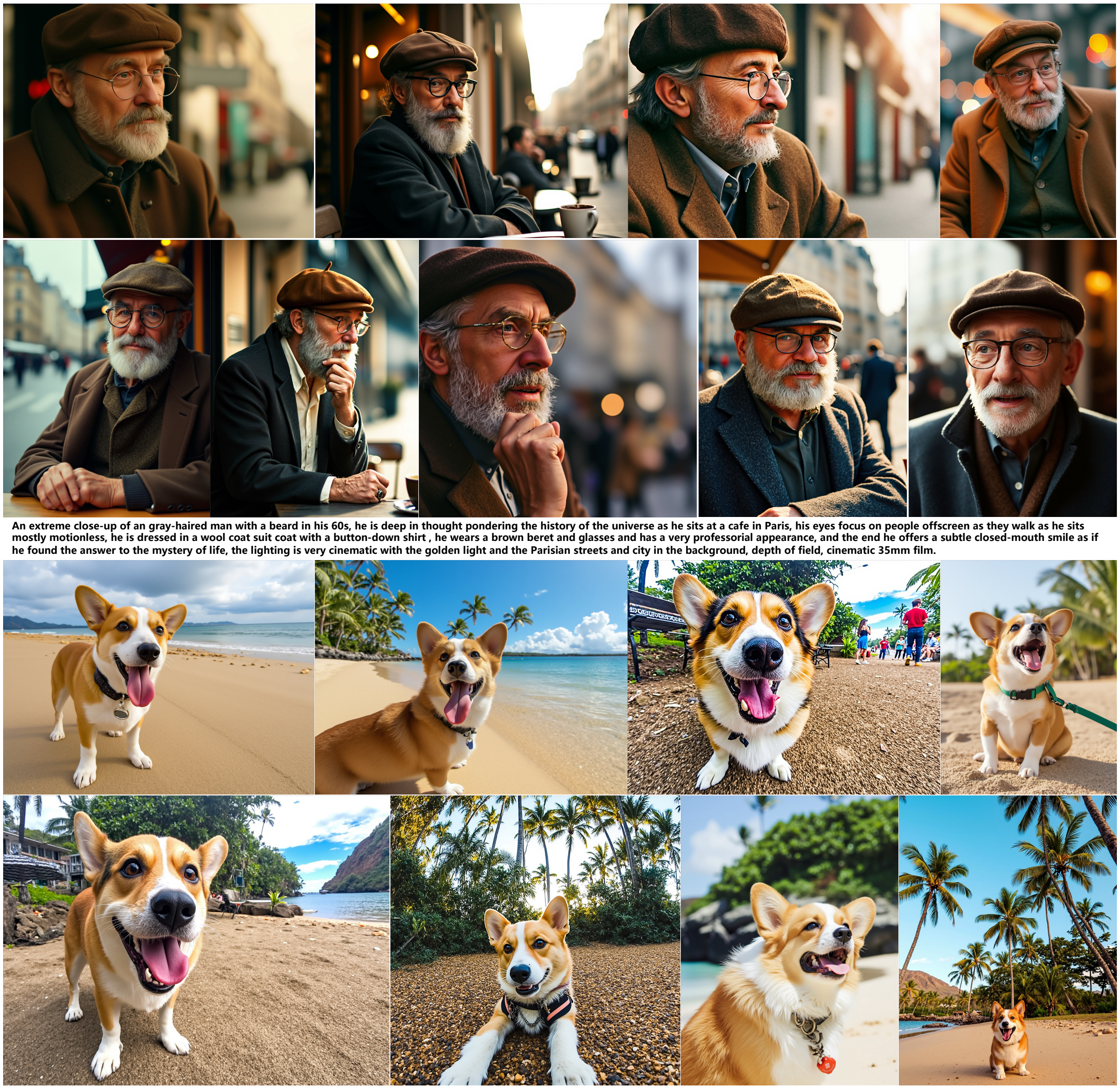
A toy with a bird-like appearance. It has a bright yellow face with large, round eyes and red cheeks. The toy is characterized by silver-gray hair and an orange beak. It wears a black top with white suspenders and pants, resembling a sporty or referee outfit. The character is holding a basketball under one arm, standing on a stone surface with a blurred outdoor background, possibly near water. The overall scene is casual and playful.

This close-up shot of a chameleon showcases its striking color changing capabilities. The background is blurred, drawing attention to the animal's striking appearance.

The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up from it's tires, the sunlight shines on the SUV as it speeds along the dirt road, casting a warm glow over the scene. The dirt road curves gently into the distance, with no other cars or vehicles in sight. The trees on either side of the road are redwoods, with patches of greenery scattered throughout. The car is seen from the rear following the curve with ease, making it seem as if it is on a rugged drive through the rugged terrain. The dirt road itself is surrounded by steep hills and mountains, with a clear blue sky above with wispy clouds.

A white and orange tabby cat is seen happily darting through a dense garden, as if chasing something. Its eyes are wide and happy as it jogs forward, scanning the branches, flowers, and leaves as it walks. The path is narrow as it makes its way between all the plants. the scene is captured from a ground-level angle, following the cat closely, giving a low and intimate perspective. The image is cinematic with warm tones and a grainy texture. The scattered daylight between the leaves and plants above creates a warm contrast, accentuating the cat's orange fur. The shot is clear and sharp, with a shallow depth of field.

Figure 10. High-quality images synthesized by our Diffusion-4k.



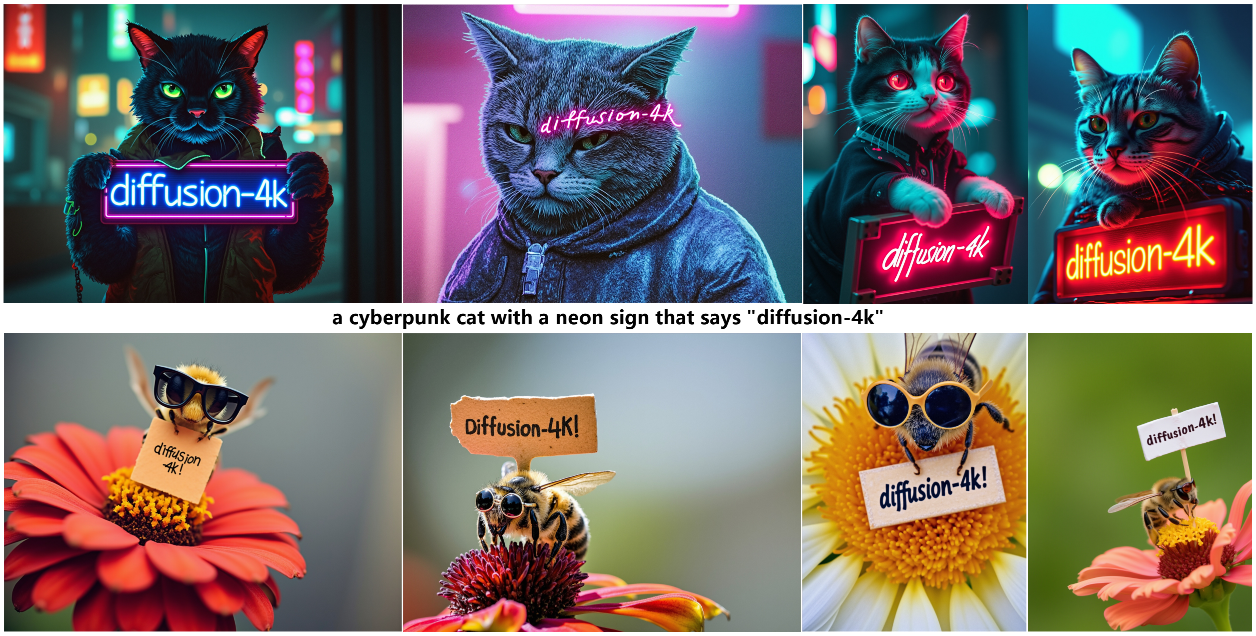
An extreme close-up of an gray-haired man with a beard in his 60s, he is deep in thought pondering the history of the universe as he sits at a cafe in Paris, his eyes focus on people offscreen as they walk as he sits mostly motionless, he is dressed in a wool coat suit coat with a button-down shirt , he wears a brown beret and glasses and has a very professorial appearance, and the end he offers a subtle closed-mouth smile as if he found the answer to the mystery of life, the lighting is very cinematic with the golden light and the Parisian streets and city in the background, depth of field, cinematic 35mm film.

A corgi vlogging itself in tropical Maui.

Figure 11. Synthesized images with different aspect ratios and random seeds.

9. More Details

Training Details of Diffusion-4K. During image preprocessing, images are resized to a shorter dimension of 4096, randomly cropped to a 4096×4096 resolution, and normalized with a mean and standard deviation of 0.5. Our partitioned VAE compresses the pixel space $\mathbb{R}^{H \times W \times 3}$ into a latent space $\mathbb{R}^{\frac{H}{F} \times \frac{W}{F} \times C}$, where $F = 16$. The encoded latents are normalized using the mean and standard deviation from the pretrained latent diffusion models, which are globally computed over a subset of the training data. The latent diffusion models are then optimized using the wavelet-based latent fine-tuning objective in Eq. (4). Regarding the text encoder, both CLIP [35] and T5-XXL [36] serve as the default models for text comprehension in SD3 [12] and Flux [5]. To conserve memory, text embeddings for latent diffusion models are pre-computed, eliminating the need to load text encoders into the GPU during the training phase. We employ a default patch size of $P = 2$ for DiTs, including SD3-2B and Flux-12B. Latent diffusion models are optimized using the WLF objective with all parameters



a cyberpunk cat with a neon sign that says "diffusion-4k"

A macro shot of a flower with a bee wearing sunglasses on it that holds a sign saying: "diffusion-4k!"

Figure 12. Synthesized images with spelled texts.

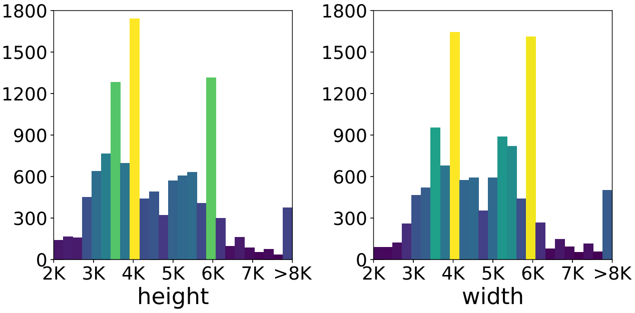


Figure 13. Histograms of image height and width in Aesthetic-4K.

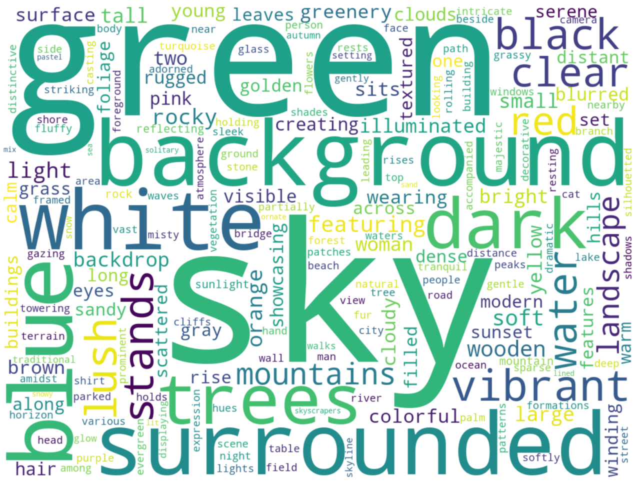


Figure 14. Word cloud of image captions from Aesthetic-4K.

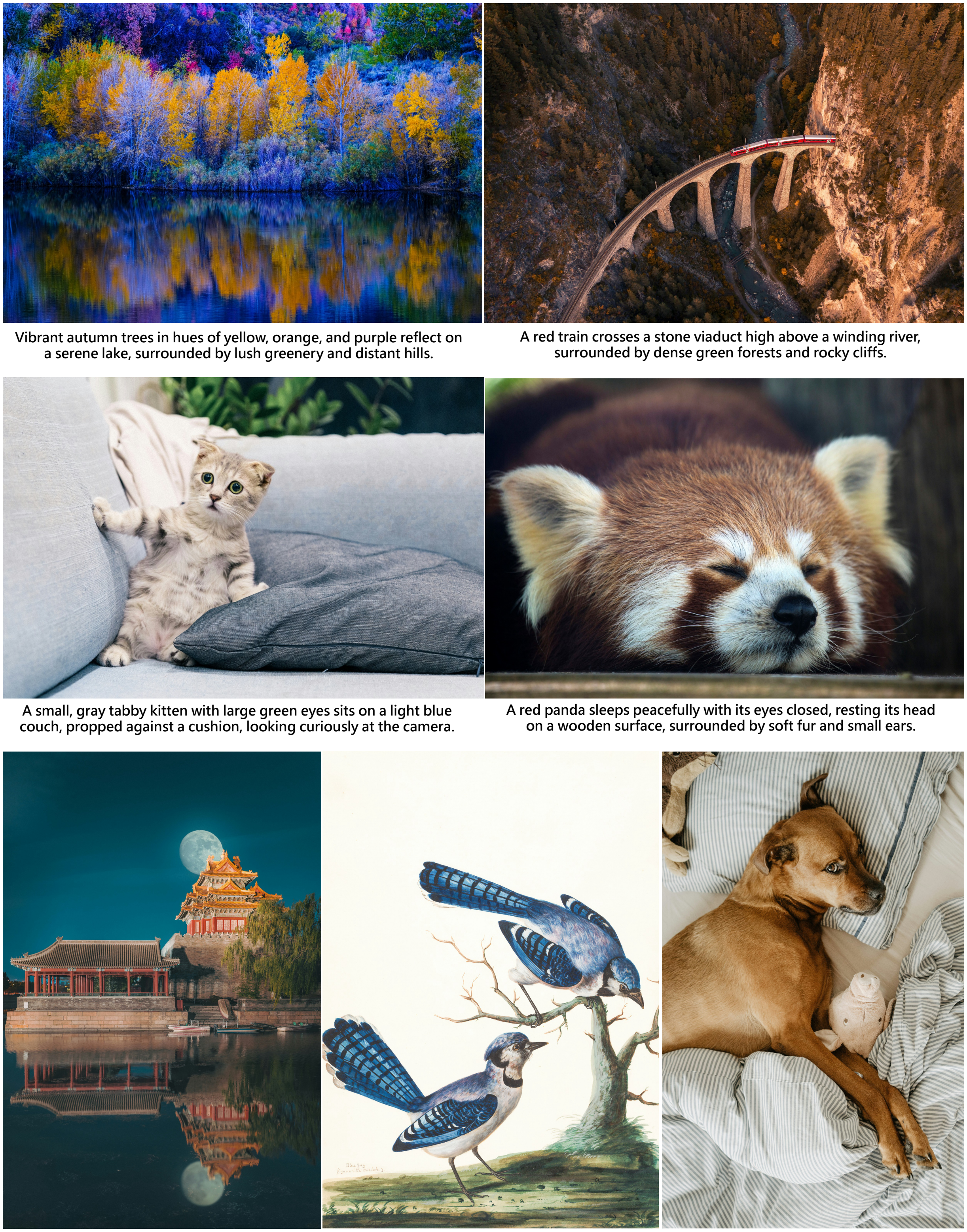
Model	SD3-2B-WLF	Flux-12B-WLF
Training steps	20K	20K
Throughput (images/s)	0.59	1.39

Table 10. Training details of SD3-2B and Flux-12B with WLF at 4096 × 4096.

unfrozen, whereas text encoders and the partitioned VAE remain fixed during training. Additionally, as shown in Tab. 10, we provide training details with SD3-2B and Flux-12B, including training steps and throughput, demonstrating the efficiency of our WLF method in handling scalable DiTs at ultra-high resolutions. Our approach requires approximately 2,000 A100 GPU hours to fine-tune a 12B diffusion model, demonstrating high computational efficiency while minimizing resource consumption. Note that we use the open-source Flux.1-dev version, which is trained using guidance distillation, and adopt the default guidance scale of 3.5 for WLF.

Detailed Prompts for GPT-4o. In Tab. 11, we provide detailed prompts for image caption using GPT-4o, to generate precise text prompts for the Aesthetic-4K dataset. Additionally, we present detailed prompts used in the preference study with GPT-4o, to evaluate AI preferences for generated images, including visual aesthetics, prompt adherence and fine details.

Details for Evaluation. During evaluation, images are gen-



Vibrant autumn trees in hues of yellow, orange, and purple reflect on a serene lake, surrounded by lush greenery and distant hills.

A red train crosses a stone viaduct high above a winding river, surrounded by dense green forests and rocky cliffs.

A small, gray tabby kitten with large green eyes sits on a light blue couch, propped against a cushion, looking curiously at the camera.

A red panda sleeps peacefully with its eyes closed, resting its head on a wooden surface, surrounded by soft fur and small ears.

A traditional Chinese pagoda sits elevated above a calm body of water, reflecting the structure and a large moon in the night sky, while a smaller pavilion and several boats are nearby.

Two blue jays with striking blue and black feathers perch on a bare, twisted branch, one in an upright position while the other leans forward on the branch, against a light background.

A brown dog lies comfortably on a striped bed, resting its head on a pillow and cuddling a plush pig toy among wrinkled bedding.

Figure 15. High-quality image-text samples in our Aesthetic-4K Dataset.

Tasks	Prompts
Image Caption	{"**text**": "Directly describe with brevity and as brief as possible the scene or characters without any introductory phrase like 'This image shows', 'In the scene', 'This image depicts' or similar phrases. Just start describing the scene please." }
Preference Study	{"**system**": "As an AI visual assistant, you are analyzing two specific images. When presented with a specific caption, it is required to evaluate visual aesthetics, prompt coherence and fine details.", "**text**": "The caption for the two images is: ⟨prompt⟩. Please answer the following questions: 1. Visual Aesthetics: Given the prompt, which image is of higher-quality and aesthetically more pleasing? 2. Prompt Adherence: Which image looks more representative to the text shown above and faithfully follows it? 3. 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. Please respond me strictly in the following format: 1. Visual Aesthetics: ⟨the first image is better⟩ or ⟨the second image is better⟩. The reason is ⟨give your reason here⟩. 2. Prompt Adherence: ⟨the first image is better⟩ or ⟨the second image is better⟩. The reason is ⟨give your reason here⟩. 3. Fine Details: ⟨the first image is better⟩ or ⟨the second image is better⟩. The reason is ⟨give your reason here⟩. "}

Table 11. Designed prompts for image caption and preference study with GPT-4o.

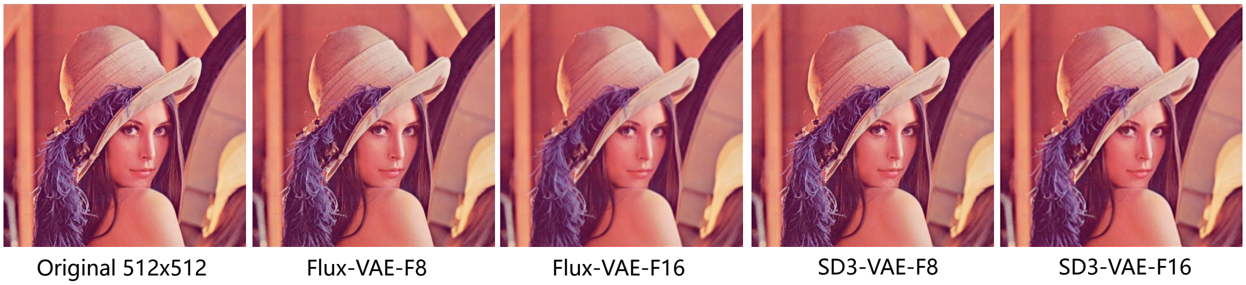


Figure 16. Reconstruction results of partitioned VAEs at 512 × 512.

erated using a guidance scale of 7.0, and flow-matching sampling introduced in SD3 [12], with 28 sampling steps for SD3-2B and 50 sampling steps for Flux-12B. The FID [17], Aesthetics [45] and CLIPScore [16] are computed using resized images at a fixed low resolution.

To quantitatively analyze the alignment between our indicators and human ratings, five participants are asked to rate extracted patches on a scale from 1 to 10 based on visual details, with the average scores used to evaluate SRCC and PLCC as presented in Tab. 1. Note that the SRCC and PLCC for the Compression Ratio are calculated using its reciprocal.

For human preference evaluation on 4K image synthesis in Fig. 6, we conduct experiment with 112 text prompts sampled from Sora [30], PixArt [9], SD3 [12], *etc*. Ten participants are asked to rate the preference for the generated images in visual aesthetics, prompt adherence and fine details respectively.

Reconstruction Results of Partitioned VAEs. We provide the reconstruction results with images at a resolution of 512 × 512 in Fig. 16, showcasing the capability of partitioned VAEs in handling low-resolution images.