

---

# UltraHR-100K: Enhancing UHR Image Synthesis with A Large-Scale High-Quality Dataset

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Appendix

Table 1: **Quantitative comparison with other baselines on our UltraHR-eval2K ( $2048 \times 2048$ ) benchmark.** The best result is highlighted in **bold**.

| Method               | Year        | FID ↓         | FID <sub>patch</sub> ↓ | IS ↑          | IS <sub>patch</sub> ↑ | CLIP ↑       |
|----------------------|-------------|---------------|------------------------|---------------|-----------------------|--------------|
| Pixart- $\sigma$ [1] | 2024        | 34.472        | 24.890                 | 13.883        | 6.981                 | 31.84        |
| SANA [2]             | 2025        | 39.540        | 34.351                 | 14.064        | 7.089                 | <b>31.92</b> |
| Diffusion4K [3]      | 2025        | 34.591        | 15.707                 | 14.664        | 5.358                 | 31.74        |
| <b>Ours</b>          | <b>2025</b> | <b>34.234</b> | <b>14.252</b>          | <b>15.439</b> | <b>8.033</b>          | 31.85        |

## 2 A More Comparisons

Figures 2 to 4 present additional high-resolution image generation results produced by our proposed method. These visual examples further demonstrate the capability of our approach to synthesize realistic and visually compelling images with fine-grained details and consistent structure, even under challenging scenarios. Furthermore, Figures 5 to 9 provide comprehensive visual comparisons between our method and several state-of-the-art (SOTA) approaches on our UltraHR-eval4K. As shown, our method consistently produces sharper textures, more accurate structural results, and superior perceptual quality compared to existing baselines. These results further validate the effectiveness and robustness of our approach in handling complex high-resolution image generation tasks. Furthermore, we introduce UltraHR-eval2K( $2048 \times 2048$ ), a large-scale text-to-image (T2I) benchmark comprising 2000 UHR images, designed to facilitate a comprehensive evaluation of existing ultra-high-resolution generation models. As reported in Table 1, our approach consistently outperforms state-of-the-art methods in terms of quantitative metrics, demonstrating its superior capability in generating high-fidelity images. Notably, to ensure a fair comparison, we directly evaluate all methods—**including our method**—using their pre-trained models from UltraHR-eval4K, without any additional training or fine-tuning.

## 18 B Training Details

Our target is to enhance the UHR generation capability of pretrained T2I models using our proposed dataset and frequency-aware training strategies. Due to computational constraints, we focus on SANA as the base model. All training experiments are conducted on 4 H20 GPUs over a period of approximately 7 days. Our 4K training configuration largely follows the settings used in SANA’s original 4K training setup. For the proposed DOTS, we adopt a Beta distribution with parameters  $\alpha = 2$  and  $\beta = 4$  for timestep sampling. As illustrated in Figure 1, compared to the popular uniform and logit-normal sampling methods, beta sampling better biases the sampling distribution.

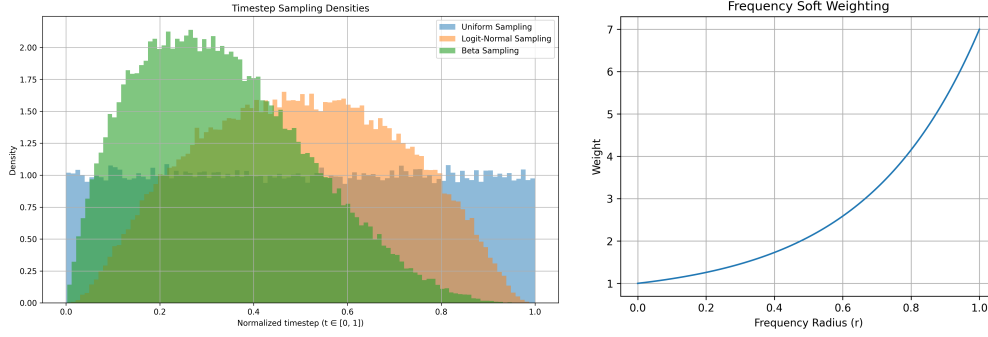


Figure 1: **Left:** Comparison of the beta sampling and the current mainstream sampling approaches. **Right:** Visualization of frequency soft weighting. By applying soft weighting in the frequency domain, better high-frequency control is achieved.

By adjusting  $\alpha$  and  $\beta$ , we can control the sampling bias: when  $\alpha < \beta$ , the distribution favors later denoising steps (closer to  $t = 0$ ); conversely,  $\alpha > \beta$  biases the sampling toward earlier steps (closer to  $t = 1$ ). For the SWFR, we employ an exponential weighting scheme, which emphasizes learning in the high-frequency domain. This weighting is particularly effective for enhancing fine-grained detail synthesis. In our experiments, we set the regularization hyperparameters to  $\lambda = 6$  and  $\gamma = 3$ . Figure 1 demonstrates that our soft weighting mechanism enables fine-grained control over high-frequency signals, where  $\lambda = 6$  and  $\gamma = 3$ .

## C Data Processing Pipeline

To construct the UltraHR-100K, we design a multi-stage data processing pipeline to ensure both scale and quality.

**Stage I: Data Collection and Preliminary Filtering.** We begin by collecting 400K high-resolution images from diverse online sources. To ensure a baseline level of visual quality, we apply preliminary filtering based on edge sharpness and contrast using Laplacian and Sobel operators. This step removes overly blurred, low-information, or artifact-heavy images, resulting in an initial subset  $S$  with acceptable visual characteristics.

**Stage II: Parallel Quality-Based Filtering.** To further enhance quality, we conduct three parallel filtering processes: *Detail Richness:* We compute GLCM statistics (e.g., contrast, entropy, correlation) across multiple directions to evaluate fine-grained texture complexity, forming subset  $S_G$ . *Content Complexity:* Using Shannon entropy, we measure the amount of information within an image to form  $S$  constitute subset  $S_E$ . *Aesthetic Quality:* We adopt the LAION Aesthetics Predictor to assess perceptual appeal to form subset  $S_A$ .

**Stage III: Dataset Finalization and Caption Annotation.** The final dataset is obtained by intersecting the three subsets:  $S_G \cap S_E \cap S_A$ , ensuring each image meets all three high-quality standards. Finally, we employ Gemini 2.0, a strong vision-language model, to annotate each image with a long and fine-grained caption, capturing both global semantics and localized details. Figure 10 and 11 show more high-quality data with fine-grained caption in our UltraHR-100K.

## 52 References

- 53 [1] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping  
54 Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\sigma$ : Weak-to-strong training of diffusion transformer  
55 for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91.  
56 Springer, 2024.
- 57 [2] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang,  
58 Muiyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with  
59 linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.
- 60 [3] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-  
61 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF*  
62 *Conference on Computer Vision and Pattern Recognition*, 2025.



Figure 2: High-quality images synthesized by our method.





Figure 3: High-quality images synthesized by our method.



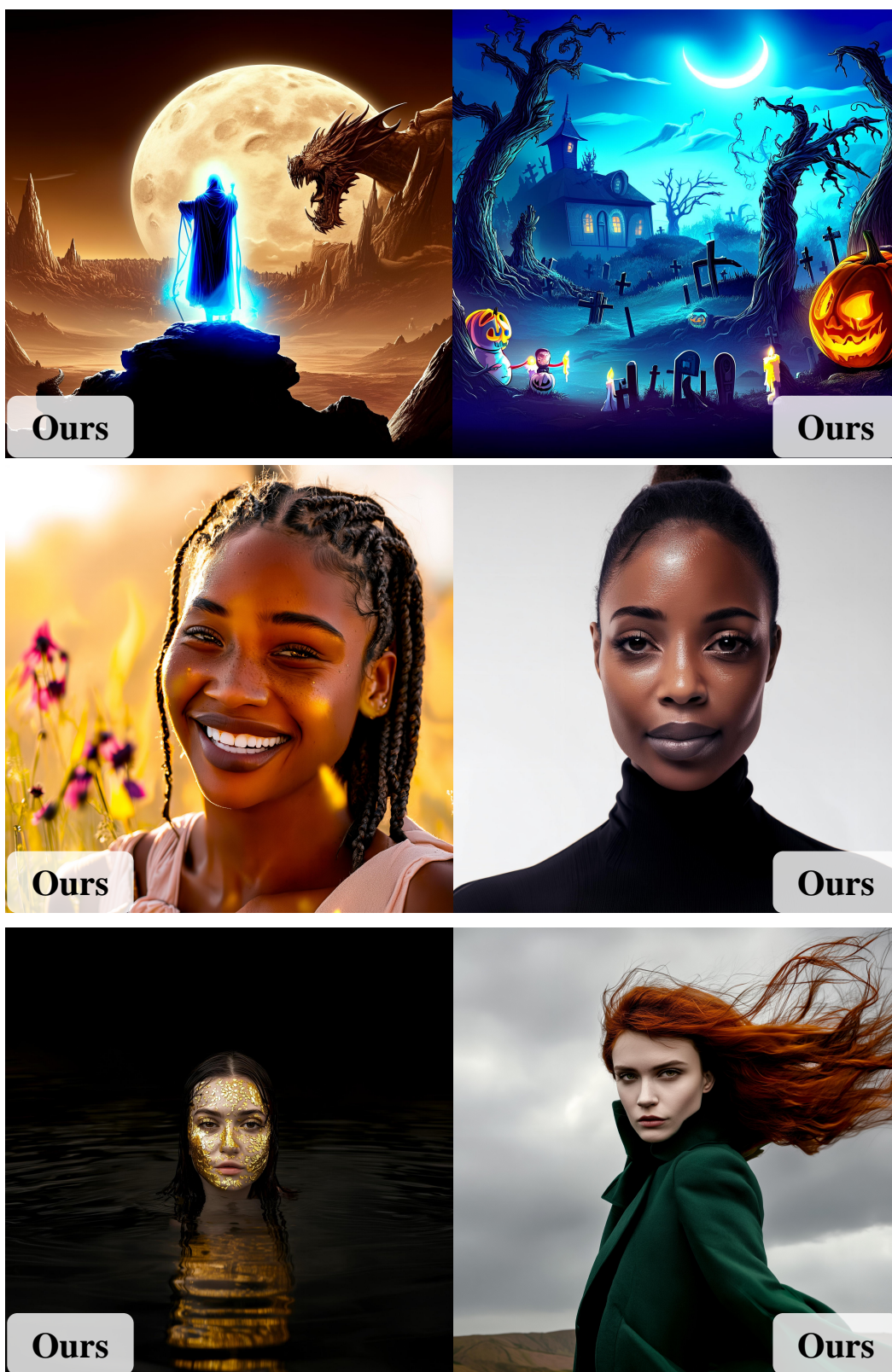


Figure 4: High-quality images synthesized by our method.



Figure 5: Qualitative comparisons with SOTA methods on our UltraHR-eval4K (4096×4096). "Prompt": "A panoramic view captures a majestic mountain range, dominated by snow-covered peaks and rocky formations under a brilliant blue sky. The scene unfolds with an array of elements, from the rugged, gray rocks in the foreground to the expansive fields of snow leading up to the imposing, craggy mountain peaks. A weathered wooden structure nestles in the snowy landscape, contrasting with the technological presence of a metallic weather station perched on the rocky edge. In the distance, rolling hills fade into a hazy horizon, marked by a solitary cloud that adds depth to the scene. A wooden post with rope is visible in the right corner."





Figure 6: Qualitative comparisons with SOTA methods on our UltraHR-eval4K (4096×4096). "Prompt": "A ruggedly handsome man in period clothing stands on a coastal cliff, gazing out at the sea. He wears a brown coat over a gray waistcoat and tan breeches, his dark, curly hair tousled by the wind, creating a romantic and windswept appearance. The dramatic landscape includes rocky cliffs covered with sparse vegetation and the blue sea stretching into the distance, contrasted against a sky with scattered clouds. The lighting is soft and natural, emphasizing the contours of the man's face and the textures of his clothing. The overall tone is one of solitary contemplation amid a vast and untamed setting, reminiscent of historical dramas. The composition balances the figure with the landscape, creating a sense of scale and isolation."



Figure 7: Qualitative comparisons with SOTA methods on our UltraHR-eval4K (4096×4096). "Prompt": "A vast, mountainous landscape is depicted under a dynamic sky, with snow-capped peaks transitioning into lower elevations of golden meadows and scattered trees. The mountain range dominates the scene, partially covered in snow and rocky outcrops, reaching towards the sky filled with dramatic, billowing clouds. The lower slopes feature a mix of golden grasses and sparse vegetation, giving way to small patches of trees turning color with the season. Sunlight breaks through the cloud cover, creating highlights and shadows that enhance the three-dimensional aspect of the scene. The artistic style captures the natural beauty and ruggedness of the mountain environment, emphasizing the contrast between the cool tones of the snow and the warm hues of the meadows below, creating a sense of depth and grandeur."



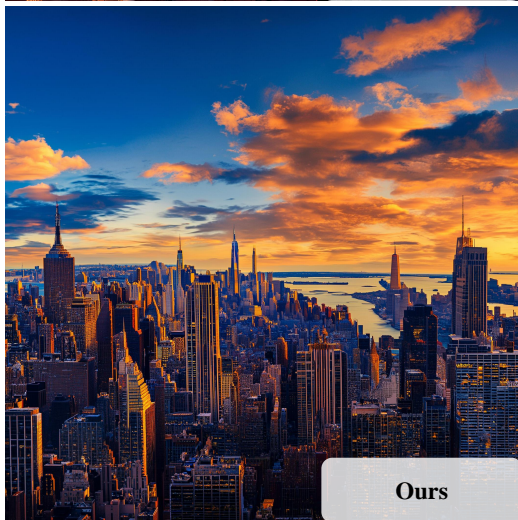


Figure 8: Qualitative comparisons with SOTA methods on our UltraHR-eval4K (4096×4096). "Prompt": "This striking panorama captures the iconic Manhattan skyline at sunset, showcasing the city's architectural grandeur under a dramatic sky. From left to right, recognizable skyscrapers pierce the horizon, including the MetLife Building and the distinctive silhouette of the Empire State Building, set against a backdrop of vibrant clouds streaked with golden light. A new skyscraper is still under construction. The city's dense urban fabric stretches into the distance, punctuated by other notable landmarks. The sky shifts from deep blues to warm oranges, reflecting off the buildings and the bodies of water surrounding the island. "



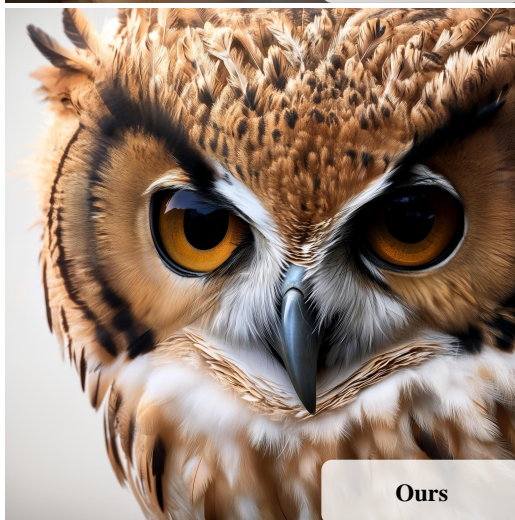


Figure 9: Qualitative comparisons with SOTA methods on our UltraHR-eval4K (4096×4096). "Prompt": "An owl's intense gaze captures attention in this detailed portrait. Its face, a mix of warm brown and soft white feathers, is framed by large, dark eyes, reflecting small, bright highlights. A subtle, downward-curving beak adds to its stern expression, the texture of its plumage detailed and slightly ruffled. The backdrop is a soft, muted blend, which further emphasizes the owl's captivating presence. The portrait showcases the bird's wild beauty through sharp focus and natural lighting, enhancing every detail from its piercing gaze to the delicate feather structure."



A panoramic view captures a rugged canyon landscape, bathed in warm sunlight highlighting layered rock formations and sparse vegetation. The canyon walls, composed of stratified rock, reveal a range of beige and orange hues, with small patches of dark green shrubs and trees scattered across the terrain. Some areas retain remnants of snow, indicating a recent or lingering cold spell. The varying light and shadow accentuate the depth and texture of the landscape, creating a striking contrast between the sunlit peaks and the darkened recesses of the canyon.

UltraHR-100K



A majestic eagle is captured in a detailed portrait, showcasing its intense gaze and intricate feather patterns. The eagle's head, marked by striking black-and-white striped plumage, turns to the right, revealing a piercing yellow eye framed by a dark, stern brow. Its black beak, tipped with a vibrant yellow, adds to the bird's powerful and predatory profile. Below the neck, the feathers cascade in shades of brown and burnt orange, creating a rich tapestry of texture and depth. The background, a muted blend of blues and grays, keeps the focus firmly on the bird.

UltraHR-100K



A striking portrait captures a woman with purple hair, framed by a fur-lined hood and set against a softly blurred snowy background. Her gaze is direct and compelling, enhanced by subtle makeup and a nose ring. The composition is symmetrical, drawing attention to her face and emphasizing the texture of the fur hood. The sleeves of her garment feature a unique patchwork design with blue and light gray fabrics, adding an element of modern style against the wintry scene. The blurred snowflakes contribute to the atmospheric depth, giving a sense of movement and realism to the otherwise stylized portrait.

Figure 10: More data in our UltraHR-100K. Captions provide more expressive descriptions, encompassing not only global summaries of the image content but also rich details.





UltraHR-100K

A serious-looking brown tabby cat rests comfortably, its intense gaze directed at the viewer. The cat's fur features rich, brown stripes across its head and body, complemented by piercing amber eyes that demand attention. Its whiskers are long and white, framing a determined expression. The background is softly blurred, creating a shallow depth of field that emphasizes the cat's prominent features and imposing stare. This composition evokes a regal and slightly intimidating presence, highlighting the cat's natural beauty and assertive demeanor through a sharp focus and warm color palette.



UltraHR-100K

A tabletop display of crepes with toppings and beverages creates a warm, inviting still life. Golden-brown crepes, folded into triangles, are artistically arranged on plates, one embellished with fresh strawberries, a dollop of cream, and almond slivers, inviting immediate indulgence. Adjacent to the main plate, a small bowl of pure white cream offers a blank canvas for customization, while glasses filled with amber-hued liquid suggest a choice between tea and syrup, their intricate glass patterns catching the light. Soft shadows cast across the textured surface add depth and dimension to the composition. The overall color palette, with its blend of warm browns, creamy whites, and occasional pops of red, creates a comforting yet elegant visual narrative.



UltraHR-100K

A young woman with dark hair and striking features poses beside a tree, her intense gaze directed towards the viewer. The soft, natural light catches the curves of her face, highlighting her dark eyes and the rich burgundy of her lipstick. Her dark, wavy hair cascades around her shoulders, some strands catching the golden light, while the rough texture of the tree bark provides a stark contrast to her smooth skin. She is dressed in a black turtleneck, adding a touch of elegance to the outdoor setting. The background is a blur of autumn colors, creating a soft, ethereal backdrop that complements the portrait's intimate and contemplative mood.

Figure 11: More data in our UltraHR-100K. Captions provide more expressive descriptions, encompassing not only global summaries of the image content but also rich details.