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

Chen Zhao[1,*], En Ci[1,*], Yunzhe Xu[1,*], Tiehan Fan[1], Shanyan Guan[2],

Yanhao Ge[2], Jian Yang[1], Ying Tai[1,†]

[1] State Key Laboratory of Novel Software Technology, Nanjing University, China
[2] vivo Mobile Communication Co., Ltd., China

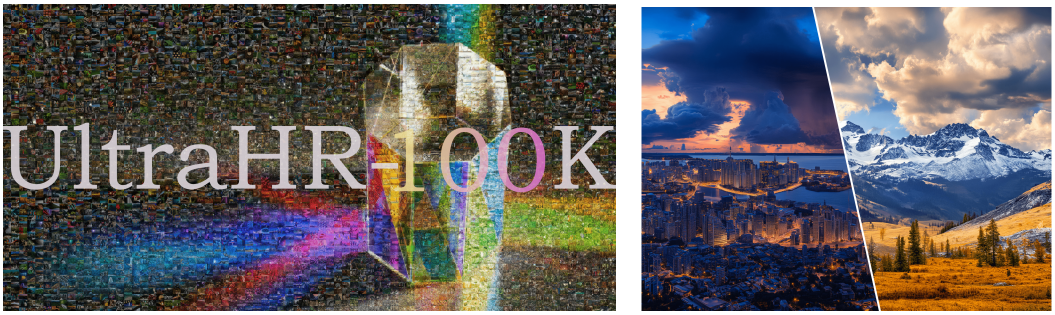

Figure 1: Our UltraHR-100K (**left**) is a large-scale high-quality dataset for ultra-high-resolution (UHR) image synthesis, featuring a diverse range of categories. Utilizing this dataset enables the generation of high-fidelity UHR images (**right**).

Abstract

Ultra-high-resolution (UHR) text-to-image (T2I) generation has seen notable progress. However, two key challenges remain : 1) the absence of a large-scale high-quality UHR T2I dataset, and (2) the neglect of tailored training strategies for fine-grained detail synthesis in UHR scenarios. To tackle the first challenge, we introduce **UltraHR-100K**, a high-quality dataset of 100K UHR images with rich captions, offering diverse content and strong visual fidelity. Each image exceeds 3K resolution and is rigorously curated based on detail richness, content complexity, and aesthetic quality. To tackle the second challenge, we propose a frequency-aware post-training method that enhances fine-detail generation in T2I diffusion models. Specifically, we design (i) *Detail-Oriented Timestep Sampling (DOTS)* to focus learning on detail-critical denoising steps, and (ii) *Soft-Weighting Frequency Regularization (SWFR)*, which leverages Discrete Fourier Transform (DFT) to softly constrain frequency components, encouraging high-frequency detail preservation. Extensive experiments on our proposed UltraHR-eval4K benchmarks demonstrate that our approach significantly improves the fine-grained detail quality and overall fidelity of UHR image generation. The code is available at here.

[*] Equal Contribution.
[†] Correspondence to: Ying Tai.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Recent advances in text-to-image (T2I) diffusion models have greatly improved image quality and controllability [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. However, most existing models are constrained to fixed resolutions (typically 1024×1024), and exhibit noticeable quality degradation and structural artifacts when directly scaled to ultra-high-resolution (UHR) image generation [12, 13, 14, 15, 16, 17, 18, 19]. This limitation poses a significant barrier for real-world applications that demand fine-grained detail and high visual fidelity, such as digital art, virtual content creation, and commercial design.

Existing solutions to face this challenge can be grouped into two main paradigms: training-free [13, 14, 20, 21, 22, 23, 24, 25, 26] and training-based methods [12, 15, 16, 17, 27]. Training-free methods attempt to generate UHR images by modifying network architectures [20, 22, 23] or by adjusting inference schemes [14, 21]. However, these techniques exhibit excessive smoothing, produce implausible details, and incur prolonged inference times—limitations that severely hinder their practical deployment [12]. Fundamentally, training-free methods depend on pre-trained T2I models [2, 3, 5, 7] that were not exposed to UHR data during training, and consequently *lack the inherent capacity to render the fine-grained, photorealistic details essential* that real-world UHR image synthesis requires.

Recently, training-based models for UHR image generation have shown promising results [15, 16, 17]. However, they still face two critical challenges: 1) *The absence of a open-source, large-scale high-quality UHR T2I dataset.* High-fidelity UHR image collection is burdensome due to the scarcity of suitable data. Although Aesthetic-4K [16] introduced the first open-source UHR T2I dataset, it remains limited in both scale (approximately 10K images) and quality (the lack of a rigorous selection criterion), constraining its generalizability and high-quality generation capabilities in real-world scenarios. Consequently, constructing a open-source, large-scale high-quality UHR T2I dataset represents both a significant challenge and a critical necessity. 2) *The neglect of tailored training strategies for UHR fine-grained detail synthesis.* Existing models primarily focus on training efficiency to fine-tune pre-trained T2I models [15, 17], overlooking the high-fidelity detail synthesis. Large-scale pre-training equips T2I models with strong semantic planning abilities, but they struggle to synthesize fine-grained details in the UHR setting [2, 3, 5, 7]. Thus, a detail-oriented training strategy is essential for achieving high-quality UHR image synthesis.

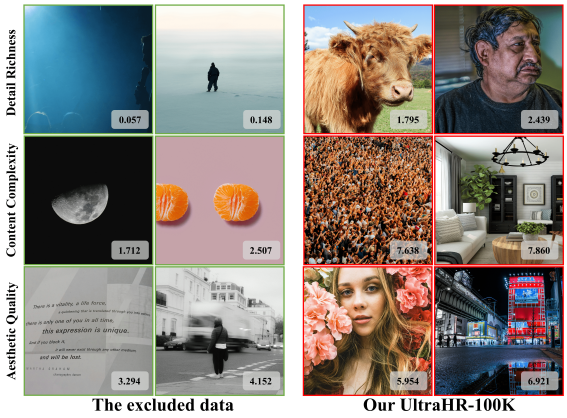

Figure 2: We perform a rigorous selection of UltraHR-100K by evaluating all collected images across three key dimensions: detail richness, content complexity, and aesthetic quality. **Left:** We present representative low-quality (bad case) examples for each dimension along with their corresponding scores, highlighting the necessity of such filtering. **Right:** In contrast, our UltraHR-100K exhibit superior texture details, semantic complexity, and aesthetic appeal.

Large-Scale High-Quality Dataset for Tackling Challenge 1: We construct UltraHR-100K, a large-scale high-quality UHR T2I dataset consisting of 100K UHR images paired with rich textual descriptions. As illustrated in Figure 1, UltraHR-100K offers the following key advantages: 1) *Scale and Diversity*: Compared to recent publicly available Aesthetic-4K [16], our UltraHR-100K is approximately 10× larger, featuring 100K images spanning a broad spectrum of categories and visual concepts. 2) *Higher Quality*: All images in UltraHR-100K are rigorously selected from three key dimensions: detail richness, content complexity, and aesthetic quality. Notably, the minimum resolution across the proposed dataset exceeds 3K (average of width and height), ensuring high-resolution content, as shown in Figure 3. 3) *Fine-Grained Captions*: To provide detailed textual annotations for each image, we leverage Gemini 2.0 [28], a powerful commercial vision-language model (VLM), to generate high-quality captions. As shown in Figure 4, our captions are significantly more detailed and semantically rich compared to those in the Aesthetic-4K [16].

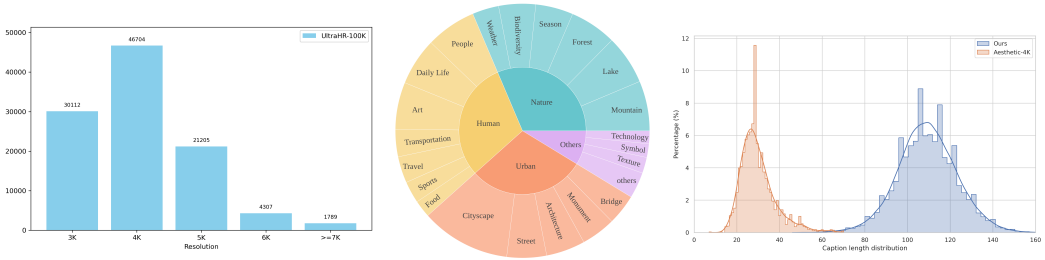

Figure 3: **Left:** Resolution distribution of our UltraHR-100K. All images have a minimum resolution of 3K, defined as the average of height and width exceeding 3000 pixels. **Middle:** Image categories across our dataset. *The proportion of each category mirrors its distribution in our dataset.* **Right:** Caption length distribution. Compared to the recent Aesthetic-4K[16], our captions are significantly longer, providing richer semantic supervision.

Table 1: Overview of our data processing pipeline. The first stage involves large-scale data collection and preliminary filtering to ensure a baseline level of visual quality. The second stage performs three parallel filtering procedures. The final high-quality dataset is obtained by taking the intersection of these subsets. We further employ a strong VLM (Gemini 2.0) to annotate each image.

Pipeline	Tool	Remark
Data collecting	Python	Get 400K high-resolution images
Preliminary data filtering	Laplacian and Sobel	Obtain subset S with basic visual quality
Detail richness	GLCM	Obtain the set S_G with rich fine-grained details
Content complexity	Shannon entropy	Obtain the set S_E with complex and diverse content
Aesthetic score	LAION aesthetic predictor	Get high aesthetic score set S_A
The final dataset	Intersection	Obtain intersection: UltraHR-100K = $S_A \cap S_E \cap S_G$
UHR image caption	Gemini 2.0	Obtain long and fine-grained descriptions for the images

Detail-Oriented Training Strategy for Tackling Challenge 2: To enhance UHR detail synthesis, we propose a frequency-aware post-training method, which consists of detail-oriented timestep sampling (DOTS) and soft-weighting frequency regularization (SWFR). DOTS improves detail synthesis in UHR image generation by directing more training focus to timesteps associated with fine-grained details. Unlike the discrete and block-based decomposition approach used in Diffusion4K [16], which relies on DWT for frequency separation, our SWFR utilizes the continuous spectrum provided by the Discrete Fourier Transform (DFT) to enable more precise frequency control. By applying a soft-weighted constraint across frequency bands, SWFR encourages the model to better reconstruct high-frequency details, without compromising low-frequency structural integrity.

Through the proposed ***dataset*** and ***training strategy***, we can ***enhance*** the synthesis capability of existing pre-trained T2I models [2, 3, 15, 7] in UHR image generation, with a particular focus on improving fine-grained detail representation. Furthermore, we construct a large 4K T2I benchmarks, UltraHR-eval4K (4096×4096), to comprehensively evaluate existing UHR generation models. Extensive experimental results demonstrate the effectiveness of our method.

2 Related Work

2.1 Text-to-Image Synthesis

Text-to-image (T2I) generation [1, 2, 3, 4, 5, 6, 7, 8, 29, 30, 31, 32, 33, 34, 35] has made notable progress owing to the emergence of diffusion-based frameworks [36, 37, 38, 39, 40, 41, 42, 43, 44, 45], which exhibit impressive ability in synthesizing visually compelling content from textual descriptions. Early methods such as Denoising Diffusion Probabilistic Models (DDPM) [46] and Denoising Diffusion Implicit Models (DDIM) [47] revealed the strength of iterative denoising procedures for producing realistic images. Subsequently, the attention to latent space diffusion [48] brought a major breakthrough, significantly lowering training complexity and enhancing scalability [2]. More recently, incorporating transformer [3, 7, 15, 49, 50] into diffusion models has further boosted image

generation quality. In this paper, we aim to enhance the generative capability of T2I models in UHR scenarios.

2.2 Ultra-High-Resolution Image Synthesis

UHR image generation plays a crucial role in practical domains such as industry and entertainment [16, 51, 52]. Due to computational constraints, current advanced latent diffusion models typically operate at a maximum resolution of 1024 × 1024 [2, 3, 6, 7, 8, 53, 54]. However, scaling to 4K resolution significantly increases computational demands, with cost growing quadratically with image size. Several training-free approaches have extended existing latent diffusion models for 4K generation by modifying the inference strategies of diffusion models. [13, 14, 55, 20, 21, 22, 23, 56]. DiffuseHigh [23] enhances the base-resolution generation by upscaling and subsequently re-denoising it, guided by structural information from the DWT. HiFlow [13] adopts a cascaded generation paradigm to effectively capture and utilize low-resolution flow characteristics. However, these techniques exhibit excessive smoothing, produce implausible details, and incur prolonged inference time [12]. Pixart-σ [17] takes a pioneering step by approaching direct 4K image generation through efficient token compression in DiT. Similarly, Sana [15] introduces a cost-effective 4K generation pipeline. Despite these advancements, existing models primarily focus on training efficiency, overlooking the high-fidelity detail synthesis.

3 Constructing UltraHR-100K

To face the challenge of the lack of high-quality text-image pairs at UHR image generation, we construct a large-scale high-quality dataset named UltraHR-100K. We begin by collecting approximately 400K high-resolution images (with a minimum resolution of 3840×2160) using a custom Python crawler built with Scrapy, sourcing images from the web and various high-resolution imaging devices. *However, high resolution alone does not guarantee high quality. We pose a central question:* ***What constitutes a high-quality image for UHR image generation?*** We argue that beyond resolution, such images should exhibit rich content complexity, fine-grained visual details, and aesthetic appeal. Accordingly, we conduct a rigorous filtering process based on three criteria—***content complexity, detail richness, and aesthetic quality***—to curate a 100K-level T2I dataset that meet these standards. The proposed UltraHR-100K provides a reliable foundation for training and evaluating models in high-fidelity UHR image generation. The data processing pipeline is provided in Table 1.

Table 2: Dataset statistical comparisons.

Dataset	Number	Height	Width
PixArt-30k [17]	30,000	1,615	1,801
Aesthetic-4K [16]	12,015	4,128	4,640
UltraHR-100K	104,117	3,648	5,119
Aesthetic-Eval@4096 [16]	195	4,912	6,449
UltraHR-eval4K	2,000	4,912	7,175

Preliminary Data Filtering. High-resolution images scraped from the web often suffer from blur, noise, or lack of texture, which can significantly degrade image quality. To eliminate such artifacts, we apply a two-stage low-level quality filter. First, we compute the Laplacian variance to assess image sharpness and discard samples below a blur threshold. Second, we apply the Sobel operator to measure edge density, removing overly flat or textureless images. This process yields a cleaned subset S with sufficient basic visual quality.

Detail Richness. Fine-grained details are essential for training generative models to preserve high-frequency content. To quantify the aspect, we compute Gray-Level Co-occurrence Matrix (GLCM) score, including contrast, entropy, and correlation across multiple directions. These metrics capture spatial pixel relationships indicative of texture complexity. We then select the top 50% of images from S with the highest aggregated GLCM scores, resulting in subset S_G.

Content Complexity. Visually complex images and diverse spatial structures are more valuable for guiding generation models to achieve rich content. We use Shannon entropy as a proxy to measure the content complexity of each image. Images with higher entropy tend to contain more varied pixel intensities. From subset S, we retain the top 50% highest-entropy images to construct subset S_E.

Figure 4: Comparison between our UltraHR-100K and Aesthetic-4K[16]. Captions in our UltraHR-100K provide more expressive descriptions, encompassing not only global summaries of the image content but also rich details that enhance semantic alignment.

Aesthetic Quality. Aesthetic appeal is an important factor in image realism and human preference. To incorporate this dimension, we adopt the LAION Aesthetic Predictor [57], a neural network trained to estimate perceptual quality. It outputs a scalar score reflecting visual composition, color harmony, and overall appeal. We rank all images in S by their aesthetic scores and retain the top 50% to form subset S_A, consisting of the most visually pleasing samples.

UltraHR-100K. To ensure that the final dataset consists of high-quality UHR images with *diverse content, rich textures, and strong aesthetic appeal*, we take the intersection of the three selected subsets. Specifically, the final dataset is defined as:

$$\text{UltraHR-100K} = S_G \cap S_E \cap S_A \tag{1}$$

This intersection guarantees that each image in UltraHR-100K simultaneously meets high standards in detail richness, content complexity, and aesthetic quality, as shown in Figure 2. In addition, we construct a evaluation subset from our dataset—**UltraHR-eval4K**—containing 2,000 images. Table 2 compares our UltraHR-100K with Aesthetic-4K [16] and PixArt-30K [17]. These statistics highlight that UltraHR-100K not only improves dataset scale, but also provides more extensive spatial content.

UHR Image Caption. UHR images typically contain significantly more visual information than standard-resolution images, making them inherently more semantically complex. However, existing datasets [16, 57] often provide only short captions, limiting the semantic expressiveness of generative models. To address this issue, we leverage Gemini 2.0 [28], a state-of-the-art commercial vision-language model (VLM), to generate rich and detailed captions for our dataset. As illustrated in Figures 3 and 4, our captions are not only substantially longer but also encompass both global summaries and fine-grained descriptions, enhancing alignment with complex image content.

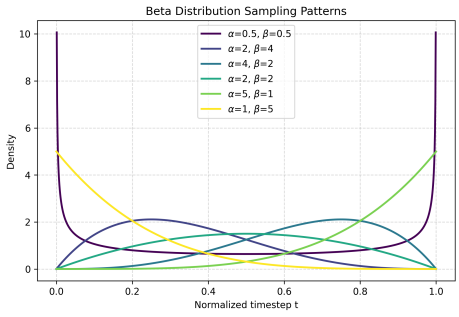

Figure 5: Relation between weighting ratio and timesteps with beta sampling strategy.

4 Frequency-Aware Post-Training

Pretrained T2I models, trained on large-scale datasets, exhibit strong capabilities in semantic and content planning. However, they often struggle to synthesize fine-grained details when extended to UHR scenarios [12, 15]. In this work, we focus on enhancing the detail synthesis ability of pretrained T2I models through tailored post-training strategies. To this end, we propose a frequency-aware post-training method (FAPT). Specifically, FAPT consists of two parts: detail-oriented timestep sampling (DOTS) and soft-weighting frequency regularization (SWFR). DOTS improves detail synthesis in UHR image generation by directing more training focus to timesteps associated with fine-grained details. Meanwhile, SWFR imposes a soft-weighted constraint across the frequency spectrum, guiding the model to better preserve and reconstruct high-frequency details.

4.1 Detail-Oriented Timestep Sampling

Motivation. Existing study [58] have validated the observation that the overall image structure (low-frequency signals) is largely reconstructed in the early denoising steps, while fine-grained details (high-frequency signals) are progressively synthesized in the later stages of the denoising process. This insight motivates us to design a sampling strategy that emphasizes the later stages of the denoising process, aiming to enhance the learning of fine-grained details during post-training stage.

DOTS. To achieve this target, we adopt a beta sampling strategy, which provides a simple yet flexible mechanism to bias the sampling distribution over denoising timesteps, as shown in Figure 5. Specifically, we first draw a timestep $t \in (0, 1)$ from a Beta distribution parameterized by shape parameters α and β:

$$t \sim \mathrm{Beta}(\alpha, \beta). \tag{2}$$

The Beta distribution yields a rich family of unimodal or skewed distributions over the interval $(0, 1)$, and its probability density function is given by:

$$\pi_{\mathrm{beta}}(t; \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} t^{\alpha-1}(1-t)^{\beta-1}, \tag{3}$$

where $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function. By adjusting α and β, we can control the bias of the sampling distribution. This sampling mechanism naturally supports adaptive emphasis in training: by emphasizing later denoising timesteps, we can guide the model to focus on high-frequency details.

4.2 Soft-Weighting Frequency Regularization

Motivation. Large pre-trained T2I models [2, 3, 5, 7] demonstrate strong semantic planning from diverse data exposure but struggle with fine-grained detail synthesis in UHR scenarios. Existing UHR T2I models focus mainly on training efficiency [15, 17], often neglecting high-fidelity detail. Diffusion4K [16] introduces DWT-based frequency decomposition to enable 4K training, yet DWT yields coarse and discontinuous frequency separation, limiting its effectiveness for UHR modeling. To overcome this, we adopt DFT-based decomposition, which provides finer, globally coherent frequency representations better suited for capturing fine-scale structures in high-resolution synthesis.

SWFR. To enhance fine-scale fidelity in UHR image synthesis, we introduce a soft-weighting frequency regularization that complements the standard diffusion loss by explicitly supervising frequency consistency, with an emphasis on high-frequency components. Formally, consider the standard diffusion process:

$$\boldsymbol{z}_t = \alpha_t \cdot \boldsymbol{x}_0 + \sigma_t \cdot \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{x}_0$ denotes the data distribution, $\boldsymbol{\epsilon}$ is sampled from standard normal distribution, and α_t, σ_t are known coefficients in the diffusion formulation. Recent T2I models [5, 7, 15] adopt rectified flows to predict velocity $\boldsymbol{v}$, with the objective as follows:

$$\boldsymbol{v}_\Theta(\boldsymbol{z}_t, t) = \boldsymbol{\epsilon} - \boldsymbol{x}_0. \tag{5}$$

To regularize the model in the frequency domain, we compute the 2D Discrete Fourier Transforms (DFT) of both prediction $\boldsymbol{x}$ and target $\boldsymbol{y}$:

$$\hat{\boldsymbol{x}} = \mathcal{F}(\boldsymbol{x}), \quad \hat{\boldsymbol{y}} = \mathcal{F}(\boldsymbol{y}), \tag{6}$$

Figure 6: Qualitative comparisons with SOTA methods on our UltraHR-eval4K (4096× 4096). Compared with previous works, our method is capable of generating visually complex images with rich semantic content. More visual examples are available in the supplementary materials.

where $\mathcal{F}(\cdot)$ denotes the DFT. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote the model prediction (e.g., $\boldsymbol{x} = \boldsymbol{v}_\Theta(\boldsymbol{z}_t, t)$) and target (e.g., $\boldsymbol{y} = \boldsymbol{\epsilon} - \boldsymbol{x}_0$), respectively. We define a frequency-domain regularization term as:

$$\mathcal{L}_{\text{freq}} = \mathbb{E}\left[|w(\boldsymbol{r}) \cdot \hat{\boldsymbol{x}} - w(\boldsymbol{r}) \cdot \hat{\boldsymbol{y}}|^2\right], \tag{7}$$

where $w(\boldsymbol{r})$ is a frequency soft weighting function designed to boost high-frequency supervision:

$$w(\boldsymbol{r}) = 1 + \lambda \cdot \frac{\exp(\gamma \boldsymbol{r}) - 1}{\exp(\gamma) - 1}, \quad \boldsymbol{r} \in [0, 1], \tag{8}$$

and $\boldsymbol{r}$ is the normalized distance from the center of the frequency plane. Hyperparameters λ and γ control the strength and steepness of high-frequency emphasis, respectively. Finally, the overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{freq}} \cdot \mathcal{L}_{\text{freq}}, \tag{9}$$

where $\mathcal{L}_{\text{diff}}$ denotes the diffusion loss, which can be instantiated as velocity prediction loss ($\|\boldsymbol{v}_\Theta(\boldsymbol{z}_t, t) - (\boldsymbol{\epsilon} - \boldsymbol{x}_0)\|^2$). λ_{freq} is a balancing coefficient that controls the strength of frequency-domain supervision. This regularization $\mathcal{L}_{\text{freq}}$ encourages the model to maintain consistent spectral power between prediction and target, especially in high-frequency bands.

5 Experiments

5.1 Implementation Details

Overall Training Setting. We adopt a two-stage training strategy. In the first stage, we follow the Logit-Normal Sampling scheme introduced in SD3 [6] and perform fine-tuning on our UltraHR100K dataset, aiming to enhance the semantic planning capability in UHR generation. In the second stage, we apply our proposed frequency-aware post-training method, which focuses on high-frequency learning to further improve the fine-grained details. We use the CAMEWrapper [15] optimizer with a constant learning rate of 1e-4, and employ mixed-precision training with a batch size of 24. The first-stage training is conducted for 4K iterations, followed by 8K iterations in the second stage. Due to computational constraints, we conduct training solely on SANA, and all experiments are performed on four H20 GPUs.

Baselines. To comprehensively evaluate our approach, we conduct extensive comparisons against SOTA methods for UHR image generation, which can be broadly categorized into three groups. The first group consists of powerful T2I models combined with super-resolution technique, BSRGAN [59]. The second group includes training-free approaches, where we evaluate FLUX[7]) using corresponding training-free generation methods, I-Max [60] and HiFlow [13]. Lastly, we compare with leading

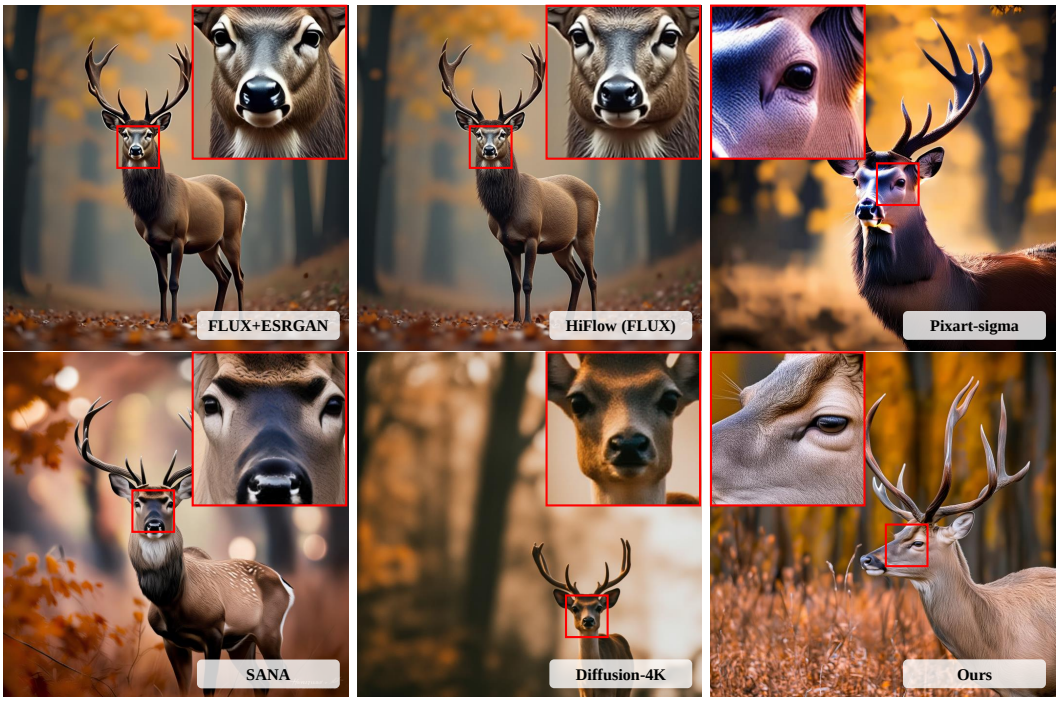

Figure 7: Qualitative comparisons with SOTA methods on our UltraHR-eval4K (4096× 4096). Compared with previous works, our method can generate realistic textures and fine-grained details.

Figure 8: More visual comparisons demonstrate that our method consistently produces high-quality results. Additional and more diverse comparisons can be found in the supplementary material.

training-based UHR generation models, including Pixart-σ [17], SANA [15], and Diffusion4K [16]. All baselines are evaluated under their official settings to ensure a fair and consistent comparison.

Evaluation. We employ several metrics to assess the quality of the generated images, with a particular focus on our evaluation sets, UltraHR-eval4K. To evaluate image-text consistency, we calculate the long CLIP score [61] and Fine-Grained (FG) CLIP score [62]. Additionally, the Fréchet Inception Distance (FID) [63] and Inception Score (IS) [64] are computed to evaluate the overall image quality of the generated images. Following previous works [13, 12], we compute the FID-patch and IS-patch to evaluate the local quality and details of the images, which are based on local image patches. These metrics provide a comprehensive evaluation of the overall quality, detail retention in the generated images.

Table 3: **Quantitative comparison with other baselines on our UltraHR-eval4K (4096 × 4096) benchmark.** The best result is highlighted in **bold**.

Method	FID ↓	$\text{FID}_{\text{patch}}$ ↓	IS ↑	IS_{patch} ↑	CLIP ↑	FG-CLIP ↑
FLUX [7] + BSRGAN [59]	37.651	43.143	11.773	5.389	31.45	28.02
SD3.5 [6] + BSRGAN [59]	31.870	25.598	12.780	5.456	31.75	28.66
I-Max(FLUX) [60]	37.667	37.835	11.991	4.391	31.49	27.78
HiFlow(FLUX) [13]	35.892	38.327	11.767	4.620	31.52	27.75
Pixart-σ [17]	33.171	32.198	12.212	5.390	31.78	28.65
SANA [15]	37.070	38.795	11.778	**5.649**	31.70	28.60
Diffusion4K [16]	39.857	38.515	10.832	3.235	31.41	26.48
Ours(UltraHR-100K)	33.995	20.932	12.502	5.020	**31.85**	28.65
Ours(UltraHR-100K+FAPT)	**31.748**	**15.795**	**12.995**	5.104	31.82	**28.68**

Table 4: Left: User study results conducted on our UltraHR-eval4K. Right: Quantitative comparison on Aesthetic-Eval@4096. The results demonstrate the superior performance of our method.

Method	Overall Quality	Detail Quality	Text-Image Alignment	Preference
Pixart-σ [17]	14%	10%	16%	18%
SANA [15]	4%	8%	8%	6%
Diffusion4K [16]	12%	4%	6%	6%
Ours	**70%**	**78%**	**72%**	**70%**

Method	FID ↓	$\text{FID}_{\text{patch}}$ ↓	CLIP ↑	FG-CLIP ↑
Pixart-σ [17]	150.593	44.702	34.88	28.48
SANA [15]	146.027	37.031	34.62	28.61
Diffusion4K [16]	152.790	39.729	33.99	26.06
Ours	**142.965**	**24.008**	**35.08**	**28.64**

5.2 Comparison to State-of-the-Art Methods

Quantitative Comparison. Table 3 summarizes the quantitative performance on our UltraHR-eval4K benchmark (4096 × 4096). Our method consistently achieves superior scores on key perceptual metrics such as FID, FID-patch and IS, indicating its strong capability in generating high-quality images with fine-grained textures. Moreover, our method achieves competitive CLIP scores, reflecting its ability to maintain semantic alignment with the input prompt. Notably, our method yields a substantial improvement in $\text{FID}_{\text{patch}}$, highlighting its effectiveness in synthesizing fine-grained details. This result demonstrates that our proposed approach significantly enhances the detail generation capability of pre-trained T2I models in UHR scenarios.

Qualitative Comparison. Figure 6 presents qualitative comparisons on UltraHR-eval4K (4096 × 4096), focusing on the overall semantic richness and spatial layout of the generated images. While existing SOTA methods struggle to produce coherent and content-rich scenes at such ultra-high resolution, our method demonstrates a strong capability in generating visually complex images with diverse and semantically meaningful elements. This highlights our model's superior capacity for global spatial reasoning and semantic planning in large-scale synthesis. In Figure 7, we further compare fine-grained textures and local details. Our method produces sharper structures and more realistic textures, faithfully preserving high-frequency information that other methods tend to miss or oversmooth. These results collectively demonstrate the effectiveness of our proposes dataset and method in enhancing both the global semantics and local fidelity for ultra-high-resolution text-to-image generation. Figure 8 presents more visual comparisons.

User Study. As shown in Table 4, we conducted a user study with 5 volunteers evaluating 50 randomly selected cases. Images were rated on overall quality, detail quality, text-image alignment and preference. The results demonstrate the superiority of our method across all aspects.

Comparisons on Public Benchmark. We conduct a quantitative comparison on the publicly available Aesthetic-4K benchmark, specifically the Aesthetic-Eval@4096 subset, as reported in Table 4. This evaluation set contains 195 image-text pairs, where all images have a short side greater than 4096 pixels. Due to the limited number of samples, the reported FID scores are relatively high. Nonetheless, the results clearly demonstrate the superior performance of our method, supporting its robustness and generalizability beyond our proposed benchmark.

Table 5: Ablation study of our key components and data scale. Model A is a baseline using full fine-tuning on our dataset. The comparison between C (trained on a partial dataset) and D (full dataset) validates the effectiveness of large-scale data.

Model	DOTS	SWFR	Dataset	**FID** ↓	**FID$_{\text{patch}}$** ↓	**CLIP** ↑
LoRA	×	×	Full	35.07	35.02	31.80
A	×	×	Full	33.99	20.93	**31.85**
B	✓	×	Full	32.57	19.95	31.79
C	✓	✓	Part	32.75	18.42	31.81
D	✓	✓	Full	**31.74**	**15.79**	31.82

5.3 Ablation Study

We conduct a comprehensive ablation study to validate the effectiveness of our proposed training strategy and the importance of large-scale data. As shown in Table 5, Model A serves as the baseline without our proposed DOTS and SWFR. Model B introduces DOTS, resulting in consistent improvements in both FID and patch-level FID, demonstrating its effectiveness in guiding the sampling process. Further incorporating SWFR (Model D) yields substantial improvements, particularly in patch-level FID, confirming that our proposed regularization enhances the detail synthesis capability of T2I models. To evaluate the impact of training data scale, we compare Model C and Model D. Model C is trained with a randomly sampled 15K subset of our UltraHR-100K using the same training strategy. The performance drop compared to Model D clearly highlights the importance of large-scale UHR data in achieving high-fidelity and semantically aligned image generation.

Analysis for DOTS. The DOTS module employs a Beta(α, β) distribution to guide timestep sampling, where α and β control the bias along the denoising trajectory. When $\alpha < \beta$, sampling favors later steps (near $t = 0$) that refine high-frequency details; when $\alpha > \beta$, it leans toward early steps (near $t = 1$) emphasizing global structure. In our experiments, we set $\alpha = 2$, $\beta = 4$, biasing sampling toward later steps to better capture fine details crucial for ultra-high-resolution generation. An ablation study (Table 6) varying α and β confirms this choice: larger α weakens detail learning, smaller α harms semantic consistency, and overly concentrated or flattened distributions reduce diversity. These results validate $(\alpha = 2, \beta = 4)$ as a balanced and effective configuration.

Table 6: Analysis for DOTS.

Method	**FID**	**FID_patch**	**CLIP**
$(\alpha = 3, \beta = 4)$	33.196	22.143	31.83
$(\alpha = 1, \beta = 4)$	33.727	25.095	31.79
$(\alpha = 2, \beta = 5)$	33.874	23.850	31.82
$(\alpha = 2, \beta = 3)$	33.638	24.638	**31.84**
$(\alpha = 2, \beta = 4)$	**31.748**	**15.795**	31.82

6 Conclusion

In this paper, we present UltraHR-100K, a curated dataset of 100K UHR images with rich textual annotations. Each image is carefully selected to ensure high levels of detail, visual complexity, and aesthetic appeal. Moreover, we introduce a frequency-aware post-training method, which includes: (i) Detail-Oriented Timestep Sampling (DOTS), and (ii) Soft-Weighting Frequency Regularization (SWFR). Experiments on our proposed UltraHR-eval4K benchmark confirm that our approach significantly boosts both the visual fidelity and fine-detail accuracy of UHR image synthesis.

Limitations and future works. Our main limitations lie in two aspects. First, while the proposed frequency-aware post-training strategy effectively enhances fine-detail synthesis, it introduces a slight degradation in text–image alignment, as shown in Table 5. Second, our dataset currently contains a relatively limited amount of portrait data, which constrains the improvement in ultra-high-resolution (UHR) portrait generation, as illustrated in Figure 8. In future work, we plan to develop more balanced training strategies to alleviate the alignment issue and expand our dataset with additional high-quality UHR portrait images to further improve performance in portrait synthesis.

Acknowledgments. This work was supported by Natural Science Foundation of China: No. 62406135, Natural Science Foundation of Jiangsu Province: BK20241198, and Gusu Innovation and Entrepreneur Leading Talents: No. ZXL2024362.

References

[1] 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] 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.

[3] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*.

[4] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

[5] 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.

[6] 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.

[7] Black-Forest Labs. Flux. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024.

[8] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, 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.

[9] Leyang Li, Shilin Lu, Yan Ren, and Adams Wai-Kin Kong. Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. *arXiv preprint arXiv:2504.12782*, 2025.

[10] Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025.

[11] Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3581–3589, 2025.

[12] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[13] Jiazi Bu, Pengyang Ling, Yujie Zhou, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Hiflow: Training-free high-resolution image generation with flow-aligned guidance. *arXiv preprint arXiv:2504.06232*, 2025.

[14] 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.

[15] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

[16] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[17] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[18] Chen Zhao, Zhizhou Chen, Yunzhe Xu, Enxuan Gu, Jian Li, Zili Yi, Qian Wang, Jian Yang, and Ying Tai. From zero to detail: Deconstructing ultra-high-definition image restoration from progressive spectral perspective. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17935–17946, 2025.

[19] Chen Zhao, Weiling Cai, Chengwei Hu, and Zheng Yuan. Cycle contrastive adversarial learning with structural consistency for unsupervised high-quality image deraining transformer. *Neural Networks*, 178:106428, 2024.

[20] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[21] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European Conference on Computer Vision*, pages 196–212. Springer, 2024.

[22] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36:70847–70860, 2023.

[23] Younghyun Kim, Geunmin Hwang, Junyu Zhang, and Eunbyung Park. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 4338–4346, 2025.

[24] Zhengqiang Zhang, Ruihuang Li, and Lei Zhang. Frecas: Efficient higher-resolution image generation via frequency-aware cascaded sampling. *The Thirteenth International Conference on Learning Representations*, 2025.

[25] Zihan Zhou, Shilin Lu, Shuli Leng, Shaocong Zhang, Zhuming Lian, Xinlei Yu, and Adams Wai-Kin Kong. Dragflow: Unleashing dit priors with region based supervision for drag editing. *arXiv preprint arXiv:2510.02253*, 2025.

[26] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. *arXiv preprint arXiv:2411.06558*, 2024.

[27] Nikai Du, Zhennan Chen, Shan Gao, Zhizhou Chen, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025.

[28] Google. Gemini. `https://gemini.google.com/`, 2025.

[29] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024.

[30] Li Zhang, Yan Zhong, Jianan Wang, Zhe Min, Liu Liu, et al. Rethinking 3d convolution in ℓ_p-norm space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[31] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc++: Advanced multi-instance generation controller for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[32] Li Zhang, Zean Han, Yan Zhong, Qiaojun Yu, Xingyu Wu, et al. Vocapter: Voting-based pose tracking for category-level articulated object via inter-frame priors. In *ACM Multimedia 2024*, 2024.

[33] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023.

[34] Ji Du, Jiesheng Wu, Desheng Kong, Weiyun Liang, Fangwei Hao, Jing Xu, Bin Wang, Guiling Wang, and Ping Li. Upgen: Unleashing potential of foundation models for training-free camouflage detection via generative models. *IEEE Transactions on Image Processing*, 2025.

[35] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

[36] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2024.

[37] 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.

[38] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.

[39] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. *arXiv preprint arXiv:2410.12669*, 2024.

[40] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8281–8291, 2024.

[41] Chen Zhao, Chenyu Dong, and Weiling Cai. Learning a physical-aware diffusion model based on transformer for underwater image enhancement. *arXiv preprint arXiv:2403.01497*, 2024.

[42] Chen Zhao, Weiling Cai, Chenyu Dong, and Ziqi Zeng. Toward sufficient spatial-frequency interaction for gradient-aware underwater image enhancement. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3220–3224. IEEE, 2024.

[43] Zhennan Chen, Rongrong Gao, Tian-Zhu Xiang, and Fan Lin. Diffusion model for camouflaged object detection. *arXiv preprint arXiv:2308.00303*, 2023.

[44] Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. *arXiv preprint arXiv:2501.02976*, 2025.

[45] Ji Du, Fangwei Hao, Mingyang Yu, Desheng Kong, Jiesheng Wu, Bin Wang, Jing Xu, and Ping Li. Shift the lens: Environment-aware unsupervised camouflaged object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19271–19282, 2025.

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

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[48] 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.

[49] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024.

[50] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. *arXiv preprint arXiv:2305.10028*, 2023.

[51] 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.

[52] Li Zhang, Weiqing Meng, Yan Zhong, Bin Kong, Mingliang Xu, Jianming Du, Xue Wang, Rujing Wang, and Liu Liu. U-cope: Taking a further step to universal 9d category-level object pose estimation. In *European Conference on Computer Vision*, pages 254–270. Springer, 2025.

[53] Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. *arXiv preprint arXiv:2503.12885*, 2025.

[54] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*, 2024.

[55] Li Zhang, Mingliang Xu, Dong Li, Jianming Du, and Rujing Wang. Catmullrom splines-based regression for image forgery localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7196–7204, 2024.

[56] Shilin Lu, Zhuming Lian, Zihan Zhou, Shaocong Zhang, Chen Zhao, and Adams Wai-Kin Kong. Does flux already know how to perform physically plausible image composition? *arXiv preprint arXiv:2509.21278*, 2025.

[57] 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.

[58] Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[59] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4791–4800, 2021.

[60] Ruoyi Du, Dongyang Liu, Le Zhuo, Qin Qi, Hongsheng Li, Zhanyu Ma, and Peng Gao. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. *arXiv preprint arXiv:2410.07536*, 2024.

[61] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024.

[62] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.

[63] 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.

[64] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We list our contributions and scope in the abstract and in the last three paragraphs of Section introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Sec. Experiments.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: In this paper, we do not present new theoretical results, as we rely on existing theories.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the detailed network components, training procedure and training settings at Sec. Experiments, respectively.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code in supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide these details at Sec. Experiments and Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars or other information about the statistical significance of the experiments. We will consider including this information in future work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the type of compute workers and memory in Sec. Experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read the NeurIPS Code of Ethics and find that the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive societal impacts in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The new assets introduced in the paper are not well documented. We will improve the documentation and provide it alongside the assets in future revisions.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy for what should or should not be described.