RealUHR: Harnessing Patch-Cascade Flows for Photorealistic Ultra-High-Resolution Synthesis

Yongsheng Yu[1], Haitian Zheng[2], Zhe Lin[2], Connelly Barnes[2], Yuqian Zhou[2], Zhifei Zhang[2], Jiebo Luo[1]

[1]University of Rochester
[2]Adobe Research

Abstract

Ultra-high-resolution (UHR) text-to-image synthesis faces significant hurdles, including immense computational costs and a scarcity of training data. To address these, we introduce **RealUHR**, an efficient and scalable framework for generating photorealistic 4K images. At its core, RealUHR employs a *Patch-Cascade Flow Matching* pipeline that ensures global coherence without costly patch fusion by initiating generation from a semantically meaningful structure. This enables highly efficient, few-step inference for independent patches. Our key contribution is *Guidance-Consistent Adaptation (GCA)*, a novel two-stage strategy to resolve the fundamental objective mismatch in guidance-distilled models. GCA allows powerful backbones like FLUX to be effectively adapted for patch-aware UHR synthesis. The framework's detail-rendering capabilities are further enhanced by a non-uniform time schedule. Experiments show that RealUHR establishes superior performance in both quality and efficiency, and excels in zero-shot applications such as creative up-sampling and generative artifact suppression.

Introduction

Deep generative models have made impressive strides in synthesizing high-fidelity and diverse visual content (Pernias et al. 2023; Podell et al. 2023; Labs 2024; Rombach et al. 2022; Chen et al. 2023; Yu et al. 2022; Zheng et al. 2025). These advances are largely driven by powerful backbones such as state-of-the-art Diffusion Transformers (Esser et al. 2024; Chen et al. 2023; Labs 2024; Peebles and Xie 2023), which excel at translating text prompts into high-quality images. This progress has sparked growing interest in Ultra-High-Resolution (UHR) image generation—often targeting resolutions up to 4K or beyond—for applications in digital art, virtual reality, scientific visualization, and high-fidelity media production (Liu et al. 2025b,a; Yu et al. 2024, 2025a).

A direct approach to UHR synthesis involves fine-tuning diffusion models on native high-resolution images. Training on complete high-resolution inputs facilitates the learning of holistic semantic structures, thereby yielding globally coherent outputs. However, this strategy typically requires extensive collections of curated, high-quality 4K data and substantial computational resources. Recent 4K Text-to-Image (T2I) methods (Chen et al. 2023; Xie et al. 2024; Ren et al. 2024) depend on millions of images and incorporate specialized lightweight architectures such as linear attention (Xie et al. 2024), compact parameter counts (Chen et al. 2023), or aggressive feature compression (Ren et al. 2024). Such adaptations usually hinder its accessibility and complicate the adoption of powerful pretrained foundation models, such as the 12B-parameter FLUX (Labs 2024), for UHR generation. Although Diffusion-4K (Zhang et al. 2025) alleviates memory constraints by fine-tuning FLUX with a more aggressive VAE compression ratio, this latent compression inevitably sacrifices high-frequency detail and textural realism, resulting in localized blurriness and reduced detail under close inspection (see Figure 4).

An alternative line of work explores training-free strategies (Du et al. 2024a; Zhang et al. 2023; Huang et al. 2024; Du et al. 2024b), which harness the generation capacity of pretrained T2I models through progressive upscaling. For

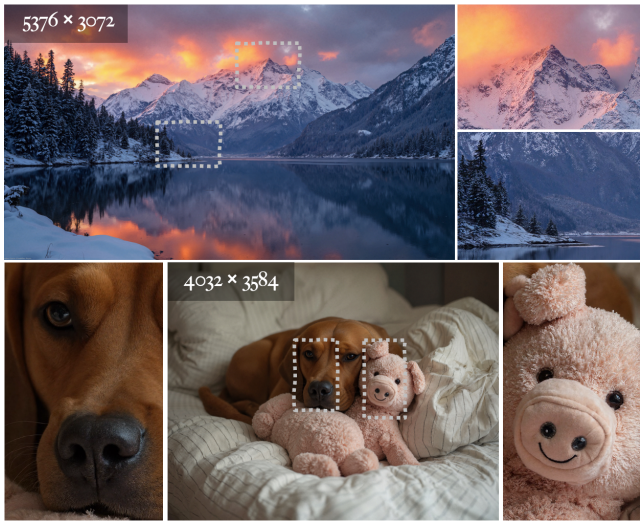

Figure 1: RealUHR: Pushing the Frontiers of Text-to-Image Synthesis. Our framework generates ultra-high-resolution images with exceptional photorealism and fine-grained detail. The examples above, synthesized at up to 4K, show the ability to render coherent scenes with intricate textures suitable for the most demanding applications.

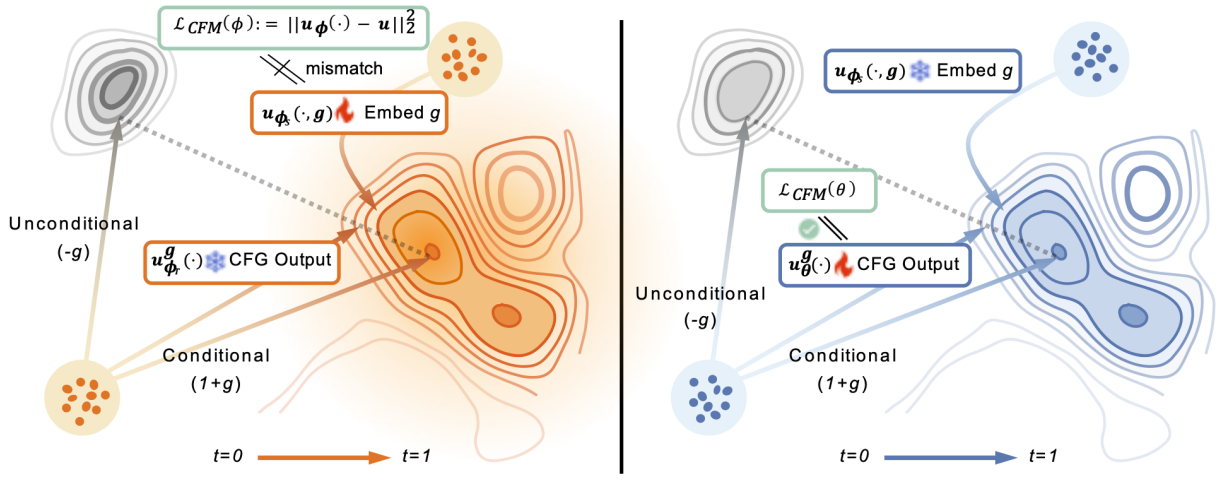

Figure 2: The Guidance-Consistent Adaptation (GCA) motivation and strategy. Left: The Flow Matching objective ($\mathcal{L}_{\text{CFM}}$) is incompatible with the guidance-aware student model ϕ_S, creating an objective mismatch during finetuning. Right: Our reverse distillation decouples the baked-in guidance by training a new model θ to explicitly output conditional and unconditional velocity fields. The resulting model θ becomes fully compatible with the $\mathcal{L}_{\text{CFM}}$ objective for effective patch-aware adaptation.

instance, DemoFusion (Du et al. 2024a) decomposes high-resolution images into inference-manageable patches and subsequently refining them in an iterative manner (Bar-Tal et al. 2023; Lee et al. 2023). The patch-based workflow supports scaling to arbitrary resolutions and adapts to diverse domains. However, to address the inconsistencies that arise from processing local patches independently, these methods require computationally expensive local patch fusion to enforce coherence, thereby sacrificing runtime efficiency.

Motivated by Flow Matching (Lipman et al. 2022), which models the generative process via straight-line trajectories in latent space, an alternative paradigm emerges that is inherently simpler, more stable, and highly efficient (Liu, Gong, and Liu 2022). The linearity of these paths allows for bypassing the initial denoising stages by commencing generation directly from a semi-noisy latent structure. To this end, we propose a Patch-cascade Flow Matching pipeline to synthesize high-resolution outputs through independent patch-wise diffusion, initiated from a semi-noisy structure. Unlike conventional methods that rely on costly local patch fusion during inference, our approach begins from a semantically meaningful structure rather than pure noise. This practically ensures global coherence from the outset. Furthermore, because the underlying ODE trajectories in Rectified Flows (Esser et al. 2024) are significantly straighter and simpler than those of traditional diffusion models, numerical solvers can employ much larger step sizes. This facilitates high-quality generation only in 4 steps, thereby enhancing runtime efficiency while preserving high fidelity.

Furthermore, to adapt powerful pretrained diffusion backbones (e.g., the 12B-parameter FLUX (Labs 2024)) to patch-aware inference, we propose a two-stage, guidance-consistent patch adaptation strategy as shown in Figure 2. Initially, we eliminate dependence on the classifier-free guidance scale embedding inherent in the original distilled model. Subsequently, we finetune the model on a carefully curated hybrid dataset consisting of approximately 80K rigorously filtered high-quality patches and 70K full-resolution images. This staged adaptation ensures a seamless transition between the pretrained CFG-distilled model and the fine-tuned patch-adapted version, preserving consistency in the optimization objectives.

The proposed RealUHR demonstrates robust scalability, supporting arbitrary resolutions up to 4K as shown in Figure 1, and exhibits excellent generalization capabilities in zero-shot scenarios. It can function as a creativity-controlled super-resolution module, allowing users to flexibly balance image fidelity and creative variation, and also serves as an effective post-processing filter for refining generative artifacts in existing T2I outputs, all without additional supervision or retraining.

In summary, our main contributions are:

- We introduce **RealUHR**, a highly efficient patch-cascade framework for UHR synthesis that achieves global coherence without costly patch fusion techniques.
- We propose *Guidance-Consistent Adaptation (GCA)*, a two-stage fine-tuning strategy. GCA resolves the objective mismatch in guidance-distilled models.
- Our method achieves strong performance on 2K/4K generation benchmarks, and we further demonstrate its practical utility in zero-shot applications such as creative upsampling and artifact suppression.

Related Work

Flow–based Diffusion Models

Text-to-image (T2I) diffusion models (Rombach et al. 2022; Podell et al. 2023; Pernias et al. 2023; Chen et al. 2024) have achieved remarkable progress in generating high-fidelity and semantically aligned images from text prompts. Recent developments in *Flow Matching* (FM) (Lipman et al.

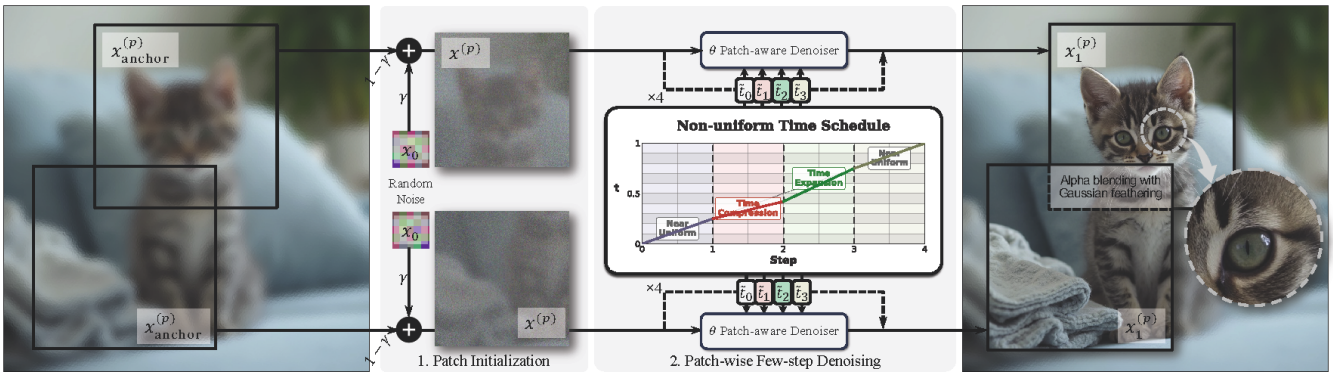

Figure 3: The Patch-Cascade Flow Matching pipeline. (1) Patch Initialization: Anchor patches are interpolated to generate semantically meaningful latent patches. (2) Patch-wise Few-Step Denoising: Each patch is independently refined by a denoiser guided by our Non-uniform Time Schedule. Finally, all patches are composed via alpha blending with Gaussian feathering.

2022) have introduced a deterministic perspective to diffusion by reformulating the process as an ordinary differential equation (ODE), enabling improved numerical stability and compatibility with advanced integrators (Labs 2024; Esser et al. 2024). While both diffusion and FM approaches require multiple evaluations of a large network through discretized ODEs or SDEs, researchers have proposed various scheduling strategies to better allocate computational resources across time steps. These include shortened trajectories (Song, Meng, and Ermon 2021), high-order solvers (Lu et al. 2022), and optimal-transport-inspired paths (Tong et al. 2024), all aimed at accelerating generation without compromising output quality. We apply FM-based denoising samplers and fine-tuning objectives to the pre-trained text-to-image model FLUX to achieve photorealistic ultra-high-resolution image generation.

Ultra-High-Resolution Image Generation

Scaling diffusion models to UHR image generation is fundamentally challenged by data scarcity and quadratic computational growth. One line of work confronts this by fine-tuning models on native UHR data (Ren et al. 2024; Chen et al. 2024; Xie et al. 2024). While effective, these methods typically require specialized, lightweight architectures to remain feasible, alongside massive datasets and GPU resources. A more resource-efficient alternative is to employ a two-stage pipeline, using a dedicated super-resolution model (Liang et al. 2021; Wang et al. 2021; Yu et al. 2025b; Kang et al. 2023) to upscale a base image. This approach, however, introduces a disconnect, as the SR model is often trained on synthetic degradations (e.g., blur, JPEG artifacts (Wang et al. 2021)) and operates with an objective distinct from the initial text-to-image generation.

A complementary paradigm avoids retraining entirely, proposing *training-free* solutions that adapt a single T2I model's inference process for progressive upscaling. Methods like DemoFusion (Du et al. 2024a) and I-Max (Du et al. 2024b) leverage techniques such as global-local denoising and latent-space flow modeling to operate on patches. Our work establishes a novel middle ground, bridging these two paradigms. We perform a highly efficient finetuning on a modest dataset, yet consolidate the entire generative process within a single T2I model, akin to training-free methods. This hybrid strategy allows us to retain the high fidelity of a specialized model while avoiding the architectural limitations and extensive resource demands of full UHR training.

Methodology

Preliminaries

Let $\mathbf{x}_0 \sim p(\mathbf{x})$ denote a source distribution (e.g., isotropic Gaussian noise) and $\mathbf{x}_1 \sim q(\mathbf{x})$ the target data distribution (e.g., natural images). FM defines an ordinary differential equation (ODE)-driven transformation $\boldsymbol{\psi}_t : \mathbb{R}^d \to \mathbb{R}^d$ governed by the learned velocity field $\mathbf{u}_t : \mathbb{R}^d \to \mathbb{R}^d$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\boldsymbol{\psi}_t(\mathbf{x}) = \mathbf{u}_t\left(\boldsymbol{\psi}_t(\mathbf{x})\right), \quad \boldsymbol{\psi}_0(\mathbf{x}) = \mathbf{x}. \tag{1}$$

By integrating Eq. (1) from $t = 0$ to $t = 1$, we obtain a trajectory $\mathbf{x}_t = \boldsymbol{\psi}_t(\mathbf{x}_0)$ that deterministically transforms noise $\mathbf{x}_0$ into data $\mathbf{x}_1 = \boldsymbol{\psi}_1(\mathbf{x}_0)$.

Conditional FM loss. Training minimizes

$$\mathbf{x}_t := (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \tag{2}$$

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\left\|\mathbf{u}_\theta(\mathbf{x}_t,t) - \mathbf{u}(\mathbf{x}_t \,|\, \mathbf{x}_1)\right\|_2^2\right], \tag{3}$$

with target field $\mathbf{u}(\mathbf{x}_t \,|\, \mathbf{x}_1) = \frac{\mathbf{x}_1 - \mathbf{x}_t}{1-t}$, which is the optimal linear transport between $\mathbf{x}_0$ and $\mathbf{x}_1$.

Uniform sampling. To sample from $p(\mathbf{x})$, Eq. (1) is discretized with an explicit Euler integrator (Euler 1768). Let $T = \{t_i\}_{i=0}^N$ be a uniform time grid with step size $\Delta t = 1/N$, then the approximate flow is computed:

$$\mathbf{x}_{t_{i+1}} = \mathbf{x}_{t_i} + \Delta t\, \mathbf{u}_\theta(\mathbf{x}_{t_i}, t_i), \quad i = 0, \ldots, N-1, \tag{4}$$

where $\mathbf{u}_\theta$ is the neural velocity field parameterized by θ. The final state $\hat{\mathbf{x}} := \mathbf{x}_{t_N}$ is returned as a sample from the model.

Patch-Cascade Flow Matching

An overview of the inference procedure for our Patch-Cascade Flow Matching framework is presented in Figure 3.

Algorithm 1 Non-Uniform Time Scheduling Inference

Input: text embedding $\mathbf{c}$, step count N, schedule parameters (α, p)
$\{t_i\}_{i=0}^{N-1} \leftarrow$ UNIFORMSCHEDULE(N)
for $i = 0$ to $N-1$ **do**
  $\tilde{t}_i \leftarrow t_i(1 + w_i)$ ..... Eq. (7)
  $\tilde{t}_i \leftarrow \min(\max(\tilde{t}_i, 0), 1)$
  $\mathbf{u} \leftarrow \mathbf{u}_\theta(\mathbf{x}_{t_i}, \tilde{t}_i, \mathbf{c})$
  $\mathbf{x}_{t_{i+1}} \leftarrow$ SCHEDULERSTEP$(\mathbf{x}_{t_i}, \mathbf{u}, t_i)$ ... Eq. (4)
end for
return $\mathbf{x}_{t_N}$

Patch Initialization. Unlike approaches (Bar-Tal et al. 2023; Du et al. 2024a) that rely on costly fusion of global and local denoising paths, we design patch initialization to maintain global semantic coherence across independently processed patches. Specifically, we initialize each patch trajectory with a latent formed by linearly interpolating between a Gaussian noise $\mathbf{x}_0$ and its *anchor patch* $\mathbf{x}^{(p)}_{\text{anchor}}$, where each anchor patch is an overlapping subregion of the base-resolution image $\mathbf{x}_{\text{base}}$ aligned to the model's training resolution (e.g., 1024×1024):

$$\mathbf{x}^{(p)} = (1-\gamma) \cdot \mathbf{x}^{(p)}_{\text{anchor}} + \gamma \cdot \mathbf{x}_0, \tag{5}$$

where $\gamma \in [0, 1]$ controls the level of noise corruption.

Non-Uniform Time Scheduling. We observe that the patch-wise inference of FLUX often leads to texture oversmoothing (see Figure 5). To address this, we leverage a key insight from diffusion model dynamics: the ODE solution trajectories are not perceptually uniform. As shown by Karras et al. (Karras et al. 2022), these trajectories exhibit significant curvature primarily in a narrow intermediate region, while remaining nearly linear at high and low noise levels. This suggests that reallocating computational effort to these critical, high-curvature periods can enhance textural realism. Therefore, we introduce a non-uniform time scheduling strategy:

$$w_i := \alpha \cdot \text{sgn}(t_i - 0.5) \cdot [\sin(\pi t_i)]^p, \tag{6}$$

$$\tilde{t}_i := t_i(1 + w_i), \tag{7}$$

where $\alpha \in (0, 0.5)$ controls the warping intensity and $p > 1$ sharpens the pulse by concentrating the reweighting toward the endpoints $t = 0$ and $t = 1$. Since the mapping $t_i \mapsto \tilde{t}_i$ is strictly monotonic, it preserves the chronological order of sampling steps, thereby maintaining compatibility with the flow-matching formulation. The full inference procedure is detailed in Algorithm 1. In summary, each latent patch $\mathbf{x}^{(p)}$ undergoes a 4-step, non-uniform FM sampling. The resulting denoised patches $\{\mathbf{x}^{(p)}_1\}$ are then blended via alpha blending with Gaussian feathering to yield the final image.

Guidance-Consistent Patch Adaptation

The FLUX.1-dev model, which we denote as ϕ_S, is a highly efficient student model. It is trained by distilling (Meng et al. 2023) the behavior of a larger, classifier-free guided (CFG) (Ho and Salimans 2022) teacher model, ϕ_T. This distillation process embeds the guidance scale, g, directly into the student model's architecture. The model ϕ_S learns to approximate the teacher's CFG-interpolated velocity field $\mathbf{u}^{\text{CFG}}_{\phi_T}$ by minimizing:

$$\mathcal{L}_{\text{distill}} = \left\| \mathbf{u}^{\text{CFG}}_{\phi_T} - \mathbf{u}_{\phi_S}(\mathbf{x}_t, t, \mathbf{c}, g) \right\|_2^2, \tag{8}$$

$$\text{where } \mathbf{u}^{\text{CFG}}_{\phi_T} = (1+g) \cdot \mathbf{u}_{\phi_T}(\mathbf{x}_t, t, \mathbf{c}) - g \cdot \mathbf{u}_{\phi_T}(\mathbf{x}_t, t). \tag{9}$$

Here, $\mathbf{u}_{\phi_S}(\mathbf{x}_t, t, \mathbf{c}, g)$ is the output of the student model, which explicitly takes the guidance scale g and condition $\mathbf{c}$ as inputs. This eliminates the need for separate forward passes, boosting sampling efficiency.

However, a challenge arises during finetuning. The standard flow matching objective (see Eq. (3)) trains a model to predict the data-conditioned velocity field $\mathbf{u}(\mathbf{x}_t \mid \mathbf{x}_1)$, which has no notion of the CFG structure. As illustrated in the left panel of Figure 2, this creates a fundamental mismatch between the distillation objective of ϕ_S and the standard finetuning objective, rendering direct finetuning ineffective.

To solve this, we propose a two-stage adaptation strategy:

Stage 1: Decoupling Guidance via Reverse Distillation. First, we initialize a new model, θ, with the weights of the pre-trained student model ϕ_S. We then finetune θ to disentangle the baked-in guidance via a "reverse distillation" process, where ϕ_S acts as a static teacher. We train θ by minimizing the following objective:

$$\mathcal{L}_{\text{reverse}} = \mathbb{E}\left[\left\| \mathbf{u}^{\text{CFG}}_{\theta} - \mathbf{u}_{\phi_S}(\mathbf{x}_t, t, \mathbf{c}, g) \right\|_2^2 \right], \tag{10}$$

$$\text{where } \mathbf{u}^{\text{CFG}}_{\theta} := (1+g) \cdot \mathbf{u}_{\theta}(\mathbf{x}_t, t, \mathbf{c}) - g \cdot \mathbf{u}_{\theta}(\mathbf{x}_t, t).$$

Here, the term $\mathbf{u}^{\text{CFG}}_{\theta}$ represents the application of classifier-free guidance to our new model θ, which produces separate conditional and unconditional outputs. The loss forces this CFG-combined output to match the single, guidance-aware output of the original FLUX model $\mathbf{u}_{\phi_S}$. This procedure, illustrated in the right panel of Figure 2, effectively makes θ independent of g as a direct input, preparing it for finetuning.

Stage 2: Patch Adaptation via Flow Matching Finetuning. With the guidance mechanism successfully decoupled, the model θ is now fully compatible with the standard flow matching objective. We can then proceed to finetune it for patch-aware adaptation using the loss from Eq. (3). The adaptation leverages a curated hybrid dataset comprising 80,000 high-quality image patches and approximately 70,000 full images drawn from LAION-Aesthetics (Schuhmann et al. 2022) and Aesthetic-4K (Zhang et al. 2025). Patch samples are rigorously filtered using a proposed pipeline that removes watermarked, overly dark, blurry, low-texture, and low-aesthetic-quality images, ensuring high visual fidelity and detail preservation.

Experiments

We validate RealUHR through comprehensive quantitative, qualitative, and user studies, benchmarking against existing high-resolution text-to-image methods. Ablation studies further analyze the impact of each module. Additionally, we show two zero-shot applications enabled by our framework.

Figure 4: Qualitative comparison against state-of-the-art models on 4K (4096×4096) synthesis. Our method, RealUHR, demonstrates superior photorealism and textural fidelity. Top row: RealUHR renders an elderly woman with striking facial realism and intricate fabric textures, details that are absent or stylized in competing results. Bottom row: Our model excels at complex natural scenes, capturing the fine fur of the panda and the structure of the bamboo.

Comparison with Existing Methods

Baselines. We compare our method against two categories of baselines: training-free approaches, such as I-Max (Du et al. 2024b), and DemoFusion (Du et al. 2024a); and models specifically trained for UHR generation, including Diffusion-4K (Zhang et al. 2025) and UltraPixel (Ren et al. 2024). Notably, Diffusion-4K, I-Max, and our proposed method share the same backbone, which is the multi-modal diffusion transformer-based model, FLUX (Labs 2024).

Quantitative Comparison. For benchmarking, we adopt the test dataset introduced in (Zhang et al. 2025), which contains 2,781 image-text pairs at 2K resolution and 195 pairs at 4K resolution. Our evaluation includes perceptual quality metrics such as FID (Heusel et al. 2017) and KID (Bińkowski et al. 2018), as well as a patch-based FID computed over multiple 299×299 crops to assess high-resolution fidelity. To evaluate text-image alignment, we employ CLIPScore (Hessel et al. 2021), while overall quality is further assessed using no-reference metrics CLIP-IQA (Wang, Chan, and Loy 2023) and QualiCLIP (Agnolucci, Galteri, and Bertini 2024).

Quantitative results are summarized in Table 1. Our method, built on the FLUX backbone, demonstrates superior performance. At 4K resolution, it achieves the best results across nearly all perceptual and quality metrics (FID, Patch FID, KID, CLIPIQA, and QualiCLIP) while also being the most efficient, with an inference time of just 108 seconds. In the 2K benchmark, our method maintains its leading position on perceptual metrics, securing the best FID, Patch FID, and KID scores. Overall, our approach establishes a new benchmark for balancing high-fidelity synthesis and computational efficiency in UHR scenarios.

Qualitative Comparison. Figure 4 provides a qualitative comparison of our method against several state-of-the-art baselines at 4K resolution. The results highlight RealUHR's superior capability in rendering photorealistic textures and fine-grained details, which are critical for convincing UHR synthesis. For instance, in the depiction of an "elderly woman" (top row), competing methods struggle with realism: DemoFusion fails to render a coherent face, UltraPixel produces a painterly effect, and I-Max also generates stylized results. In stark contrast, our method synthesizes a highly realistic portrait with crisp details in both the subject's face and the texture of her scarf. Similarly, when generating a "panda chewing bamboo" (bottom row), our model produces exceptionally fine fur textures and realistic eye details, avoiding the artificial smoothness or blurriness seen in the outputs of DemoFusion, UltraPixel, and Diffusion-4K. These examples underscore our model's effectiveness in generating complex scenes with high fidelity and detail.

User Study. To assess the effectiveness of our method in UHR image generation, we conducted a user study focusing on five evaluation criteria: *text alignment*, *photorealism*, *aesthetic*, *detail clarity*, and an aggregate metric, *Overall Quality*. Participants provided binary (*positive*) ratings for each criterion. Diverse textual prompts were used to ensure comprehensive coverage. For each method, we report the percentage of positive ratings per criterion. As summarized

	Method	Backbone	Perceptual metrics↓			Quality metrics↑		CLIPScore↑	Time(s)↓
			FID	Patch FID	KID	CLIPIQA	QualiCLIP		
4K	UltraPixel (Ren et al. 2024)	Cascade	143.20	46.50	5.56	0.508	0.376	**34.83**	151
	I-Max (Du et al. 2024b)	FLUX	141.25	50.99	5.83	0.433	0.303	33.96	751
	Diffusion-4K (Zhang et al. 2025)	FLUX	149.33	58.44	6.25	0.478	0.266	32.99	122
	Ours	FLUX	**139.29**	**46.15**	**3.93**	**0.636**	**0.437**	34.19	**108**
2K	DemoFusion (Du et al. 2024a)	SDXL	41.61	28.69	12.14	**0.615**	**0.596**	13.45	147
	UltraPixel (Ren et al. 2024)	Cascade	43.14	32.62	10.77	0.576	0.437	13.52	40
	I-Max (Du et al. 2024b)	FLUX	43.16	30.08	9.55	0.552	0.539	13.55	96
	Diffusion-4K (Zhang et al. 2025)	FLUX	42.38	30.58	9.62	0.581	0.434	**13.77**	**22**
	Ours	FLUX	**40.29**	**26.10**	**9.42**	0.613	0.495	13.54	38

Table 1: Quantitative comparison of UHR text-to-image models on 4K and 2K benchmarks. Best results are in bold, and second best are underlined. ↓ indicates lower is better, and ↑ indicates higher is better.

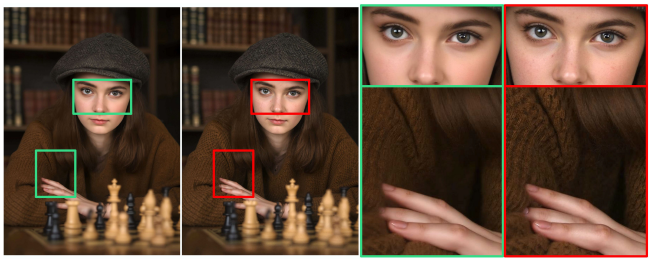

Figure 5: The impact of our Non-uniform Time Scheduling on textural realism. Compared to a standard uniform scheduler (Left), our method (Middle) significantly enhances fine-grained details and mitigates the texture oversmoothing effect.

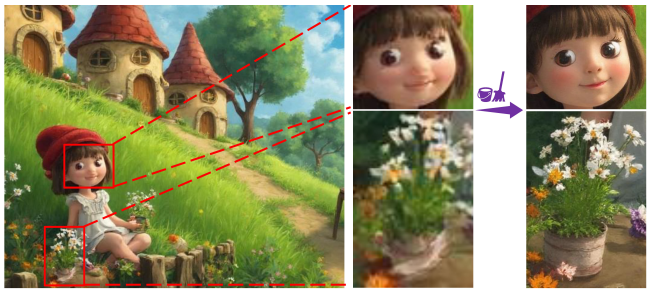

Figure 6: Application in generative artifact suppression. RealUHR can effectively correct common defects, such as distorted faces, in images generated by other models while preserving the original composition and semantic integrity.

in Figure 8, our approach consistently outperforms existing methods, achieving the highest positive rating percentages across all evaluation dimensions. Notably, our model excels in *photorealism*, indicating its strong capability in generating globally coherent semantics while yielding natural and realistic local textures.

Ablation Study

We conduct a series of ablation studies to validate the effectiveness of our key contributions, with detailed results presented in Table 2 and Table 3. First, we analyze our Non-uniform Time Scheduling against several alternatives. As shown in Table 2, our proposed scheduler consistently outperforms standard uniform, logarithmic, and cosine schedulers, confirming that strategically reallocating computation during inference is crucial for enhancing high-resolution detail. To determine the optimal settings, we also performed a systematic hyperparameter search for the scheduler's warping intensity α and pulse sharpness p, identifying $\alpha = 0.15$ and $p = 2.0$ as the ideal configuration.

In addition, we dissect the contributions of our framework's core components. As summarized in Table 3, removing the Patch Initialization (w/o Patch Init) and starting from pure noise leads to a severe degradation across all metrics, particularly the global FID score. This is due to a significant loss of semantic coherence between the independently generated patches. We also observe that removing the full GCA strategy results in a notable drop in performance, as the model is no longer effectively adapted for patch-wise synthesis, leading to degraded patch-specific details. Disabling only Stage 1 of GCA (the reverse distillation) causes a more subtle but still drop in perceptual and quality metrics, validating the importance of resolving the objective mismatch before fine-tuning. These studies confirm that all proposed components are critical and complementary in achieving the final strong results.

Applications

Generative artifact removal. Text-to-image diffusion models often produce high-resolution outputs with local structural artifacts, such as distorted facial features, malformed hands, or inconsistent textures. These issues are especially common in fine-scale regions and become more prominent under close inspection. RealUHR can serve as a post-processing module to refine such generated images. By performing a single denoising trajectory conditioned on the

Figure 7: Application in creative up-sampling. RealUHR can upscale real-world images (left) by a factor of $4\times$ with controllable creativity. As the noise strength parameter γ increases from 0.3 to 0.8, the output transitions from a faithful reconstruction to an imaginative reinterpretation, all while maintaining semantic coherence.

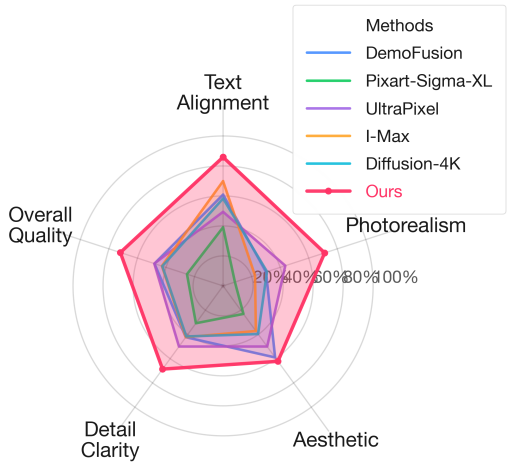

Figure 8: User study: percentage of positive ratings (%).

	Scheduling Function	FID↓	Patch FID↓	KID↓	QualiCLIP↑
2K	Uniform	40.97	29.72	9.69	<u>0.494</u>
	Logarithmic	41.22	30.15	9.75	0.489
	Cosine	<u>40.61</u>	<u>27.82</u>	<u>9.53</u>	0.492
	Ours	**40.29**	**26.09**	**9.42**	**0.495**

Table 2: Ablation study on non-uniform time scheduling. We compare ours against uniform, cosine, and logarithmic schedulers. Best results are in bold, second best are <u>underlined</u>.

original image, RealUHR corrects generation-specific artifacts while preserving the global structure and semantics. As shown in Figure 6, RealUHR significantly improves semantic coherence in critical regions (e.g., faces, flowers) without altering the overall composition.

Creative up-sampling. RealUHR can be used as a drop-in, creative upsampler for any real-world input image. By modulating a creativity control parameter, also referred to as noise strength $\gamma \in [0, 1]$, the degree of novel content injection can be adjusted. Small values of γ recover canonical super-resolution, whereas larger values progressively introduce new textures and structures. Figure 7 illustrates this

	Ablation Variant	FID↓	Patch FID↓	KID↓	QualiCLIP↑
4K	w/o GCA Stage 1	141.03	48.52	4.81	0.425
	w/o GCA (Full)	141.53	50.48	5.15	0.412
	w/o Patch Init	145.82	53.11	6.23	0.385
	Ours (Full Model)	**139.29**	**46.15**	**3.93**	**0.437**
2K	w/o GCA Stage 1	41.54	29.03	9.70	0.488
	w/o GCA (Full)	42.14	29.83	9.84	0.485
	w/o Patch Init	43.51	31.24	10.15	0.478
	Ours (Full Model)	**40.29**	**26.10**	**9.42**	**0.495**

Table 3: Ablation study validating the contributions of our core components: Patch Initialization and Guidance-Consistent Adaptation (GCA).

behavior for both an anime frame and a photorealistic portrait: as γ increases from 0.3 to 0.8 (left to right), the outputs transition smoothly from faithful reconstruction to imaginative reinterpretation while preserving global semantics. RealUHR is well-suited for concept exploration, asset re-use at 4K, and rapid variant generation.

Conclusion

In this paper, we present **RealUHR**, a novel framework that significantly advances ultra-high-resolution (UHR) text-to-image synthesis. By leveraging a *Patch-Cascade Flow Matching* pipeline, RealUHR efficiently generates globally coherent images without requiring computationally expensive patch fusion techniques. Our core innovation, the *Guidance-Consistent Adaptation* strategy, successfully overcomes the objective mismatch inherent in guidance-distilled models. This breakthrough enables the adaptation of powerful foundation models like FLUX for high-fidelity patch-aware generation. Our method's effectiveness is validated by state-of-the-art results on 2K and 4K benchmarks, where it demonstrates a superior balance of quality and speed. The versatility of RealUHR is further highlighted through its successful application to zero-shot tasks, including artifact suppression and creative up-sampling. These contributions establish RealUHR as a robust and scalable solution for the next generation of UHR synthesis. Future work will explore its extension to other modalities, such as video.

References

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

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. In *ICLR*.

Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; and Li, Z. 2024. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, 74–91. Springer.

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

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

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

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

Euler, L. 1768. *Institutionum calculi integralis*. Number Bd. 1 in Institutionum calculi integralis. imp. Acad. imp. Saènt.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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

Huang, L.; Fang, R.; Zhang, A.; Song, G.; Liu, S.; Liu, Y.; and Li, H. 2024. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European Conference on Computer Vision*, 196–212. Springer.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10124–10134.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *NeurIPS*.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Lee, Y.; Kim, K.; Kim, H.; and Sung, M. 2023. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36: 50648–50660.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

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

Liu, H.; Xu, Z.; Hong, F.-T.; Huang, H.-P.; Zhou, Y.; and Zhou, Y. 2025a. Video Motion Graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13730–13740.

Liu, H.; Yang, X.; Akiyama, T.; Huang, Y.; Li, Q.; Kuriyama, S.; and Taketomi, T. 2025b. TANGO: Co-Speech Gesture Video Reenactment with Hierarchical Audio Motion Embedding and Diffusion Interpolation. In *The Thirteenth International Conference on Learning Representations*.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS*.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Pernias, P.; Rampas, D.; Richter, M. L.; Pal, C. J.; and Aubreville, M. 2023. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*.

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

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis,

C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35: 25278–25294.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

Tong, A.; Fatras, K.; Malkin, N.; Huguet, G.; Zhang, Y.; Rector-Brooks, J.; Wolf, G.; and Bengio, Y. 2024. Improving and generalizing flow-based generative models with minibatch optimal transport. *TMLR*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 1905–1914.

Xie, E.; Chen, J.; Chen, J.; Cai, H.; Tang, H.; Lin, Y.; Zhang, Z.; Li, M.; Zhu, L.; Lu, Y.; et al. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*.

Yu, Y.; Zeng, Z.; Hua, H.; Fu, J.; and Luo, J. 2024. PromptFix: You Prompt and We Fix the Photo. In *NeurIPS*.

Yu, Y.; Zeng, Z.; Zheng, H.; and Luo, J. 2025a. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. In *ICCV*.

Yu, Y.; Zhang, L.; Fan, H.; and Luo, T. 2022. High-fidelity image inpainting with gan inversion. In *ECCV*.

Yu, Y.; Zheng, H.; Zhang, Z.; Zhang, J.; Zhou, Y.; Barnes, C.; Liu, Y.; Xiong, W.; Lin, Z.; and Luo, J. 2025b. ZipIR: Latent Pyramid Diffusion Transformer for High-Resolution Image Restoration. *arXiv preprint arXiv:2504.08591*.

Zhang, J.; Huang, Q.; Liu, J.; Guo, X.; and Huang, D. 2025. Diffusion-4K: Ultra-High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Zhang, S.; Chen, Z.; Zhao, Z.; Chen, Z.; Tang, Y.; Chen, Y.; Cao, W.; and Liang, J. 2023. HiDiffusion: Unlocking High-Resolution Creativity and Efficiency in Low-Resolution Trained Diffusion Models. *arXiv preprint arXiv:2311.17528*.

Zheng, H.; Yao, Y.; Yu, Y.; Zhou, Y.; Luo, J.; and Lin, Z. 2025. PixPerfect: Seamless Latent Diffusion Local Editing with Discriminative Pixel-Space Refinement. In *NeurIPS*.