HiWave: Training-Free High-Resolution Image Generation via Wavelet-Based Diffusion Sampling

TOBIAS VONTOBEL, ETH Zürich, Switzerland
SEYEDMORTEZA SADAT, ETH Zürich, Switzerland and Disney Research Studios, Switzerland
FARNOOD SALEHI, Disney Research Studios, Switzerland
ROMANN WEBER, Disney Research Studios, Switzerland

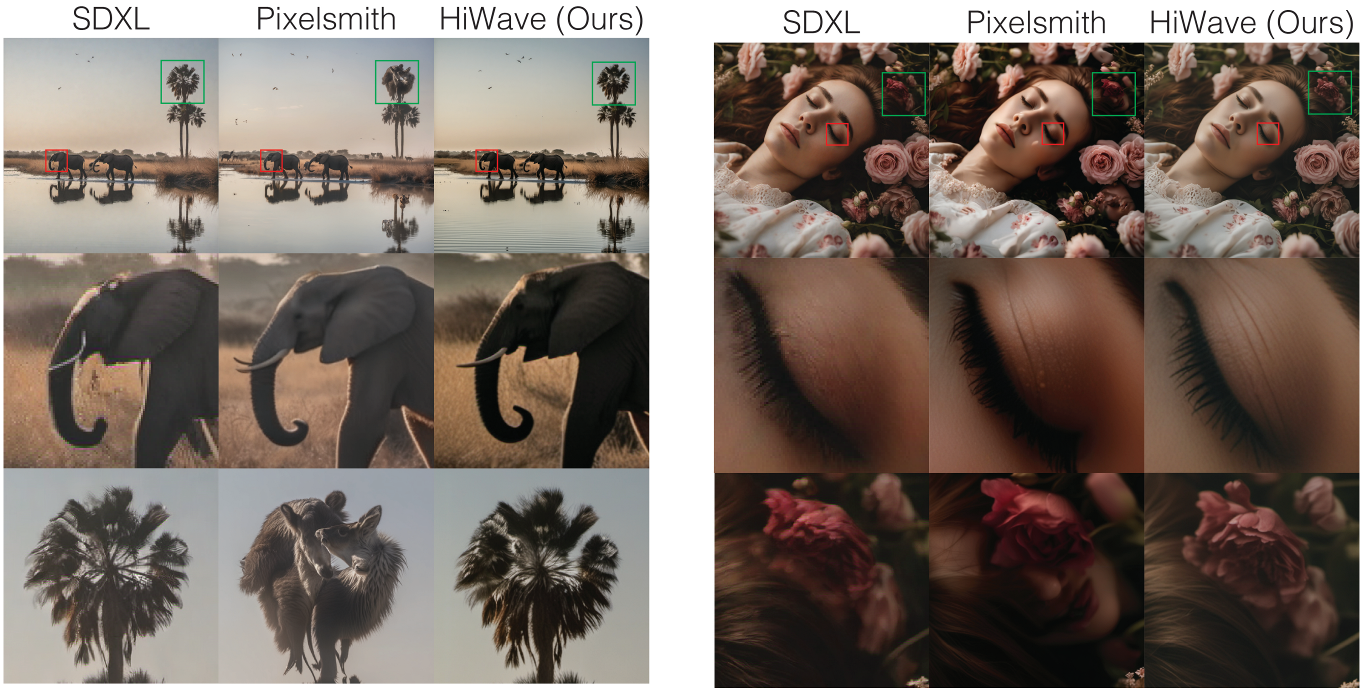

Fig. 1. We propose HiWave, a novel training-free approach for high-resolution image generation using pretrained diffusion models. While standard Stable Diffusion XL (SDXL) can produce globally coherent images, it lacks fine details when upscaled to 4096×4096 resolution (left column). Existing training-free methods (e.g., Pixelsmith [Tragakis et al. 2024]) enhance details in SDXL outputs but often introduce duplicated objects and visual artifacts (middle column). In contrast, HiWave leverages a patch-wise DDIM inversion strategy combined with a wavelet-based detail enhancer module to produce high-quality images with rich details and minimal duplication artifacts. The second and third rows show 10× and 5× magnified views of the red and green boxed regions, respectively.

Diffusion models have emerged as the leading approach for image synthesis, demonstrating exceptional photorealism and diversity. However, training diffusion models at high resolutions remains computationally prohibitive, and existing zero-shot generation techniques for synthesizing images beyond training resolutions often produce artifacts, including object duplication and spatial incoherence. In this paper, we introduce HiWave, a training-free, zero-shot approach that substantially enhances visual fidelity and structural coherence in ultra-high-resolution image synthesis using pretrained diffusion models. Our method employs a two-stage pipeline: generating a base image from the pretrained model followed by a patch-wise DDIM inversion step and a novel wavelet-based detail enhancer module. Specifically, we first utilize inversion methods to derive initial noise vectors that preserve global coherence from the base image. Subsequently, during sampling, our wavelet-domain detail enhancer retains low-frequency components from the base image to ensure structural consistency, while selectively guiding high-frequency components to enrich fine details and textures. Extensive evaluations using Stable Diffusion XL demonstrate that HiWave effectively mitigates common visual artifacts seen in prior methods, achieving superior perceptual quality. A user study confirmed HiWave's performance, where it was preferred over the state-of-the-art alternative in more than 80% of comparisons, highlighting its effectiveness for high-quality, ultra-high-resolution image synthesis without requiring retraining or architectural modifications.

Authors' Contact Information: Tobias Vontobel, ETH Zürich, Zürich, Switzerland, votobias@ethz.ch; Seyedmorteza Sadat, ETH Zürich, Zürich, Switzerland and Disney Research Studios, Zürich, Switzerland, ssadat@ethz.ch; Farnood Salehi, Disney Research Studios, Zürich, Switzerland, farnood.salehi@disneyresearch.com; Romann Weber, Disney Research Studios, Zürich, Switzerland, romann.weber@disneyresearch.com.

SA Conference Papers '25, Hong Kong, Hong Kong
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2137-3/25/12
https://doi.org/10.1145/3757377.3763979

CCS Concepts: • **Computing methodologies → Computer vision**.

Additional Key Words and Phrases: Image synthesis, diffusion models, high-resolution generation, discrete wavelet transform

ACM Reference Format:
Tobias Vontobel, Seyedmorteza Sadat, Farnood Salehi, and Romann Weber. 2025. HiWave: Training-Free High-Resolution Image Generation via Wavelet-Based Diffusion Sampling. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3757377.3763979

1 Introduction

Since the introduction of diffusion models, generative image synthesis has reached unprecedented levels of photorealism and creative control. Recent models such as Stable Diffusion [Esser et al. 2024; Podell et al. 2023a] can generate stunning images at resolutions up to 1024×1024 pixels. Despite these improvements in image quality, producing outputs beyond 1024×1024 remains technically challenging due to the substantial computational demands associated with training at higher resolutions.

Diffusion models typically rely on large-scale networks with high parameter counts to achieve optimal image quality and prompt alignment [Esser et al. 2024]. As input resolution increases, the computational cost of training these models becomes prohibitive—especially for attention-based architectures [Esser et al. 2024; Peebles and Xie 2023], where complexity scales quadratically with spatial dimensions. Moreover, most datasets used to train large-scale diffusion models lack high-resolution content beyond 1024×1024. As a result, current state-of-the-art models are typically limited to moderate resolutions, restricting their applicability in domains such as advertising and film production, where ultra high-resolution outputs (e.g., 4K) are required.

Motivated by these limitations, recent work has explored methods to extend pretrained diffusion models to higher resolutions—i.e., beyond their native training sizes. These approaches fall into two main categories: *patch-based* techniques that process image regions independently, such as Pixelsmith [Tragakis et al. 2024] and DemoFusion [Du et al. 2024], and *direct inference* methods that modify model architectures, such as HiDiffusion [Zhang et al. 2023] and FouriScale [Huang et al. 2024]. However, these methods face fundamental limitations. Patch-based approaches often produce duplicated objects (Figure 2a), while direct inference methods struggle to maintain global coherence at very high resolutions (e.g., beyond 2048×2048, see Figure 2b). This highlights the need for a training-free, high-resolution generation approach that ensures both global coherence and robustness to duplication and artifacts.

In this paper, we propose HiWave, a novel pipeline for training-free high-resolution image generation that produces globally coherent outputs without object duplication. HiWave adopts a two-stage, patch-based approach that preserves the coherent structure of a base image generated by a pretrained model, while enhancing the fine details required for higher resolutions. In the first stage, a base image is generated at a standard resolution (e.g., 1024×1024) using a pretrained diffusion model. This image is then upscaled in the image domain to the target high resolution, though the upscaled result lacks fine-grained details. To enrich the image with high-frequency details, we introduce a novel sampling module based on patch-wise DDIM inversion. Specifically, each image patch is inverted using DDIM to recover the corresponding latent noise that would generate the given input. Sampling is then performed starting from this inverted noise. However, to prevent the model from simply reproducing the original image, we incorporate a *detail enhancer* into the sampling process. This component is based on DWT, and it preserves the low-frequency content of the base image to maintain global structure while guiding the high-frequency components to synthesize realistic additional details suitable for the target resolution.

Our approach enables standard diffusion models trained at a resolution of 1024×1024 to generate images at 4096×4096 resolution—a 16× increase in total pixel count. HiWave leverages the high-frequency priors inherently captured by pretrained diffusion models, as observed by Du et al. [2024], to generate coherent images with fine details. Compared to existing methods, HiWave excels at maintaining structural coherence while significantly reducing hallucinations and duplicated artifacts, addressing a key limitation in current zero-shot high-resolution generation pipelines. We validate the effectiveness of HiWave using Stable Diffusion XL [Podell et al. 2023b] and compare it against Pixelsmith [Tragakis et al. 2024], the current state-of-the-art in training-free high-resolution image generation. In a user study, HiWave was preferred over Pixelsmith in more than 80% of cases, highlighting its superior visual fidelity.

In summary, our main contributions are the following:

- We introduce HiWave, a training-free, two-stage pipeline for high-resolution image synthesis that extends pretrained diffusion models to ultra high-resolutions, without requiring architectural modifications or additional training.
- We propose a novel patch-wise DDIM inversion framework coupled with a wavelet-based detail enhancer, which preserves global structure from the base image while selectively enriching its high-frequency details.
- We demonstrate that HiWave effectively mitigates common artifacts such as duplicated objects and structural inconsistencies that persist in prior methods.
- We conduct extensive evaluations with Stable Diffusion XL and show that HiWave outperforms current state-of-the-art methods in qualitative comparisons and a preference study, with users favoring HiWave outputs in over 80% of cases.

2 Related Work

Diffusion models have become a dominant framework for image synthesis due to their strong generative capabilities [Denton et al. 2015; Ho et al. 2020; Song et al. 2020b]. They have rapidly surpassed previous generative modeling techniques in terms of fidelity and diversity [Dhariwal and Nichol 2021; Nichol and Dhariwal 2021], achieving state-of-the-art performance across a wide range of applications, including unconditional image synthesis [Dhariwal and Nichol 2021; Karras et al. 2022], text-to-image generation [Balaji et al. 2022; Esser et al. 2024; Podell et al. 2023b; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022b; Yu et al. 2022], video synthesis [Blattmann et al. 2023a,b; Gupta et al. 2023], image-to-image translation [Liu et al. 2023; Saharia et al. 2022a], motion synthesis

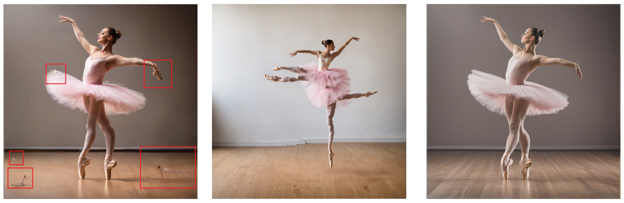

Fig. 2. Qualitative comparison of high-resolution image generation methods at 4096×4096 resolution. DemoFusion, a patch-based method, exhibits object duplication artifacts (highlighted with red rectangles). HiDiffusion, a direct inference method, lacks structural coherence and fine details. In contrast, our method produces coherent generations with rich detail.

[Tevet et al. 2023; Tseng et al. 2023], and audio synthesis [Chen et al. 2021; Huang et al. 2023; Kong et al. 2021].

Despite this progress, diffusion models often incur high computational costs and long training times [Chen et al. 2023], especially when handling high-resolution data. Latent diffusion models (LDMs) [Rombach et al. 2022] alleviate some of this burden by compressing inputs into a smaller latent space using a pretrained autoencoder. However, LDMs typically scale only up to moderate resolutions (e.g., 1024×1024). While some recent works have explored training LDMs at higher resolutions [Chen et al. 2024; Xie et al. 2024], they continue to face challenges related to prolonged training times and the limited availability of high-quality ultra-high-resolution datasets. These constraints have spurred growing interest in training-free methods that exploit pretrained models to generate images beyond their native resolutions.

One popular direction decomposes high-resolution generation into patches that are processed independently and later combined to form the full image. Examples include DemoFusion [Du et al. 2024], AccDiffusion [Lin et al. 2024], and Pixelsmith [Tragakis et al. 2024]. This patch-based strategy allows diffusion models to operate at their native resolution for each patch, thereby preserving the detail and expressiveness of the base model. However, these methods often fail to maintain global coherence at resolutions beyond 2048×2048, frequently producing boundary artifacts, duplicated content, and semantic inconsistencies across patches.

Another line of work modifies the architecture or inference process of pretrained diffusion models—without additional training—to enable single-pass generation of high-resolution images. Notable examples include FouriScale [Huang et al. 2024], MegaFusion [Wu et al. 2025], and HiDiffusion [Zhang et al. 2023]. These methods introduce attention scaling or downsampling blocks into the pretrained network to better align intermediate features with those seen at the model's native resolution. While such methods avoid patch-based artifacts, their performance often degrades at ultra-high resolutions (e.g., 4096×4096). As the gap between training and inference resolutions widens, these approaches struggle to preserve global coherence—leaving patch-based strategies as the current state-of-the-art for zero-shot high-resolution image generation.

In summary, current training-free approaches either fail to maintain global coherence compared to the base diffusion model or suffer from duplicated objects and artifacts at ultra-high resolutions (e.g., 4096×4096). HiWave proposes a novel patch-based strategy aimed at achieving coherent, high-quality image generation at such resolutions while avoiding the limitations of existing methods.

3 Background

Diffusion models. Let $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$ represent a data sample, and let $t \in [0, T]$ denote a continuous time variable. In the forward diffusion process, noise is incrementally added as $\mathbf{z}_t = \mathbf{x} + \sigma(t)\boldsymbol{\epsilon}$, where $\sigma(t)$ is a time-dependent noise scale. This noise schedule gradually corrupts the data, with $\sigma(0) = 0$ (clean data) and $\sigma(T) = \sigma_{\max}$ (maximum corruption). As shown in Karras et al. [2022], this process corresponds to the following differential equation:

$$\mathrm{d}\mathbf{z} = -\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)\mathrm{d}t, \tag{1}$$

where $p_t(\mathbf{z}_t)$ denotes the distribution over noisy samples at time t, transitioning from $p_0 = p_{\text{data}}$ to $p_T = \mathcal{N}(\mathbf{0}, \sigma_{\max}^2\mathbf{I})$. Sampling from the data distribution involves solving this ODE in reverse (from $t = T$ to $t = 0$), provided the score function $\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)$ is known. Since this score function is intractable, it is approximated by a neural denoiser $D_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$, trained to reconstruct the clean signal $\mathbf{x}$ from its noisy observation $\mathbf{z}_t$. For conditional generation, the denoiser is extended with a conditioning variable $\mathbf{y}$—such as a label or text prompt—resulting in $D_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{y})$.

Inversion methods. Inversion techniques, such as DDIM inversion [Song et al. 2021] solve the forward-time version of the diffusion ODE in Equation 1 to map an observed image $\mathbf{x}$ to its corresponding noise vector $\mathbf{z}_T$. By integrating Equation 1 from $t = 0$ to $t = T$, one obtains a deterministic mapping from the image back to the noise space (in the limit of small steps). This enables applications like image editing, as it provides a way to anchor sampling around a given input while retaining its global layout.

Classifier-free guidance. Classifier-free guidance (CFG) [Ho and Salimans 2022] is an inference-time technique that improves generation quality by blending predictions from conditional and unconditional models. During sampling, CFG adjusts the denoiser output as:

$$\hat{D}_{\text{CFG}}(\mathbf{z}_t, t, \mathbf{y}) = D_{\boldsymbol{\theta}}(\mathbf{z}_t, t) + w(D_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{y}) - D_{\boldsymbol{\theta}}(\mathbf{z}_t, t)), \tag{2}$$

where w is a guidance strength parameter (with $w = 1$ representing unguided sampling). The unconditional model $D_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$ is typically learned by randomly dropping conditioning inputs $\mathbf{y}$ during training. Alternatively, separate unconditional models can be used [Karras et al. 2023]. Analogous to the truncation trick in GANs [Brock et al. 2019], CFG improves visual fidelity, but may lead to oversaturation [Sadat et al. 2025] or reduced diversity [Sadat et al. 2024].

Discrete wavelet transform. Discrete wavelet transforms (DWT) [Brewster 1993] are a fundamental tool in signal processing, commonly used to analyze spatial-frequency content in data. The transformation utilizes a pair of filters—a low-pass filter L and a high-pass filter H. For 2D signals, these are combined to form four distinct filter operations: $LL^\top$, $LH^\top$, $HL^\top$, and $HH^\top$. When applied to an image $\mathbf{x}$, the 2D wavelet transform decomposes it into one low-frequency component $\mathbf{x}_L$ and three high-frequency components $\mathbf{x}_H, \mathbf{x}_V, \mathbf{x}_D$, capturing horizontal, vertical, and diagonal details, respectively.

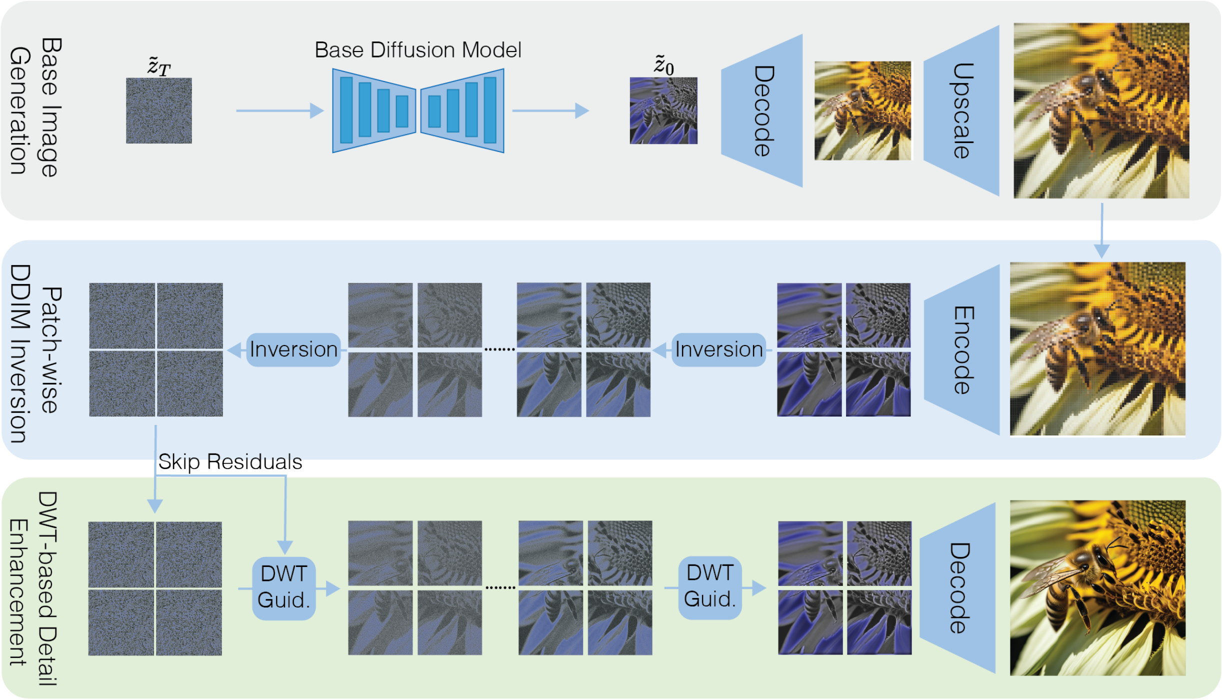

Fig. 3. Overview of HiWave, our training-free high-resolution image generation pipeline. We first generate a base image using a pretrained model through a standard sampling process that transforms random noise (z_T) into a clean image ($\tilde{z}_0$) conditioned on a text prompt. This image is then upscaled in the image domain using Lanczos interpolation and encoded back into the latent space via the VAE encoder to enrich the base image with additional details. A patch-wise DDIM inversion process is then performed, mapping the upscaled image back to its corresponding noise representation. Finally, our DWT-based detail enhancement approach applies frequency-selective guidance during denoising, using wavelet decomposition to independently control low-frequency structure and guide high-frequency components for finer details. Skip residuals are also incorporated during the early sampling steps to further preserve the global coherence of the base image. This pipeline enables HiWave to generate high-quality, high-resolution images without duplications.

Each of these sub-bands has spatial dimensions of $H/2 \times W/2$ for an image of size $H \times W$. Multiscale decomposition can be achieved by recursively applying the transform to the low-frequency component $\mathbf{x}_L$. The transformation is fully invertible, enabling exact reconstruction of the original image $\mathbf{x}$ from the set $\{\mathbf{x}_L, \mathbf{x}_H, \mathbf{x}_V, \mathbf{x}_D\}$ using the inverse discrete wavelet transform (iDWT).

4 Method

We now describe the details of HiWave, our framework for high-resolution image generation using pretrained diffusion models. An overview of the complete pipeline is shown in Figure 3, and we detail each component below. We define $D_c(\mathbf{z}_t) \doteq D_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{y})$, $D_u(\mathbf{z}_t) \doteq D_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$, and $\hat{D}_{\text{CFG}}(\mathbf{z}_t) \doteq \hat{D}_{\text{CFG}}(\mathbf{z}_t, t, \mathbf{y})$ to represent the conditional, unconditional, and CFG outputs, respectively.

4.1 Base image generation

We begin by generating a base image at the native resolution of the pretrained diffusion model, typically 1024×1024. This image is then upscaled in the image domain using Lanczos interpolation. Unlike most previous works that perform upscaling in the latent space [Du et al. 2024], we opt for image-domain upscaling to avoid artifacts commonly introduced by latent-space interpolation. These artifacts arise because standard VAEs used in diffusion pipelines are not equivariant to scaling operations, leading to inconsistencies when upscaling is applied in latent space [Kouzelis et al. 2025]. An example of such artifacts is illustrated in Figure 4. To avoid this, we first upscale the base image to the target resolution (e.g.,

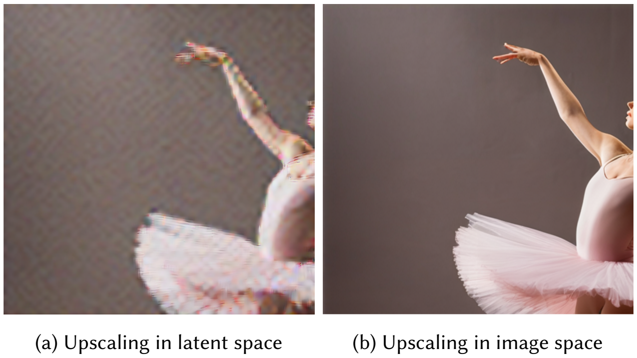
(a) Upscaling in latent space (b) Upscaling in image space

Fig. 4. Comparison of upscaling in image space vs latent space. Interpolation performed directly in latent space introduces severe spatial artifacts, as standard VAEs are not equivariant to scaling operations. In contrast, interpolating in the image space after decoding preserves structural consistency and visual quality.

4096×4096) in the image domain. At this point, we obtain an image at the target resolution, albeit lacking fine-grained details. We then encode this upscaled image into the latent space via the VAE encoder and apply a patch-based sampling process (described next) to refine high-resolution details while preserving the global structure of the original image.

4.2 Patch-wise DDIM inversion

To preserve structural coherence during patch-wise generation, we initialize the diffusion process using DDIM inversion [Song et al. 2020a] instead of random noise. This inversion retrieves noise vectors for each image patch by integrating the diffusion ODE forward in time:

$$\mathbf{z}_{t+1} \approx \mathbf{z}_t - \dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)\Delta t, \tag{3}$$

This deterministic initialization provides two key benefits: (1) Controlled noise for frequency decomposition: DDIM-inverted noise retains meaningful structural information and spatial layout from the original image, enabling our detail enhancement module to selectively refine high-frequency textures while preserving the low-frequency structural content. (2) Consistent patch initialization: Neighboring patches receive compatible noise vectors, which helps maintain continuity and avoid visible seams across patch boundaries. In contrast, initializing with random Gaussian noise loses the structural context of the base image and often introduces artifacts or structural inconsistencies. This controlled inversion lays the groundwork for our DWT-based detail enhancer in the next step, enabling detail enhancement while preserving global coherence.

4.3 Detail enhancement with DWT guidance

In the next step, we begin sampling from the inverted noise of each patch, with the goal of progressively adding details to the final image. A central challenge in patch-based high-resolution generation is maintaining a balance between global coherence and detailed texture synthesis. Existing methods often fall short in one of these areas, resulting in either artifact-prone high-frequency patterns or globally consistent images that lack detail.

To address this, we introduce a DWT-based detail enhancement module that leverages the complementary roles of the low- and high-frequency components of each latent. We argue that low-frequency components typically capture structural coherence, while high-frequency components convey fine details and textures. This distinction is especially critical in patch-wise generation, where each patch must integrate seamlessly into the global image for overall coherence while still containing sufficient detail at high resolutions.

Since we used DDIM inversion to obtain the final noise, following the sampling process using only the conditional prediction $D_c(\mathbf{z}_t)$ would reproduce the base image, i.e., globally coherent but lacking in fine detail. This motivates our decision to preserve the low-frequency bands from $D_c(\mathbf{z}_t)$, which retain much of the base image's global layout due to the DDIM inversion. To enrich each patch with the required details for high-quality generation, we guide the high-frequency components adaptively using a modified CFG strategy.

Specifically, we apply the DWT to both the conditional and unconditional predictions to get

$$\mathtt{DWT}(D_c(\mathbf{z}_t)) = \left\{D_c^L(\mathbf{z}_t), D_c^H(\mathbf{z}_t), D_c^V(\mathbf{z}_t), D_c^D(\mathbf{z}_t)\right\}, \tag{4}$$

$$\mathtt{DWT}(D_u(\mathbf{z}_t)) = \left\{D_u^L(\mathbf{z}_t), D_u^H(\mathbf{z}_t), D_u^V(\mathbf{z}_t), D_u^D(\mathbf{z}_t)\right\}. \tag{5}$$

We then construct the guided prediction in the frequency domain as follows:

$$\tilde{D}_{\mathrm{CFG}}^L(\mathbf{z}_t) = D_c^L(\mathbf{z}_t), \tag{6}$$

$$\tilde{D}_{\mathrm{CFG}}^H(\mathbf{z}_t) = D_u^H(\mathbf{z}_t) + w_d(D_c^H(\mathbf{z}_t) - D_u^H(\mathbf{z}_t)), \tag{7}$$

$$\tilde{D}_{\mathrm{CFG}}^V(\mathbf{z}_t) = D_u^V(\mathbf{z}_t) + w_d(D_c^V(\mathbf{z}_t) - D_u^V(\mathbf{z}_t)), \tag{8}$$

$$\tilde{D}_{\mathrm{CFG}}^D(\mathbf{z}_t) = D_u^D(\mathbf{z}_t) + w_d(D_c^D(\mathbf{z}_t) - D_u^D(\mathbf{z}_t)). \tag{9}$$

Finally, we apply the inverse DWT to reconstruct the full guided signal:

$$\tilde{D}_{\mathrm{CFG}}(\mathbf{z}_t) = \mathtt{iDWT}(\left\{\tilde{D}_{\mathrm{CFG}}^L(\mathbf{z}_t), \tilde{D}_{\mathrm{CFG}}^H(\mathbf{z}_t), \tilde{D}_{\mathrm{CFG}}^V(\mathbf{z}_t), \tilde{D}_{\mathrm{CFG}}^D(\mathbf{z}_t)\right\}). \tag{10}$$

This frequency-aware guidance strategy enables precise enhancement of details while preserving the global structure of the base image.

4.4 Skip residuals

To further preserve global structure during early denoising, we incorporate skip residuals by mixing the latents obtained from DDIM inversion with those from the sampling process for each patch. Let $\mathbf{z}_t$ denote the current latent in the sampling process, $\mathbf{z}_t^s$ the corresponding DDIM-inverted latent, τ a time step threshold, and $c_1 = ((1 + \cos(\frac{T-t}{T}\pi))/2)^\alpha$ a cosine-decay weighting factor. The skip residual update is defined as

$$\hat{\mathbf{z}}_t = \begin{cases} c_1 \times \mathbf{z}_t + (1 - c_1) \times \mathbf{z}_t^s & t \geq \tau, \\ \mathbf{z}_t & t < \tau. \end{cases} \tag{11}$$

Unlike prior work that applies skip residuals throughout all diffusion steps, we adopt a more conservative strategy: they are only used during the initial denoising phase. This allows the model to leverage the base image's structure early on, and then progressively diverge to synthesize novel details guided by the DWT-enhanced predictions. In contrast, applying skip residuals at all time steps–as done in previous work–can suppress detail synthesis, while omitting them entirely causes duplication artifacts. Our DWT-based enhancer mitigates this trade-off by explicitly guiding different frequency bands, enabling the generation of rich textures while preserving global consistency.

4.5 Implementation Details

We use the sym4 wavelet for DWT due to its effective balance between spatial and frequency localization. The detail guidance strength is set to $w_d = 7.5$, enhancing high-frequency features while preserving the low-frequency structure from $D_c(\mathbf{z}_t)$.

Image generation is performed progressively—first at 1024×1024 (the model's native resolution), then at 2048×2048, and finally at 4096×4096. While some prior works report increased duplication artifacts with iterative upscaling [Tragakis et al. 2024], we did not observe this in our experiments, likely due to the combination of patch-wise DDIM inversion and our DWT-based guidance.

While our evaluation focuses on square resolutions for standardized comparison, HiWave naturally extends to arbitrary aspect ratios. The framework first generates a rectangular base image at the desired aspect ratio using the pretrained model, then applies the same patch-wise DDIM inversion and DWT-based guidance to rectangular patches that conform to the target dimensions. This maintains the same overlap strategy and frequency-domain enhancement across all patches regardless of the final image aspect ratio (see Figure 5 in Appendix for some examples).

Skip residuals are applied only during the first 15 timesteps (out of 50) for 2048×2048 resolution and the first 30 timesteps for 4096×4096. This conservative strategy contrasts with methods that apply skip residuals across the entire diffusion process. By limiting their use to early steps, we preserve the global structure initially while allowing the model to synthesize novel details in later time steps.

For patch processing, we use a 50% overlap between adjacent patches to ensure smooth transitions. To optimize memory usage, we employ a streaming approach in which patches are processed in batches rather than all at once, enabling 4096×4096 image generation on consumer GPUs with 24GB of VRAM.

5 Results

We now evaluate HiWave both qualitatively and quantitatively against state-of-the-art high-resolution image generation methods. Our evaluation focuses on three main aspects: (1) avoiding the duplication artifacts commonly seen in previous patch-based methods, (2) achieving global image coherence and high fidelity at 4096×4096 resolution, and (3) enhancing both quality and details over the original Stable Diffusion XL outputs.

5.1 Experimental setup

We select Pixelsmith [Tragakis et al. 2024] as the leading patch-based approach and HiDiffusion [Zhang et al. 2023] as the representative of direct inference methods. All methods were used to generate 4096×4096 resolution images from the same prompts and identical random seeds, using Stable Diffusion XL as the base model. To ensure a fair comparison, we employed the official codebases of all baselines. All experiments were conducted on a single RTX 4090 GPU with 24GB of VRAM.

To benchmark the methods, we used 1000 randomly sampled prompts from the LAION/LAION2B-en-aesthetic [Schuhmann et al. 2022] dataset, covering a diverse range of content including natural landscapes, human portraits, animals, architectural scenes, and close-up textures. This diversity allowed us to assess performance across a broad set of generation challenges.

For quantitative evaluation, prior work has shown that existing metrics are often unreliable at high resolutions, as they typically downscale images to lower resolutions (e.g., 224×224) before computing the score [Du et al. 2024; Tragakis et al. 2024; Zhang et al. 2023]. Consequently, we rely primarily on a human study, which provides the most reliable assessment of perceptual quality in this setting. Standard quantitative metrics are also reported in the Appendix for completeness.

5.2 Qualitative comparison with prior methods

Figure 5 presents a comprehensive visual comparison of our method against Pixelsmith and HiDiffusion across nine diverse test cases. Each row corresponds to a different subject, with the left columns displaying the full-resolution outputs and the right columns showing magnified regions (highlighted by green boxes) to facilitate detailed inspection of fine structures. HiDiffusion produces outputs that lack coherent structure and exhibit blurred textures across all nine examples. This demonstrates that, while direct inference methods avoid duplication and patch-based artifacts, they often struggle to produce globally coherent outputs with fine-grained detail at high resolutions. In contrast, Pixelsmith generates more detailed images but frequently suffers from object duplication. For instance, it produces duplicated humans in rows 1 through 4 and a phantom figure in the grass background of row 5. By comparison, HiWave consistently generates high-quality images free from artifacts and duplication. This illustrates that HiWave effectively addresses a key limitation of prior approaches, enabling coherent and detailed high-resolution generation without duplication.

5.3 Detail enhancement over the base image

Figure 7 illustrates how HiWave enhances fine details and can improve semantic plausibility in base images generated by the SDXL model at the original resolution. Our method retains the same overall composition but successfully extracts and amplifies fine details that were merely suggested in the original SDXL generations. In row 1, the intricate blue patterns on the porcelain bottle show substantially improved definition, with clear brushwork details and subtle variations in the blue pigment that are barely distinguishable in the original image. Row 2 illustrates how our method reveals individual yarn strands and stitch patterns of a knitted toy mouse with great clarity. In row 3, a child in the flower field shows how our method can handle natural scenes with finer hair detail, clothing texture, and surrounding flora detail without losing scene composition. These results demonstrate that HiWave effectively adds the fine-grained details necessary for high-quality high-resolution image generation.

5.4 Human evaluation study

To validate our qualitative observations, we conducted a comprehensive human preference study comparing HiWave against Pixelsmith. Using 32 image pairs generated with identical prompts and seeds at 4096×4096 resolution, we presented participants with randomized blind A/B tests. The 32 prompts were selected solely based on the SDXL base outputs–prior to running any methods–to avoid selection bias. The participants were asked to select their preferred image based on overall quality, coherence, and absence of artifacts. Figure 8 shows the preference scores for HiWave and Pixelsmith. Across 548 independent evaluations, HiWave was preferred in 81.2% of responses (445 out of 548), with seven test cases achieving 100% preference for our method. The full set of per-question preference percentages is provided in the Appendix. This strong preference aligns with our qualitative findings regarding artifact reduction and coherence preservation in Figure 5.

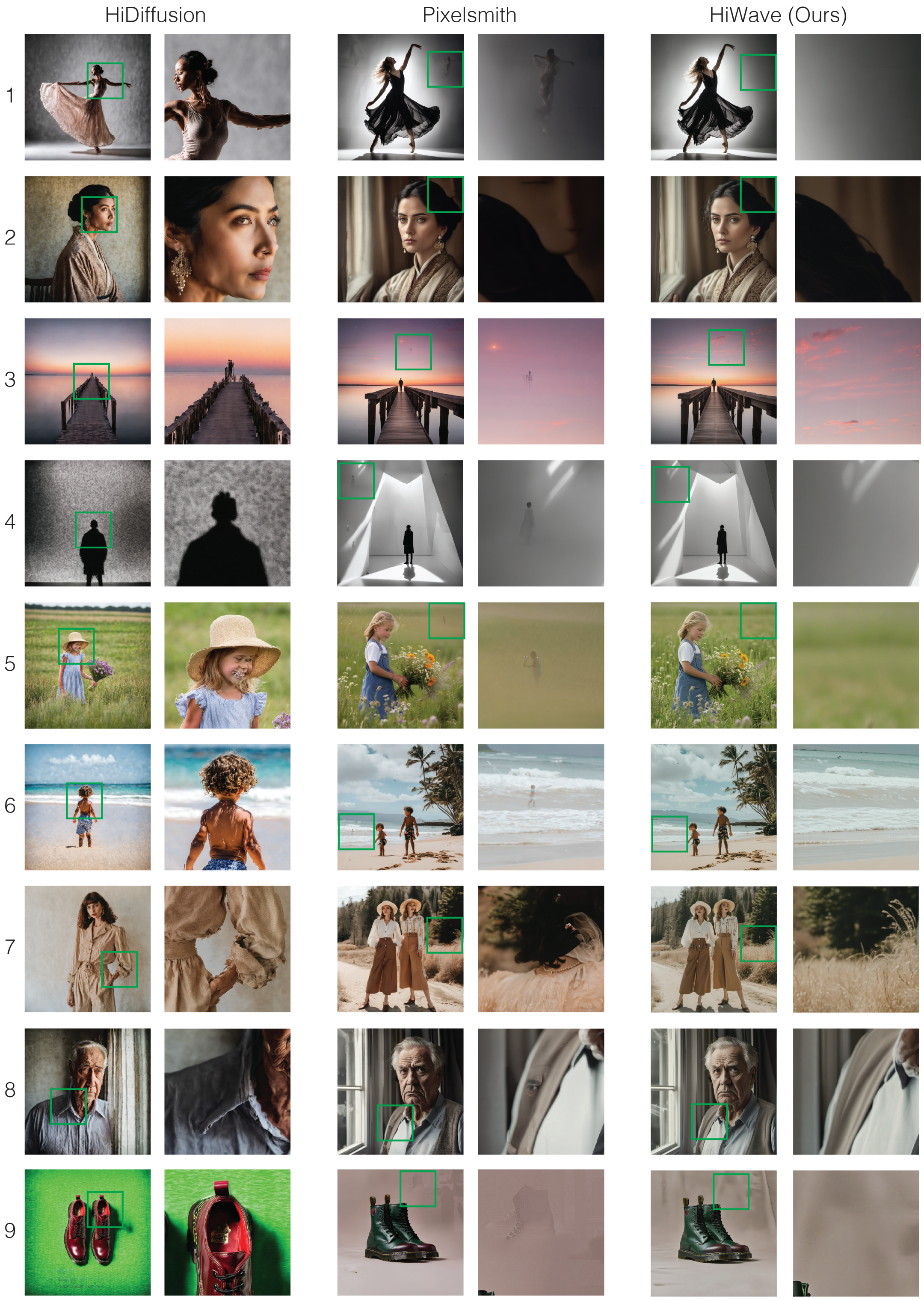

Fig. 5. Qualitative comparison of high-resolution (4096×4096) image generation across three methods. HiDiffusion (left column) consistently struggles to produce realistic details and coherent structures, leading to blurry textures and distorted features. Pixelsmith (middle column) generally generates high-quality details but exhibits noticeable duplication artifacts—particularly in background elements and textures—as highlighted in the zoomed regions (green boxes). In contrast, HiWave (right column) maintains structural coherence and delivers sharp, artifact-free generations without duplications.

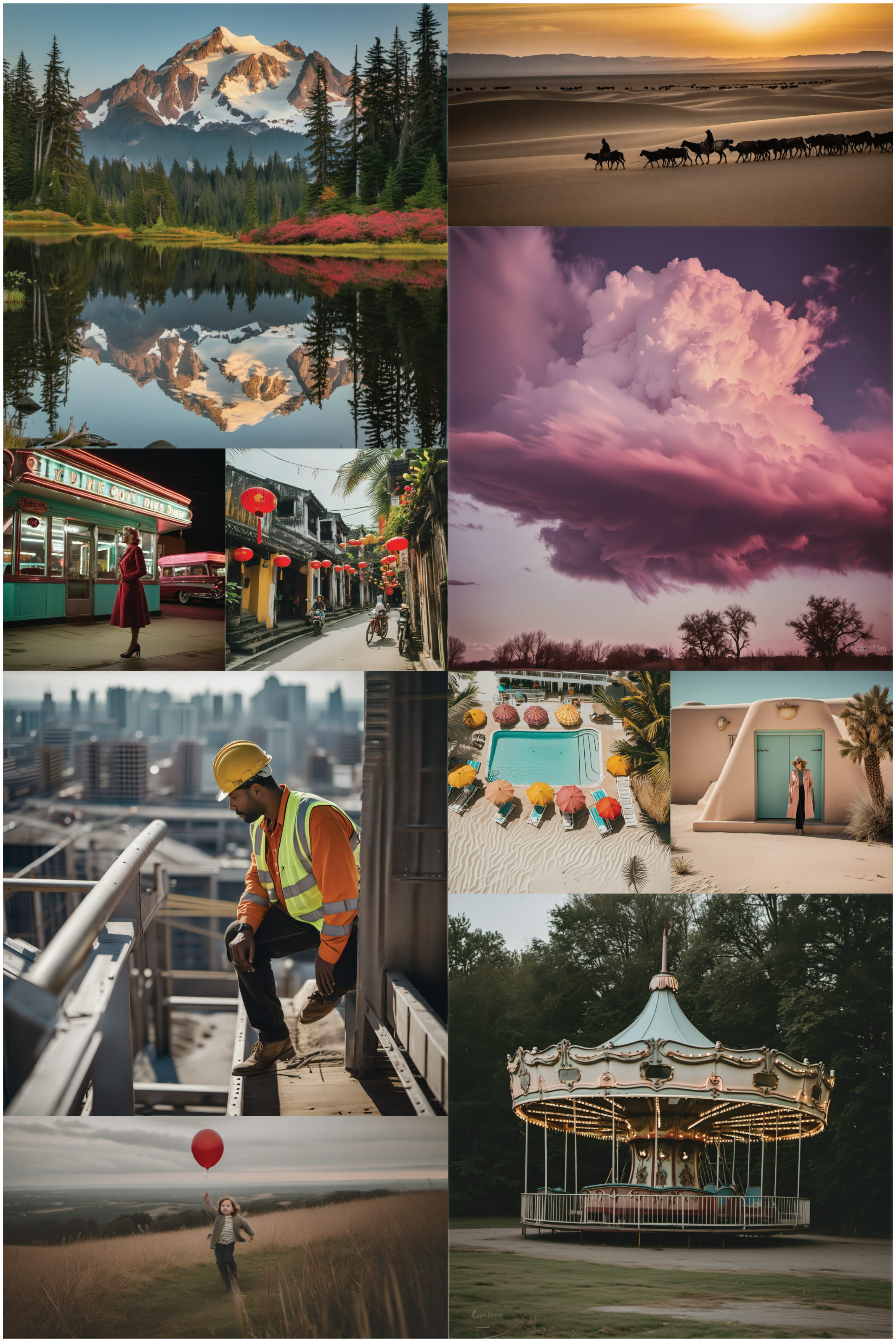

Fig. 6. Examples of high-resolution (4096×4096 and 2048×4096) images generated by our method, illustrating a variety of subjects across diverse visual motifs.

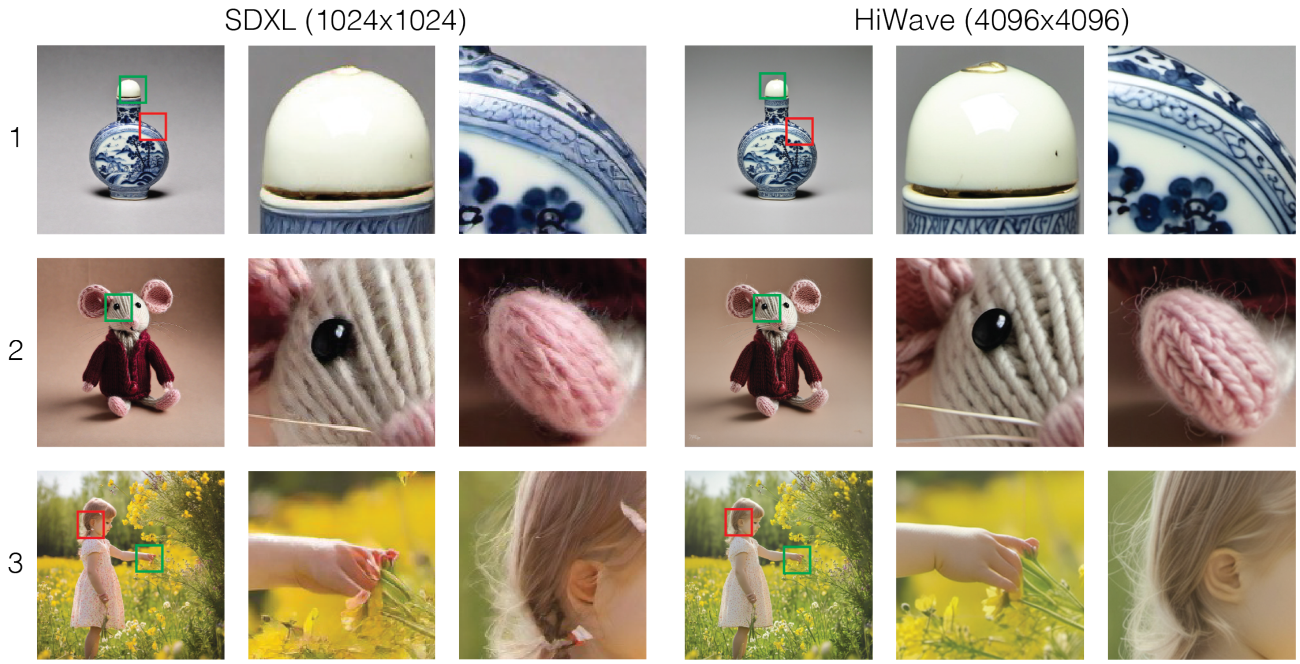

Fig. 7. Comparison between our HiWave method at 4096×4096 resolution (left) and the SDXL base model at 1024×1024 resolution (right). Each row displays a different image along with corresponding zoomed-in regions to highlight detail enhancement. Row 1 features a porcelain bottle with intricate blue patterns; Row 2 shows a knitted toy mouse with clearly visible yarn texture; and Row 3 depicts a child in a flower field, with enhanced hair and fabric details. The zoomed regions illustrate how HiWave preserves the overall composition generated by SDXL while significantly enhancing fine details that are only partially present in the original generations.

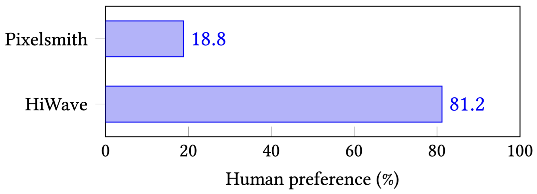

Fig. 8. User preference comparison between HiWave and Pixelsmith. HiWave was preferred by participants in over 80% of cases, validating its effectiveness in generating high-quality images at high resolutions.

5.5 Ablation Study

We conduct comprehensive ablation studies to analyze HiWave's performance and validate the contribution of each component. Our experiments demonstrate HiWave's compatibility across different architectures, including Stable Diffusion v2.1, DiT-XL, and SDXL-Flash with reduced sampling steps, confirming the broad applicability of our approach.

Our ablation studies reveal several key insights. First, HiWave improves both global structure and fine details compared to the base image generated with SDXL at 1024×1024. Second, removing guidance from low-frequency components effectively eliminates duplication artifacts. Third, initializing sampling from DDIM-inverted noise is crucial to avoid geometric and color inconsistencies across patches, while this step alone is insufficient—DWT-based frequency guidance remains essential for fully mitigating duplication and producing highly detailed, artifact-free images.

Additionally, we confirm that multistep generation produces sharper details than single-step generation, without introducing duplication in our setup, unlike what has been reported in prior work [Tragakis et al. 2024]. We also demonstrate HiWave's ability to upscale natural (non-AI-generated) images in a zero-shot manner and generate artifact-free images at ultra-high resolutions up to 8192×8192. Detailed results for all ablation studies are provided in the supplementary materials.

6 Limitations and Future Work

Since our method builds on a pretrained diffusion model, its performance naturally reflects the quality of the base generation. While HiWave consistently improves resolution and enhances details, certain structural inconsistencies in the original outputs (e.g., misshapen hands or minor facial asymmetries) may be partially retained, though often at a reduced severity. These cases typically align with well-known challenges of diffusion models, such as rendering hands, maintaining perfect facial symmetry, or handling text and repetitive structures.

Additionally, HiWave may produce slightly different color tones compared to the base SDXL generation. This occurs because our DWT-based guidance preserves low-frequency components (which encode color information) without applying CFG guidance, while the original SDXL uses CFG across all frequencies. Since CFG typically increases saturation, our approach yields different color tones. For users preferring the original color characteristics, we provide an optional post-processing step that adjusts the color statistics of the HiWave output to match the base image.

Our frequency-aware guidance mechanism substantially reduces duplication artifacts relative to existing approaches, though in rare and complex cases some residual artifacts may persist. This is particularly relevant for natural elements that inherently repeat within a scene (e.g., clusters of trees, rocks, or architectural motifs) where distinguishing between true repetition and artifacts remains challenging. Potential avenues for future work include adaptive parameter tuning that dynamically adjusts guidance strength based on content, as well as hierarchical patch frameworks for more robust inter-patch consistency at extreme resolutions.

Overall, these considerations represent manageable challenges rather than fundamental barriers. HiWave advances the state of high-fidelity, high-resolution generative modeling by demonstrating that training-free approaches can achieve compelling 4K and beyond outputs, significantly broadening the practical applicability of diffusion-based generation.

7 Conclusion

In this work, we introduced HiWave, a novel zero-shot pipeline that enables pretrained diffusion models to generate ultra-high-resolution images (e.g., 4096×4096) beyond their native resolution, without requiring architectural modifications or additional training. HiWave employs a two-stage, patch-based strategy that first generates a base image using a pretrained diffusion model and subsequently refines individual patches of the upscaled image to achieve higher resolutions. Specifically, HiWave leverages a patch-wise DDIM inversion approach to recover the latent noise from the base image and incorporates a novel wavelet-based detail enhancement module that selectively guides high-frequency components while preserving or enhancing the global structure via low-frequency components. With this frequency-aware guidance mechanism, HiWave overcomes common pitfalls of existing high-resolution methods—namely, object duplication and structural incoherence—enabling models trained at 1024×1024 to generate coherent and detailed outputs at 4096×4096. Extensive experiments with Stable Diffusion XL demonstrated that HiWave not only produces visually compelling 4K images with fine details and strong global consistency, but also significantly outperforms prior zero-shot high-resolution approaches. A user study further supported these findings, with participants preferring HiWave's results in more than 80% of cases. By enabling ultra-high-resolution synthesis without retraining, HiWave unlocks practical applications in domains where 4K (and beyond) outputs are essential. Future work could explore extensions to video, runtime optimizations, and the integration of more advanced guidance strategies within the detail enhancer module.Overall, HiWave represents a significant step toward democratizing high-fidelity, high-resolution generative modeling.

References

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *CoRR* abs/2211.01324 (2022). doi:10.48550/arXiv.2211.01324 arXiv:2211.01324

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023a. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *CoRR* abs/2311.15127 (2023). doi:10.48550/ARXIV.2311.15127 arXiv:2311.15127

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 22563–22575.

M. E. Brewster. 1993. An Introduction to Wavelets (Charles K. Chui). *SIAM Rev.* 35, 2 (1993), 312–313. doi:10.1137/1035061

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* OpenReview.net. https://openreview.net/forum?id=B1xsqj09Fm

Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. 2024. PIXART-δ: Fast and Controllable Image Generation with Latent Consistency Models. *ArXiv* abs/2401.05252 (2024). https://api.semanticscholar.org/CorpusID:266902626

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. 2023. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426* (2023).

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. 2021. WaveGrad: Estimating Gradients for Waveform Generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.* OpenReview.net. https://openreview.net/forum?id=NsMLjcFaO8O

Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems* 28 (2015).

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 8780–8794. https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning.*

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. 2023. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662* (2023).

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

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *CoRR* abs/2207.12598 (2022). doi:10.48550/arXiv.2207.12598 arXiv:2207.12598

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

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Havnø Frank, Jesse H. Engel, Quoc V. Le, William Chan, and Wei Han. 2023. Noise2Music: Text-conditioned Music Generation with Diffusion Models. *CoRR* abs/2302.03917 (2023). doi:10.48550/arXiv.2302.03917 arXiv:2302.03917

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. (2022). https://openreview.net/forum?id=k7FuTOWMOc7

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. 2023. Analyzing and Improving the Training Dynamics of Diffusion Models. arXiv:2312.02696 [cs.CV]

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.* OpenReview.net. https://openreview.net/forum?id=a-xFK8Ymz5J

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. 2025. EQ-VAE: Equivariance Regularized Latent Space for Improved Generative Image Modeling. *arXiv preprint arXiv:2502.09509* (2025).

Zhihang Lin, Mingbao Lin, Meng Zhao, and Rongrong Ji. 2024. Accdiffusion: An accurate method for higher-resolution image generation. In *European Conference on Computer Vision*. Springer, 38–53.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. 2023. I^2SB: Image-to-Image Schrödinger Bridge. *CoRR* abs/2302.05872 (2023). doi:10.48550/arXiv.2302.05872 arXiv:2302.05872

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8162–8171. http://proceedings.mlr.press/v139/nichol21a.html

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023a. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023b. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CoRR* abs/2307.01952 (2023). doi:10.48550/ARXIV.2307.01952 arXiv:2307.01952

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR* abs/2204.06125 (2022). doi:10.48550/arXiv.2204.06125 arXiv:2204.06125

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. doi:10.1109/CVPR52688.2022.01042

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. 2024. CADS: Unleashing the Diversity of Diffusion Models through Condition-Annealed Sampling. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=zMoNrajk2X

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. 2025. Eliminating Oversaturation and Artifacts of High Guidance Scales in Diffusion Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=e2ONKX6qzJ

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. 2022a. Palette: Image-to-Image Diffusion Models. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, Munkhtsetseg Nandigjav, Niloy J. Mitra, and Aaron Hertzmann (Eds.). ACM, 15:1–15:10. doi:10.1145/3528233.3530757

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. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=M3Y74vmsMcY

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=St1giarCHLP

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. (2023). https://openreview.net/pdf?id=SJ1kSyO2jwu

Athanasios Tragakis, Marco Aversa, Chaitanya Kaul, Roderick Murray-Smith, and Daniele Faccio. 2024. Is One GPU Enough? Pushing Image Generation at Higher-Resolutions with Foundation Models. *arXiv preprint arXiv:2406.07251* (2024).

Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. 2023. EDGE: Editable Dance Generation From Music. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 448–458. doi:10.1109/CVPR52729.2023.00051

Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. 2025. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 3944–3953.

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

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=AFDcYJKhND

Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Zhenyuan Chen, Yao Tang, Yuhao Chen, Wengang Cao, and Jiajun Liang. 2023. Hidiffusion: Unlocking high-resolution creativity and efficiency in low-resolution trained diffusion models. *CoRR* (2023).