AccDiffusion: An Accurate Method for Higher-Resolution Image Generation

Zhihang Lin[1], Mingbao Lin[2], Meng Zhao[3], and Rongrong Ji[1⋆]

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China.
[2] Skywork AI.
[3] Tencent Youtu Lab.
zhihanglin@stu.xmu.edu.cn, linmb001@outlook.com, arthurrizar8421@gmail.com, rrji@xmu.edu.cn

Abstract. This paper attempts to address the object repetition issue in patch-wise higher-resolution image generation. We propose AccDiffusion, an accurate method for patch-wise higher-resolution image generation without training. An in-depth analysis in this paper reveals an identical text prompt for different patches causes repeated object generation, while no prompt compromises the image details. Therefore, our AccDiffusion, for the first time, proposes to decouple the vanilla image-content-aware prompt into a set of patch-content-aware prompts, each of which serves as a more precise description of an image patch. Besides, AccDiffusion also introduces dilated sampling with window interaction for better global consistency in higher-resolution image generation. Experimental comparison with existing methods demonstrates that our AccDiffusion effectively addresses the issue of repeated object generation and leads to better performance in higher-resolution image generation. Our code is released at `https://github.com/lzhxmu/AccDiffusion`.

Keywords: Image Generation · High Resolution · Diffusion Model

1 Introduction

Diffusion models have garnered significant attention and made notable advancements with the emergence of works such as DDPM [10], DDIM [28], ADM [3], and LDMs [21], owing to their outstanding generative ability and wide range of applications. However, stable diffusion models entail tremendous training costs primarily due to the large number of timestamps required and the quadratic relationship between computing costs and resolution. Consequently, it is common to limit the resolution to a relatively low level, such as 512^2 for SD 1.5 [20] and 1024^2 for SDXL [17], during training. Even at such low resolution, stable diffusion 1.5 still entails over 20 days of training on 256 A100 GPUs [20]. Nonetheless, high-resolution generation finds widespread application in real-life scenarios, such as advertisements. The demand for generating high-resolution images clashes with the expensive training costs involved.

⋆ Corresponding author

Fig. 1: Comparison of image quality and GPU overhead for existing higher-resolution generation methods. The GPU memory of Attn-SF [12] and ScaleCrafter [6] significantly increases with resolution, while patch-wise denoising methods ,*e.g.*, MultiDiffusion [1] and DemoFusion [4] suffer object repetition issue. Best viewed zoomed in.

Therefore, researchers have shifted their focus to training stable diffusion models with low resolution and subsequently applying fine-tuning [30, 33] or training-free [1, 4, 6, 14] methods to achieve image generation extrapolation. A naive approach is to directly use pre-trained stable diffusion models to generate higher-resolution images. However, the resulting images from this approach are proved to suffer from issues such as object repetition and inconsistent object structures [4, 12]. Previous methods attempted to achieve image generation extrapolation from the perspectives of attention entropy [12] or the receptive field of stable diffusion model [6]. However, these methods have been proven to be less practical in two folds, as shown in Fig. 1(b,c): (1)a substantial increase in GPU memory consumption [33] as the resolution rises and (2) poor quality of the generated images [4]. Thanks to stable diffusion's outstanding local detail generation ability, recent works [1, 4, 14] have started conducting higher-resolution image generation in a patch-wise fashion for the sake of less GPU memory consumption. Previous works MultiDiffusion [1] and SyncDiffusion [14] fuse multiple overlapped patch-wise denoising results to generate higher-resolution panoramic images without a seam. However, the direct application of these approaches to generate higher-resolution object-centric images leads to repeated and distorted results lacking global semantic coherence, as shown in Fig. 1(d). Recently,

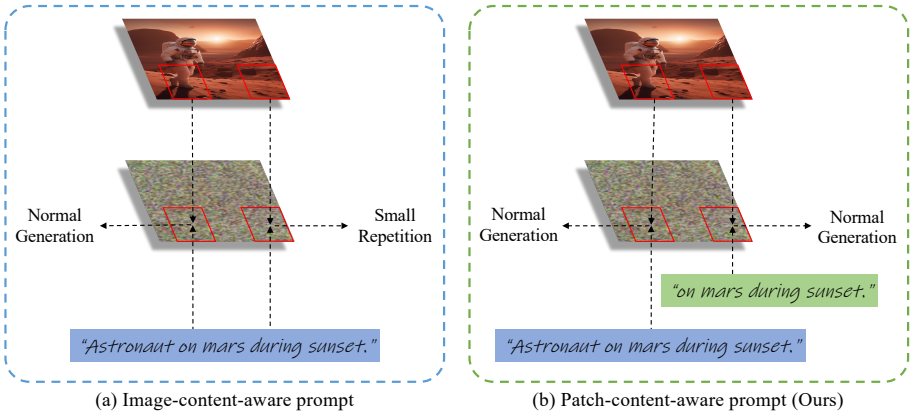

Fig. 2: Image-content-aware prompt *v.s.* Patch-content-aware prompt.

DemoFusion [4] has introduced global semantic information into the patch-wise higher-resolution image generation through residual connection and dilated sampling. It only partially solves the problem of repeated object generation and still exhibits small object repetition in ultra-high image generation as depicted in Fig. 1(e). How to resolve the issue of repeated object generation completely in patch-wise higher-resolution image generation remains an unresolved problem.

In this paper, our in-depth analysis of DemoFusion [4] indicates, as illustrated in Fig. 2(a), small object repetition generation is the adversarial outcome of an identical text prompt on all patches, encouraging to generate repeated objects, and global semantic information from residual connection and dilated sampling, suppressing the generation of repeated objects. To address the above issues, we propose AccDiffusion, an accurate method for higher-resolution image generation, with its major novelty in two folds:

(1) To completely solve small object repetition, as illustrated in Fig. 2(b), we propose to decouple the vanilla image-content-aware prompt into a set of patch-content-aware substrings, each of which serves as a more precise prompt to describe the patch contents. Specifically, we utilize the cross-attention map from the low-resolution generation process to determine whether a word token should serve as the prompt for a patch. If a word token has a high response in the cross-attention map region corresponding to the patch, it should be included in the prompt, and vice versa.

(2) Through visualization, we observe that the dilated sampling operation in DemoFusion generates globally inconsistent and noisy information, disrupting the generation of higher-resolution images. Such inconsistency stems from the independent denoising of dilation samples without interaction. To address this, we employ a position-wise bijection function to enable interaction between the noise from different dilation samples. Experimental results show that our dilated sampling with interaction leads to the generation of smoother global semantic information (see Fig. 3(c,d)).

We have conducted extensive experiments to verify the effectiveness of AccDiffusion. The qualitative results demonstrate that AccDiffusion effectively addresses the issue of repeated object generation in higher-resolution image generation. And the quantitative results show that AccDiffusion achieves state-of-the-art performance in training-free image generation extrapolation.

2 Related Work

2.1 Diffusion Models

Diffusion models [3,10,21,28] are generative probabilistic models that transform Gaussian noise into samples through gradual denoising steps. DDPM [10] is a pioneering model that demonstrates impressive image generation capabilities using Markovian forward and reverse processes. Based on DDPM, DDIM [28] utilizes non-Markovian reverse processes to decrease sampling time effectively. Furthermore, LDMs [21] incorporate the diffusion process into the latent space, resulting in efficient training and inference. Subsequently, a series of LDMs-based stable diffusion models are open-sourced and achieve state-of-the-art image synthesis capability. This has led to widespread applications in various downstream generative tasks, including images [3, 10, 16, 22, 28], audio [5, 11], video [9, 26] and 3D objects [15, 18, 31], *etc.*

2.2 Training-Free Higher-Resolution Image Generation

Although stable diffusion demonstrates impressive results, its training cost limits low-resolution training and thus generates low-fidelity images when the inference resolution differs from the training resolution [4, 6, 12]. Recent works [1, 4, 6, 12] have attempted to utilize pre-trained diffusion models for generating higher-resolution images. These works [1, 4, 6, 12] can be broadly categorized into two categories: direct generation [6, 12] and indirect generation [1, 4]. Direct generation methods scale the input of the diffusion models to the target resolution and then perform forward and reverse processes directly on the target resolution. These kinds of methods require modifications to the fundamental architecture, such as adjusting the attention scale factor [12] and the receptive field of convolutional kernels [6], to prevent repetition generation. However, the generated images fail to yield the higher-resolution detail desired. Additionally, direct generation methods encounter out-of-memory errors when generating ultra-high resolution images (*e.g.* 8K) on consumer-grade GPUs, due to the quadratic increase in memory overhead as the latent space size grows. Indirect generation methods generate higher-resolution images through multiple overlapped denoising paths of LDMs and are capable of generating images of any resolution on consumer-grade GPUs. However, these mothods [1, 14] suffer from local repetition and structural distortion. Du *et al.* [4] tried to address repeated generation by introducing global structural information from lower-resolution image.

3 Method

3.1 Backgrounds

Latent Diffusion Models (LDMs). LDMs [3] apply an autoencoder $\mathcal{E}$ to encode an image $\mathbf{x}_0 \in \mathbb{R}^{H\times W\times 3}$ into a latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0) \in \mathbb{R}^{h\times w\times c}$, where the regular diffusion process is constructed as:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\{\alpha_t\}_{t=1}^T$ is a set of prescribed variance schedules and $\bar{\alpha}_t = \Pi_{i=1}^t \alpha_i$. To perform conditional sequential denoising, a network ε_θ is trained to predict added noise, constrained by the following training objective:

$$\min_\theta \mathbb{E}_{\mathcal{E}(x_0),\varepsilon\sim\mathcal{N}(0,1),t\sim\text{Uniform}(1,T)}\Big[\big\|\varepsilon - \varepsilon_\theta\big(\mathbf{z}_t, t, \tau_\theta(y)\big)\big\|_2^2\Big], \tag{2}$$

in which $\tau_\theta(y) \in \mathbb{R}^{M\times d_\tau}$ is an intermediate representation of condition y and M is the number of word tokens in the prompt y. The $\tau_\theta(y)$ is then mapped to keys and values in cross-attention of U-Net ε_θ:

$$\begin{aligned} &Q = W_Q \cdot \varphi(z_t), \quad K = W_K \cdot \tau_\theta(y), \quad V = W_V \cdot \tau_\theta(y), \\ &\mathcal{M} = \text{Softmax}(\frac{QK^T}{\sqrt{d}}), \quad \text{Attention}(Q, K, V) = \mathcal{M} \cdot V. \end{aligned} \tag{3}$$

Here, for simplicity, we omit the expression of multi-head cross-attention and $\varphi(z_t) \in \mathbb{R}^{N\times d_\epsilon}$ denotes an intermediate representation of noise in the U-Net. Here $N = h \times w$ represents the pixel number of the latent noise z_t. $W_Q \in \mathbb{R}^{d\times d_\epsilon}, W_K \in \mathbb{R}^{d\times d_\tau}$, and $W_V \in \mathbb{R}^{d\times d_\tau}$ are learnable projection matrices. $\mathcal{M} \in \mathbb{R}^{N\times M}$ is the cross-attention maps.

In contrast, the denoising process aims to recover the cleaner version $\mathbf{z}_{t-1}$ from $\mathbf{z}_t$ by estimating the noise, which can be expressed as:

$$\mathbf{z}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\mathbf{z}_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta\big(\mathbf{z}_t, t, \tau_\theta(y)\big). \tag{4}$$

During inference, a decoder $\mathcal{D}$ is employed at the end of the denoising process to reconstruct the image from the latent representation $\mathbf{x}_0 = \mathcal{D}(\mathbf{z}_0)$.

Patch-wise Denoising. MultiDiffusion [1] achieve higher-resolution image generation by fusing multiple overlapped denoising patches. In simple terms, given a latent representation $\mathcal{Z}_t \in \mathbb{R}^{h'\times w'\times c}$ of higher-resolution image with $h' > h$ and $w' > h$, MultiDiffusion utilizes a shifted window to sample patches from $\mathcal{Z}_t$ and results in a series of patch noise $\{\mathbf{z}_t^i\}_{i=1}^{P_1}$, where $\mathbf{z}_t^i \in \mathbb{R}^{h\times w\times c}$ and $P_1 = (\frac{h'-h}{d_h} + 1) \times (\frac{w'-w}{d_w} + 1)$ is the total number of patches, d_h and d_w is the vertical and horizontal stride, respectively. Then, MultiDiffusion performs patch-wise denoising via Eq. (4) and obtains $\{\mathbf{z}_{t-1}^i\}_{i=1}^{P_1}$. Then $\{\mathbf{z}_{t-1}^i\}_{i=1}^{P_1}$ is reconstructed to get $\mathcal{Z}_{t-1}$, where the overlapped parts take the average. Eventually, a higher-resolution image can be obtained by directly decoding $\mathcal{Z}_0$ into image $\mathbf{X}_0$.

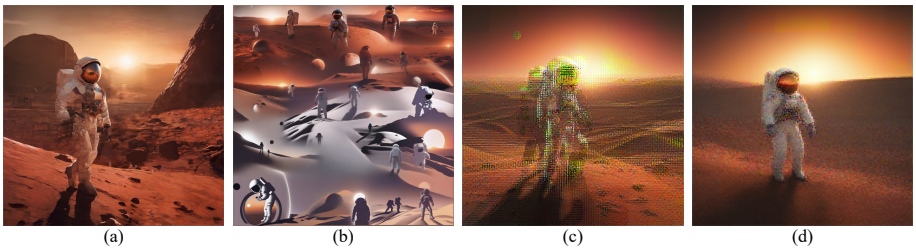

Fig. 3: Results of higher-resolution image generation. (a) The result of DemoFusion without text prompt. (b)The result of DemoFusion without residual connection and dilated sampling. (c) The result of dilated sampling without window interaction. (d)The result of our dilated sampling with window interaction.

Based on MultiDiffusion, DemoFusion [4] additionally introduces: 1) progressive upscaling to gradually generate higher-resolution images; 2) residual connection to maintain global consistency with the lower-resolution image by injecting the intermediate noise-inversed representation. 3) dilated sampling to enhance global semantic information of higher-resolution images.

3.2 In-depth Analysis of Small Object Repetition

DemoFusion demonstrates the possibility of using pre-trained LDMs to generate higher-resolution images. However, as shown in Fig. 1(e), small object repetition continues to challenge the performance of DemoFusion.

Delving into an in-depth analysis, we respectively: 1) remove the text prompt during higher-resolution generation of DemoFusion and the resulting Fig. 3(a) indicates the disappearance of repeated objects but more degradation in details. 2) remove the operations of residual connection & dilated sampling in DemoFusion and the resulting Fig. 3(b) denotes severe large object repetition. Therefore, we can make a safe conclusion that small object repetition is the adversarial outcome of an identical text prompt on all patches and operations of residual connection & dilated sampling. The former encourages to generate repeated objects while the latter suppresses the generation of repeated objects. Consequently, DemoFusion tends to generate small repeated objects.

Overall, text prompts play a significant role in image generation. It is not a viable solution to address small object repetition by removing text prompts during the higher-resolution generation, as it would lead to a decline in image quality. Instead, we require more accurate prompts specifically tailored for each patch. That is, if an object is not present in a patch, the corresponding word in the text prompts should not serve as a prompt for that patch.

To this end, in Sec. 3.3, we eliminate the restriction of having an identical text prompt for all patches in previous patch-wise generation approaches. Instead, we generate more precise patch-content-aware prompts that adapt to the content of different patches. In Sec. 3.4, we introduce how to enhance the global structure information to generate higher-resolution images without repetition.

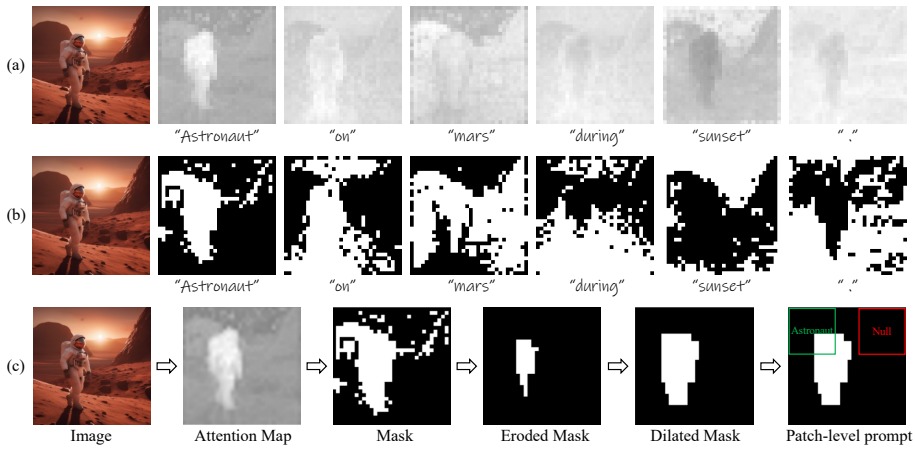

Fig. 4: Visualization of averaged attention map from the up blocks and down blocks in U-Net. We reshape the attention map into a 2D shape before visualization. (a) Cross-attention map visualization using open source code [7]. (b) Highly responsive regions of each word. (c) The illustration of the patch-level prompt generation process, including morphological operations to eliminate small connected areas. Here we use the word "Astronaut" as an example. All words in the prompt will go through the above process.

3.3 Patch-Content-Aware Prompts

Considering the significance of text prompt in higher-resolution generation, we explore patch-content-aware substring set $\{\gamma^i\}_{i=1}^{P_1}$ of the entire text prompt, each of which is responsible for injecting a condition to the corresponding patch. In general, it is challenging to know in advance what content a patch generates, but in DemoFusion [4], the global information from low-resolution image is injected into the high-resolution image generation through residual connections. Therefore, the structure of the generated higher-resolution image is similar to that of the low-resolution image. This inspires us to decide patch contents from the low-resolution image. A direct but cumbersome approach is to manually observe the patch content of low-resolution image and then set the prompt for each patch, which undermines the usability of stable diffusion. Another approach is to use SAM [13] to segment the upscaled low-resolution image and determine whether each object appears in the patch, introducing huge storage and computational costs of the segmentation model. How to automatically generate patch-content-aware prompts without external models is the key to success.

Inspired by image editing [7], instead we consider the cross-attention maps in low-resolution generation $\mathcal{M} \in \mathbb{R}^{N \times M}$, to determine patch-content-aware prompts. Recall N represents the pixel number of the latent noise z_t and M denotes the number of word tokens in the prompt y. Thus, the column $\mathcal{M}_{:,j}$ represents the attentiveness of latent noise to the j-th word token. The basic principle lies in that the attentiveness ($\mathcal{M}_{i,j}$) of image regions is mostly higher than others if it is attended by the j-th word token, as shown in Fig. 4(a). To

find the highly relevant region of each word token, we convert the attention map $\mathcal{M}$ into a binary mask $\mathcal{B} \in \mathbb{R}^{N \times M}$ as:

$$\mathcal{B}_{i,j} = \begin{cases} 1, & \text{if } \mathcal{M}_{i,j} > \overline{\mathcal{M}}_{:,j}, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where i and j enumerate N and M, respectively. The threshold $\overline{\mathcal{M}}_{:,j}$ is the mean of $\mathcal{M}_{:,j}$, which design is elaborated in Sec. 4.4. Regions with values above the threshold are considered highly responsive, while regions with values below the threshold are considered less responsive.

Next, we obtain word-level masks $\{\mathcal{B}_j\}_{j=1}^{M}$ using the following equation:

$$\hat{\mathcal{B}}_j = \text{Reshape}(\mathcal{B}_{:,j}, (h_a, w_a)), \tag{6}$$

where $h_a = \frac{h}{s}$ and $w_a = \frac{w}{s}$ represent the height and width of the attention map, respectively. Recall h and w represent the height and width of the noise, respectively. The "s" corresponds to the down-sampling scale in the corresponding block of the U-Net model. The mask $\mathcal{B}_j$ oriented for the j-th word token is reshaped into a 2d shape for further processing.

After obtaining the highly responsive regions for each word, we observe that they contain many small connected areas, as shown in Fig. 4(b). To alleviate the influence of these small connected areas, we apply the opening operation $\mathcal{O}(\cdot)$ from mathematical morphology [27], resulting in the final mask for each word, as shown in Fig. 4(c). The processed mask $\{\tilde{\mathcal{B}}_j\}_{j=1}^{M}$ can be formulated as:

$$\tilde{\mathcal{B}}_j = \mathcal{O}(\hat{\mathcal{B}}_j) = \omega(\delta(\hat{\mathcal{B}}_j)), \tag{7}$$

where $\delta(\cdot)$ and $\omega(\cdot)$ is erosion operation and dilation operation, respectively. Next, we interpolate $\tilde{\mathcal{B}}_j \in \mathbb{R}^{h_a \times w_a}$ to $\tilde{\mathcal{B}}'_j \in \mathbb{R}^{h'_a \times w'_a}$, where $h'_a = \frac{h'}{s}$ and $w'_a = \frac{w'}{s}$. Recall h' and w' are the size of higher-resolution latent representation as defined in Sec. 3.1. Inspired by MultiDiffusion [1], we use a shifted window to sample patches from $\tilde{\mathcal{B}}'_j$, resulting in a series of patch masks $\{\{\mathbf{m}_j^i\}_{i=1}^{P_1}\}_{j=1}^{M}$, where $\mathbf{m}_j^i \in \mathbb{R}^{h_a \times w_a}$ and P_1 is the total number of patches. It is important to note that each $\mathbf{m}_i^j$ corresponds to a specific patch noise $\mathbf{z}_t^i$.

Recall if an object is not present in a patch, the corresponding word token in the text prompts should not serve as a prompt for that patch. So, we can determine the patch-content-aware prompt γ^i, a sub-sequence of prompt y, for each patch $\mathbf{z}_t^i$ using the following formulation:

$$\begin{cases} y_j \in \gamma^i, & \text{if } \frac{\sum(\mathbf{m}_j^i)_{:,:}}{h_a \times w_a} > c, \\ y_j \notin \gamma^i, & \text{otherwise,} \end{cases} \tag{8}$$

where j and i enumerates M and P_1, respectively. The pre-given hyper-parameter $c \in (0, 1)$ determines whether a highly responsive region's proportion of a word y_j exceeds the threshold for inclusion in the prompts of patch z_t^i. We then concatenate all words that should appear in a patch together, resulting in patch-content-aware prompts $\{\gamma^i\}_{i=1}^{P_1}$ for noise patches $\{z_t^i\}_{i=1}^{P_1}$ during patch-wise denoising.

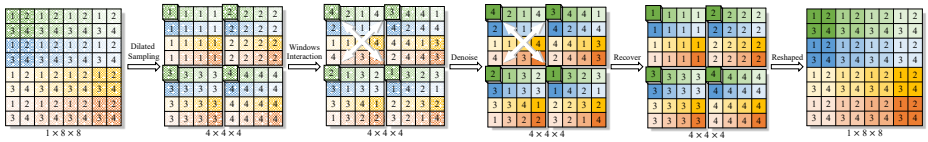

Fig. 5: Illustration of dilated sampling with window interaction: 8×8 higher-resolution and 4×4 low-resolution. The number $\{1, 2, 3, 4\}$ represent the different positions within the same window (same color). The interaction operation is conducted in the window.

3.4 Dilated Sampling with Window Interaction

Recall $\mathcal{Z}_t \in \mathbb{R}^{h'\times w'\times c}$ stands for the latent representation of a higher-resolution image in Sec. 3.1. In this section, we continue proposing dilated sampling with window interaction, for a set of patch samples $\{D_t^k\}_{k=1}^{P_2}$, to improve the global semantic information in the latent representation $\mathcal{Z}_t$. In DemoFusion [4], each sample D_t^k is a subset of the latent representation $\mathcal{Z}_t$, formulated as:

$$\mathcal{D}_t^k = (\mathcal{Z}_t)_{i::h_s,j::w_s,:}, \tag{9}$$

where $k = i\times w_s + j + 1$, and k ranges from 1 to P_2. The variables i and j range from 0 to $h_s - 1$ and $w_s - 1$, respectively. The sampling stride is determined by $h_s = \frac{h'}{h}$ and $w_s = \frac{w'}{w}$. Recall $\{h', w'\}$ and $\{h, w\}$ are the height and width of higher and low resolution latent representation. DemoFusion independently performs denoising on $\mathcal{D}_t$ via Eq. (4) and obtains $\mathcal{D}_{t-1} \in \mathbb{R}^{P_2\times h\times w\times c}$, where $P_2 = h_s \times w_s$. Then $\{\mathcal{D}_{t-1}^k\}_{k=1}^{P_2}$ is reconstructed as $G_{t-1} \in \mathbb{R}^{h'\times w'\times c}$ and added to patch-wise denoised latent representation $\mathcal{Z}_{t-1}$ using:

$$\hat{\mathcal{Z}}_{t-1} = (1-\eta)\cdot\mathcal{Z}_{t-1} + \eta\cdot G_{t-1}, \tag{10}$$

where $(G_{t-1})_{i::h_s,j::w_s,:} = \mathcal{D}_{t-1}^k$ and η decreases from 1 to 0 using a cosine schedule. Due to the lack of interaction between different samples during the denoising process, the global semantic information is non-smooth, as depicted in Fig. 3(c). The sharp global semantic information disturbs the higher-resolution generation.

To solve above issue, as illustrated in Fig. 5, we enable window interaction between different samples before each denoising process through bijective function:

$${\mathcal{D}_t}^{k,h,w} = {\mathcal{D}_t}^{f_t^{h,w}(k),h,w}, \quad f_t^{h,w} : \{1, 2, \cdots, P_2\} \Rightarrow \{1, 2, \cdots, P_2\}, \tag{11}$$

where $f_t^{h,w}$ is bijective function, and the mapping varies based on the position or time step. We then perform normal denoising progress on $\{\mathcal{D}_t^k\}_{k=1}^{P_2}$ to obtain $\{\mathcal{D}_{t-1}^k\}_{k=1}^{P_2}$. Before applying Eq. (10) to $\{\mathcal{D}_{t-1}^k\}_{k=1}^{P_2}$, we use the inverse mapping $(f_t^{h,w})^{-1}$ of $f_t^{h,w}$ to recover the position as:

$${\mathcal{D}_{t-1}}^{k,h,w} = {\mathcal{D}_{t-1}}^{(f_t^{h,w})^{-1}(k),h,w}, \quad (f_t^{h,w})^{-1} : \{1, 2, \cdots, P_2\} \Rightarrow \{1, 2, \cdots, P_2\}. \tag{12}$$

4 Experimentation

4.1 Experimental Setup

AccDiffusion is a plug-and-play extension to stable diffusion without additional training costs. We mainly validate the feasibility of AccDiffusion using the pre-trained SDXL [17]. More results for other stable diffusion variants are in the supplementary material. AccDiffusion follows the pipeline of DemoFusion [4] (SOTA higher-resolution generation) and uses the patch-content-aware prompts during the progress of higher-resolution image generation. Additionally, AccDiffusion enhances dilated sampling with window interaction. For fairness, we adhere to the default setting of DemoFusion, as described in the supplementary material. In Sec. 4.2, the hyper-parameter c in Eq. (8) is set to 0.3. Considering the training-free nature of AccDiffusion, the methods we compare include: **SDXL-DI** [17], **Attn-SF** [12], **ScaleCrafter** [6], **MultiDiffusion** [1], and **DemoFusion** [4]. We do not compare with image super-resolution methods [23, 29, 32] which take images as input and have been proven to lack texture details [4, 6].

4.2 Quantitative Comparison

For quantitative comparison, we use three widely-recognized metrics: FID (Frechet Inception Distance) [8], IS (Inception Score) [24], and CLIP Score [19]. Specifically, FID_r measures the Frechet Inception Distance between generated high-resolution images and real images. IS_r represents the Inception Score of generated high-resolution images. Given that FID_r and IS_r necessitate resizing images to 299^2, which may not be ideal for assessing high-resolution images. Motivated by [2, 4], we crop 10 local patches at $1\times$ resolution from each generated high-resolution image and subsequently resize them to calculate FID_c and IS_c. The CLIP score measures the cosine similarity between image embedding and text prompts. We randomly selected 10,000 images from the Laion-5B [25] dataset as our real images set and randomly chose 1,000 text prompts from Laion-5B as inputs for AccDiffusion to generate a set of high-resolution images.

As shown in Table 1, AccDiffusion achieves the best results and obtains state-of-the-art performance. Since the implementation of AccDiffusion is based on DemoFusion [4], it exhibits similar quantitative results and inference times with DemoFusion. However, AccDiffusion outperforms DemoFusion due to its more precise patch-content-aware prompt and more accurate global information introduced by dilated sampling with interaction, especially in high-resolution generation scenarios. Compared to other training-free image generation extrapolation methods, the quantitative results of AccDiffusion are closer to quantitative results calculated at pre-trained resolutions (1024×1024), demonstrating the excellent image generation extrapolation capabilities of AccDiffusion. Note that FID, IS, and CLIP-Score do not intuitively reflect the degree of repeated generation in the resulting images, so we conduct a qualitative comparison to validate the effectiveness of AccDiffusion in eliminating repeated generation.

Table 1: Comparison of quantitative metrics between different training-free image generation extrapolation methods. We use **bold** to emphasize the best result and underline to emphasize the second best result.

Resolusion	Method	$FID_r \downarrow$	$IS_r \uparrow$	$FID_c \downarrow$	$IS_c \uparrow$	CLIP↑	Time
1024 × 1024 (1×)	SDXL-DI	58.49	17.39	58.08	25.38	33.07	<1 min
2048 × 2048 (4×)	SDXL-DI	124.40	11.05	88.33	14.64	28.11	1 min
	Attn-SF	124.15	11.15	88.59	14.81	28.12	1 min
	MultiDiffusion	81.46	12.43	44.80	20.99	31.82	2 min
	ScaleCrafter	99.47	12.52	74.64	15.42	28.82	1 min
	DemoFusion	60.46	16.45	38.55	24.17	32.21	3 min
	AccDiffusion	**59.63**	**16.48**	**38.36**	**24.62**	**32.79**	3 min
3072 × 3072 (9×)	SDXL-DI	170.61	7.83	112.51	12.59	24.53	3 min
	Attn-SF	170.62	7.93	112.46	12.52	24.56	3 min
	MultiDiffusion	101.11	8.83	51.95	17.74	29.49	6 min
	ScaleCrafter	131.42	9.62	105.79	11.91	27.22	7 min
	DemoFusion	62.43	16.41	47.45	20.42	32.25	11 min
	AccDiffusion	**61.40**	**17.02**	**46.46**	**20.77**	**32.82**	11 min
4096 × 4096 (16×)	SDXL-DI	202.93	6.13	119.54	11.32	23.06	9 min
	Attn-SF	203.08	6.26	119.68	11.66	23.10	9 min
	MultiDiffusion	131.39	6.56	61.45	13.75	26.97	10 min
	ScaleCrafter	139.18	9.35	116.90	9.85	26.50	20 min
	DemoFusion	65.97	15.67	59.94	16.60	33.21	25 min
	AccDiffusion	**63.89**	**16.05**	**58.51**	**16.72**	**33.79**	26 min

4.3 Qualitative Comparison

In Fig. 6, AccDiffusion is compared with other training-free text-to-image generation extrapolation methods, such as MultiDiffusion [1], ScaleCrafter [6], and DemoFusion [4]. We provide more results in supplementary material. MultiDiffusion can generate seamless images but suffers severe repeated and distorted generation. ScaleCrafter, while avoiding the repetition of astronauts, suffers from structural distortions as highlighted in the red box, resulting in local repetition and a lack of coherence. DemoFusion tends to generate small, repeated astronauts, with the frequency of repetition escalating with image resolution, thereby significantly degrading image quality. Conversely, AccDiffusion demonstrates superior performance in generating high-resolution images without such repetitions. As Attn-SF [12] and SDXL-DI [17] cannot alleviate the repetition issue, their qualitative results are not compared here.

4.4 Ablation Study

In this section, we first perform ablation studies on the two core modules proposed in this paper, and then discuss the settings of the threshold for the binary mask in Eq. (5) and the threshold c for deciding patch-content-aware prompt in

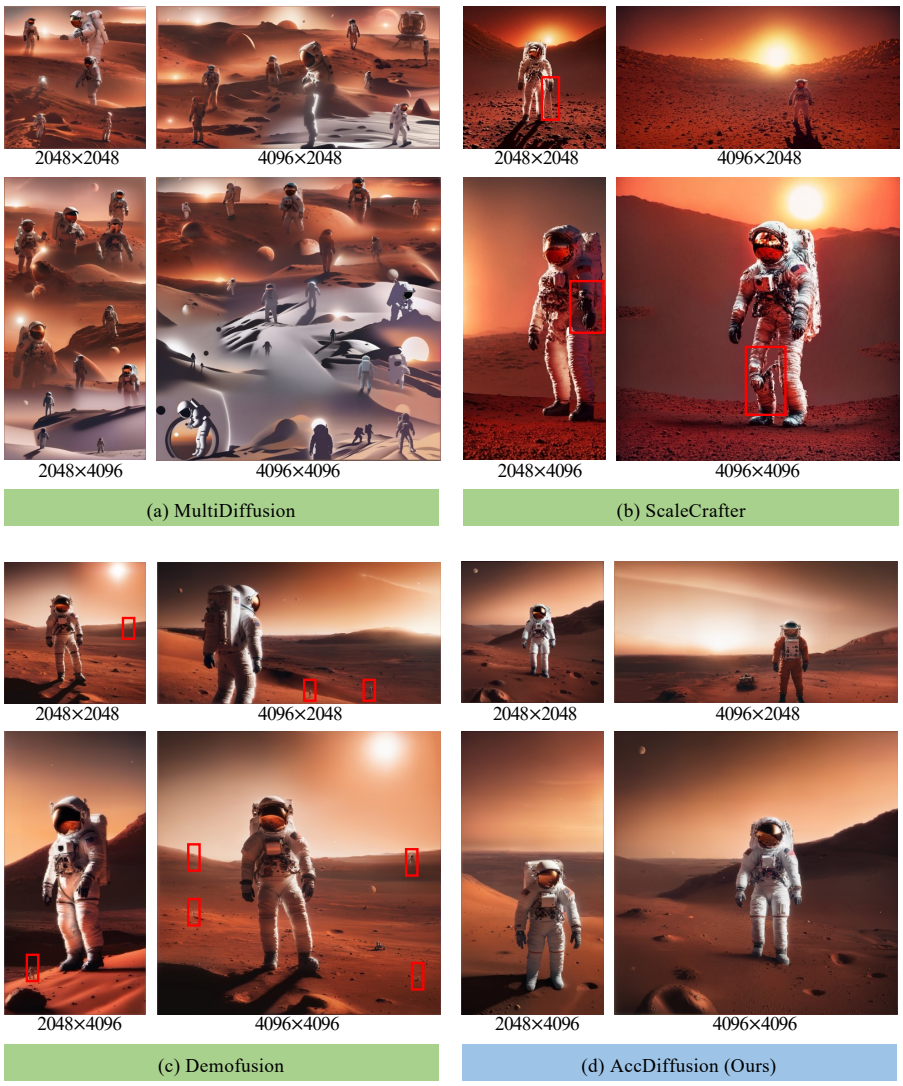

Fig. 6: Qualitative comparison of our AccDiffusion with existing training-free image generation extrapolation methods [1, 4, 6]. We draw a red box upon the generated images to highlight the repeated objects. Best viewed zoomed in.

Eq. (8). All experiments are carried out at a resolution of 4096^2 ($16\times$). Considering the fact that existing quantitative metrics are unable to accurately reflect the extent of object repetition, we choose to provide visualizations to demonstrate the effectiveness of our core modules in preventing repeated generation.

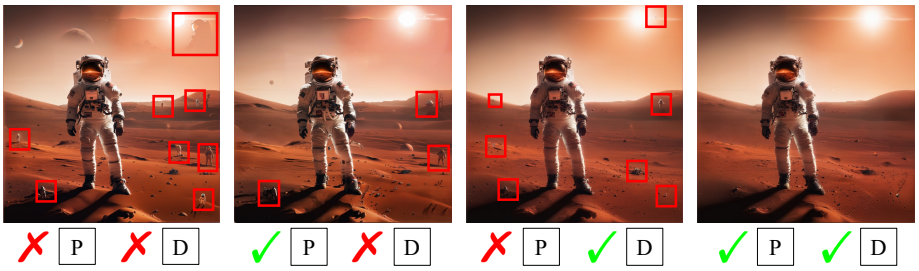

Fig. 7: Ablations of Patch-content-aware prompts (P) and Dilated sampling with window interaction (D). The "✗"/"✓" denotes removing/preserving the component. The repeated objects are highlighted by a red box. Best viewed zoomed in.

Table 2: Statistics of cross-attention maps $\mathcal{M}$ using prompt $y =$ "Astronaut on mars during sunset." as an example. Each word $\{y_j\}_{j=1}^{6}$ has a cross-attention map $\{\mathcal{M}_{:,j}\}_{j=1}^{6}$.

Statistics	"Astronaut" ($j=1$)	"on" ($j=2$)	"mars" ($j=3$)	"during" ($j=4$)	"sunset" ($j=5$)	"." ($j=6$)
Min($\mathcal{M}_{:,j}$)	0.1274	0.0597	0.2039	0.0457	0.0921	0.0335
Mean($\mathcal{M}_{:,j}$)	0.1499	0.0676	0.2533	0.0521	0.1189	0.0386
Max($\mathcal{M}_{:,j}$)	0.2096	0.0779	0.2979	0.0585	0.1499	0.0419

Ablations on Core Modules. As illustrated in Fig. 7, the absence of any module leads to a decline in generation quality. Without patch-content-aware prompts, the resulting image contains numerous repeated small objects, emphasizing the importance of patch-content-aware prompts in preventing the generation of repetitive elements. Conversely, without our window interaction in dilated sampling, the generated small object becomes unrelated to the image, indicating that dilated sampling with window interaction enhances the image's semantic consistency and suppresses repetition. The maximum number of repeated objects is produced when both modules are removed, while employing both modules simultaneously generates an image free of repetitions. This implies that the two modules work together to effectively alleviate repetitive objects.

Ablations on Hyper-Parameters. As depicted in Table 2, there is a significant variation in the range of different cross-attention maps $\mathcal{M}_j$. When using a fixed threshold for these maps, two scenarios may occur. If the threshold is too high, some words will not have highly responsive regions in their corresponding attention maps, resulting in their absence from the patch-content-aware prompt. Conversely, if the threshold is too low, the entire attention map consists of highly responsive regions, causing those words to consistently appear in the patch-content-aware prompt. By considering the average $\overline{\mathcal{M}}_{:,j}$, we can ensure that each word has suitable highly responsive regions, as demonstrated in Fig. 4(b).

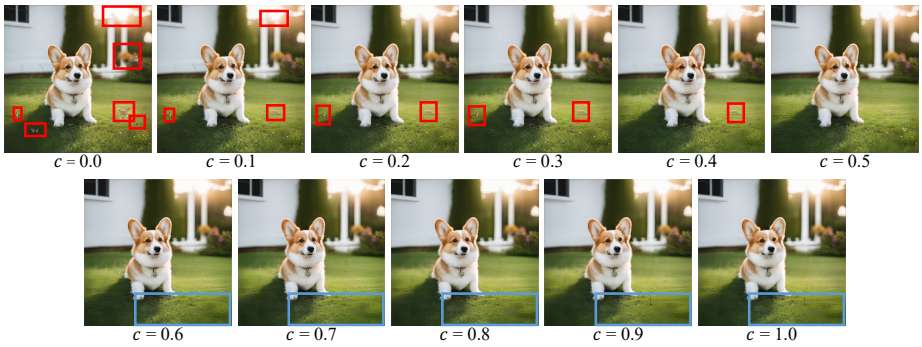

Fig. 8: Visual results of different threshold c, prompted by "A cute corgi on the lawn." The repeated objects are highlighted with a red box and the detail degradation is stressed with a blue box. Best viewed zoomed in.

Recall in Eq. (8), the c determines whether the proportion of a highly responsive region for a word y_j surpasses the threshold required for inclusion in the prompts of patch z_t^i. A very small value of c leads to more words being included in the patch prompt, potentially causing object repetition. Conversely, a very large value of c simplifies the patch prompt, which may lead to degradation of details. Our analysis is demonstrated in Fig. 8. It should be noted that this is a user-specific hyper-parameter, adjustable to suit different application scenarios.

5 Limitations and Future work

AccDiffusion is limited in: (1) As it follows the DemoFusion pipeline, similar drawbacks arise such as inference latency from progressive upscaling and overlapped patch-wise denoising. (2) AccDiffusion focuses on image generation extrapolation, meaning the fidelity of high-resolution images depends on the pretrained diffusion model. (3) Relying on LDMs' prior knowledge of cropped images, it may produce local irrational content in sharp close-up image generation.

Future studies could explore the possibility of developing non-overlapped patch-wise denoising techniques for efficiently generating high-resolution images.

6 Conclusion

In this paper, we propose AccDiffusion to address the object-repeated generation issue in higher-resolution image generation without training. AccDiffusion first introduces patch-content-aware prompts, which makes the patch-wise denoising more accurate and can avoid repeated generation from the root. And then we further propose dilated sampling with window interaction to enhance the global consistency during higher-resolution image generation. Extensive experiments, including qualitative and quantitative results, show that AccDiffusion can successfully conduct higher-resolution image generation without object repetition.

Acknowledgements

This work was supported by National Science and Technology Major Project (No. 2022ZD0118202), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. U23A20383, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2022J06001).

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. In: ICML (2023)
2. Chai, L., Gharbi, M., Shechtman, E., Isola, P., Zhang, R.: Any-resolution training for high-resolution image synthesis. In: ECCV (2022)
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
4. Du, R., Chang, D., Hospedales, T., Song, Y.Z., Ma, Z.: Demofusion: Democratising high-resolution image generation with no $$$. In: CVPR (2024)
5. Ghosal, D., Majumder, N., Mehrish, A., Poria, S.: Text-to-audio generation using instruction-tuned llm and latent diffusion model. In: ACM MM (2023)
6. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In: ICLR (2024)
7. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. In: ICLR (2023)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
9. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
11. Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z.: Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In: ICML (2023)
12. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free diffusion model adaptation for variable-sized text-to-image synthesis. In: NeurIPS (2023)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023)
14. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via synchronized joint diffusions. In: NeurIPS (2023)
15. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR (2023)
16. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)

17. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
18. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: ICLR (2023)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
20. Robin Rombach, P.E.: Stable diffusion v1-5 model card, `https://huggingface.co/runwayml/stable-diffusion-v1-5`
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
22. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
23. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. TPAMI (2022)
24. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS (2016)
25. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
26. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: ICLR (2023)
27. Soille, P., et al.: Morphological image analysis: principles and applications, vol. 2. Springer (1999)
28. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
29. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015 (2023)
30. Xie, E., Yao, L., Shi, H., Liu, Z., Zhou, D., Liu, Z., Li, J., Li, Z.: Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In: ICCV (2022)
31. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In: CVPR (2023)
32. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: CVPR (2021)
33. Zheng, Q., Guo, Y., Deng, J., Han, J., Li, Y., Xu, S., Xu, H.: Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. In: AAAI (2024)