DemoFusion: Democratising High-Resolution Image Generation With No $$$

Ruoyi Du[1,4†], Dongliang Chang[2*], Timothy Hospedales[3], Yi-Zhe Song[4], Zhanyu Ma[1]
[1]PRIS, Beijing University of Posts and Telecommunications, China
[2]Tsinghua University, China [3]University of Edinburgh, UK [4]SketchX, University of Surrey, UK
{duruoyi, mazhanyu}@bupt.edu.cn, changdongliang@pris-cv.cn,
t.hospedales@ed.ac.uk, y.song@surrey.ac.uk
https://ruoyidu.github.io/demofusion/demofusion.html

Figure 1. **Selected landscape samples of DemoFusion *versus* SDXL** [24] (all images in the figure are presented at their actual sizes). SDXL can synthesize images up to a resolution of 1024^2, while DemoFusion extends SDXL to generate images at $4\times$, $16\times$, and even higher resolutions without any fine-tuning or prohibitive memory demands. All generated images are produced using a single RTX 3090 GPU. Best viewed **ZOOMED-IN**.

Abstract

High-resolution image generation with Generative Artificial Intelligence (GenAI) has immense potential but, due to the enormous capital investment required for training, it is increasingly centralised to a few large corporations, and hidden behind paywalls. This paper aims to democratise high-resolution GenAI by advancing the frontier of high-resolution generation while remaining accessible to a broad audience. We demonstrate that existing Latent Diffusion Models (LDMs) possess untapped potential for higher-resolution image generation. Our novel DemoFusion framework seamlessly extends open-source GenAI models, employing Progressive Upscaling, Skip Residual, and Dilated Sampling mechanisms to achieve higher-resolution image generation. The progressive nature of DemoFusion requires more passes, but the intermediate results can serve as "previews", facilitating rapid prompt iteration.

1. Introduction

Generating high-resolution images with Generative Artificial Intelligence (GenAI) models has demonstrated remarkable potential [1, 19, 23]. However, these capabilities are increasingly centralised. Training high-resolution image gen-

[†]The work is done while Ruoyi Du visiting the People-Centred AI Institute at the University of Surrey.
[*]Corresponding Author

eration models requires substantial capital investments in hardware, data, and energy that are beyond the reach of individual enthusiasts and academic institutions. For example, training Stable Diffusion 1.5, at a resolution of 512^2, entails over 20 days of training on 256 A100 GPUs [1]. Companies that make these investments understandably want to recoup their costs and increasingly hide the resulting models behind paywalls. This trend toward centralisation and pay-per-use access is accelerating as GenAI image synthesis advances in quality since the investment required to train image generators increases rapidly with image resolution.

In this paper we reverse this trend and re-democratise GenAI image synthesis by introducing *DemoFusion*, which pushes the frontier of high-resolution image synthesis from 1024^2 in SDXL [24], Midjourney [19], DALL-E [23], *etc* to 4096^2 or more. DemoFusion requires no additional training and runs on a single consumer-grade RTX 3090 GPU (hardware for the "working class" in the GenAI era), as shown in Fig. 1. The only trade-off? A little more patience.

Specifically, we start with the open source SDXL [24] model, capable of generating images of 1024^2. DemoFusion is a plug-and-play extension to SDXL that enables $4\times$, $16\times$, or more increase in generation resolution (Fig 1) – all with zero additional training, and only a few simple lines of code. Off-the-shelf SDXL fails if directly prompted to produce higher-resolution images (Fig. 2 (a)). However, we observe that text-to-image LDMs encounter many cropped photos during their training process. These cropped photos either exist inherently in the training set or are intentionally cropped for data augmentation. Consequently, models like SDXL occasionally produce outputs that focus on localised portions of objects [24], as illustrated in Fig. 2 (b). In other words, existing open-source LDMs already contain sufficient prior knowledge to generate high-resolution images, if only we can unlock them by fusing multiple such high-resolution patches into a complete scene.

However, achieving coherent patch-wise high-resolution generation is non-trivial. A recent study, MultiDiffusion [2] showcased the potential of fusing multiple overlapped denoising paths to generate panoramic images. Yet, when directly applying this approach to generate specific high-resolution object-centric images, results are repetitive and distorted without global semantic coherence [42], as illustrated in Fig. 2 (c). We conjecture the underlying reason is that overlapped patch denoising merely reduces the seam issue without a broad perception of the global context required for semantic coherence. DemoFusion builds upon the same idea of fusing multiple denoising paths from a pre-trained SDXL model to achieve high-resolution generation. It introduces three key mechanisms to achieve global semantic coherence together with rich local detail (Fig. 2 (d) vs (a, c)): (i) *Progressive Upscaling*: Starting with the low-resolution input, DemoFusion iteratively enhances

(a) SDXL Direct Inference (b) SDXL Patch-based Inference

(c) MultiDiffusion (d) DemoFusion (Ours)

Figure 2. **Examples of $4\times$ (2048^2) generation based on SDXL** [24]. (a) Directly prompting SDXL to generate a $4\times$ image. (b) SDXL [24] inferences on non-overlapping patches at the original resolution. It fails, but reveals that the SDXL possesses prior knowledge of localized patches at higher resolutions. (c) MultiDiffusion [2] fuses multiple overlapping denoising paths to generate higher-resolution images without edge effects, but lacks the global context for semantic coherence. (d) Our proposed DemoFusion achieves global semantic coherence in high-resolution generation.

images through an "upsample-diffuse-denoise" loop, using the noise-inversed lower-resolution image as a better initialisation for generating the higher-resolution image. (ii) *Skip Residual*: Within the same iteration, we additionally utilise the intermediate noise-inversed representations as skip residuals, maintaining global consistency between high and low-resolution images. (iii) *Dilated Sampling*: We extend MultiDiffusion to increase global semantic coherence by using dilated sampling of denoising paths. These three techniques to modify inference are simple to implement on a pre-trained SDXL and provide a dramatic boost in high-resolution image generation quality and coherence. Fig. 3 illustrates the framework.

The caveat is that generating high-resolution images does require more runtime (users need to exercise more patience). This is partially due to the progressive upscaling requiring more passes; however, primarily because the time required grows exponentially with resolution (as per any patch-wise LDM [2]), and thus, the highest resolution pass dominates the cost. Nevertheless, the memory cost is low enough for consumer-grade GPUs, and progressive generation allows the users to preview low-resolution results rapidly, facilitating rapid iteration on the prompt until satisfaction with the general layout and style, prior to wait-

ing for a full high-resolution generation.

2. Related Work

With the progress of several years, diffusion model (DM) [33] has recently reached its own "tipping point" – with the emergence of works like DDPM [34], DDIM [34], ADM [5], DM has shown great potential in image generation due to its outstanding generation quality and diversity. Subsequently, using a pre-trained autoencoder, the latent diffusion model (LDM) [29] applies a diffusion model in the latent space, achieving efficient training and inference. This enabled the emergence of high-performance generative models trained on billions of data, such as the Stable Diffusion series. LDM's excellent generalisation capability has led to subsequent research on controllable generation [21, 22, 30, 41] and editable generation [3, 9, 20]; it has also been widely applied in numerous downstream generative tasks, such as text-to-video [8, 11, 37], text-to-3D [18, 25, 38], text-to-avatar [6, 17, 36], and text-to-human sketch [14, 26, 27], *etc*.

Despite achieving numerous successes, current LDMs like Stable Diffusion 1.5 and Stable Diffusion XL are still confined to generating images at resolutions of 512^2 and 1024^2, respectively [24]. Escalating resolution significantly increases training expenses and computational load, making such models impractical for most researchers and users. An intuitive solution to generate high-resolution images involves using LDMs for initial image generation, followed by enhancement through a super-resolution (SR) model. Cascaded Diffusion Models [12] cascades several diffusion-based SR models behind a diffusion model, but its application remains capped at 256^2 resolution images. We attempted to enhance state-of-the-art LDMs with SR models [35, 40], but found that images generated at lower resolutions were deficient in detail. Upscaling these images with SR failed to yield the high-resolution detail desired. Another attempt is to retrain/fine-tune open-source DMs to achieve satisfactory results [13, 42], but fine-tuning still brings a non-negligible cost.

Recently, MultiDiffusion [2] fuses multiple overlapped denoising paths of LDMs, achieving seamless panorama generation in a training-free manner. Subsequently, SyncDiffusion [16] further constrains the consistency between denoising paths using a gradient descent approach. However, these methods are limited to generating scene images through repetition; when applied to generating specific objects, they lead to local repetition and structural distortion. Valuing the training-free characteristic of such methods, we proposed DemoFusion based on MultiDiffusion in this paper towards democratising high-resolution generation.

Note that a recent concurrent work, SCALECRAFTER [7], with the same motivation, proposed a tuning-free framework for high-resolution image generation. It ingeniously adapts the diffusion model for higher resolutions by dilating its convolution kernels at specific layers. Despite a smart move, our experiments indicate that SCALECRAFTER somewhat degrades the model's performance and does not bring about the local details expected at higher resolutions. In contrast, DemoFusion has demonstrated better results.

3. Methodology

3.1. Preliminaries

Latent Diffusion Model: Given an image $\mathbf{x}$, an LDM first encodes it to the latent space via the encoder of the pre-trained autoencoder, *i.e.*, $\mathbf{z} = \mathcal{E}(\mathbf{x}), \mathbf{z} \in \mathbb{R}^{c\times h\times w}$.

Following this, the two core components of the diffusion model, the diffusion and the denoising process, take place in the latent space. The diffusion process comprises a sequence of T steps with Gaussian noise incrementally introducing into the latent distribution at each step $t \in [0, T]$. With a prescribed variance schedule $\beta_1, \cdots, \beta_T$, the diffusion process can be formulated as

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}). \quad (1)$$

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

$$p_\theta(\mathbf{z}_{t-i}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \mathbf{\Sigma}_\theta(\mathbf{z}_t, t)), \quad (2)$$

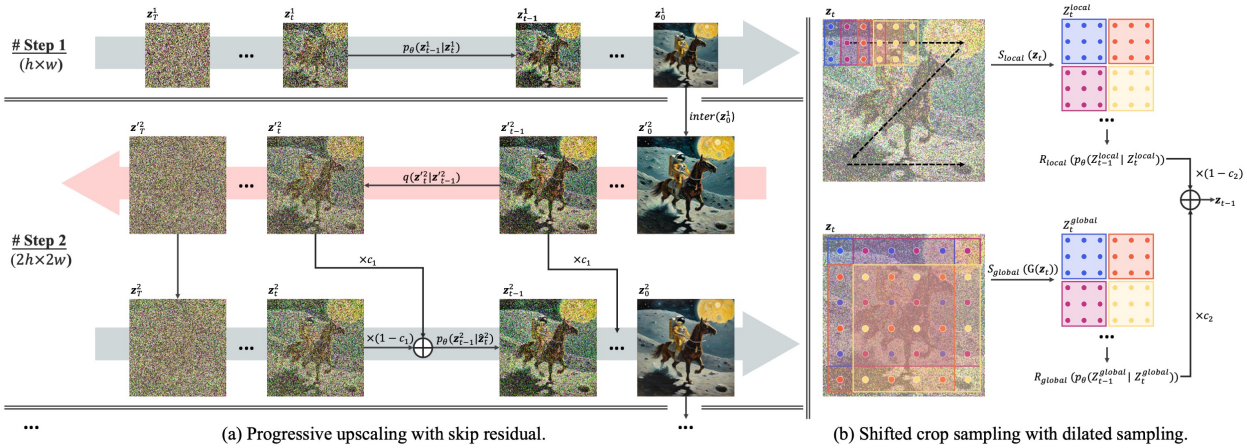
where μ_θ and $\mathbf{\Sigma}_\theta$ are determined through estimation procedures and θ denotes the parameters of the denoise model.

MultiDiffusion: MultiDiffusion [2] extends LDMs such as SDXL to produce high-resolution panoramas by overlapped patch-based denoising.

In simple terms, MultiDiffusion defines a latent space $\mathbb{R}^{c\times H\times W}$ with $H > h$ and $W > w$. For arbitrary denoising step t with $\mathbf{z}_t \in \mathbb{R}^{c\times H\times W}$, MultiDiffusion first applies a shifted crop sampling $\mathcal{S}_{local}(\cdot)$ to obtain a series of local latent representations, *i.e.*, $Z_t^{local} = [\mathbf{z}_{0,t}, \cdots, \mathbf{z}_{n,t}, \cdots, \mathbf{z}_{N,t}] = \mathcal{S}_{local}(\mathbf{z}_t)$, $\mathbf{z}_{n,t} \in \mathbb{R}^{c\times h\times w}$, where $N = (\frac{(H-h)}{d_h} + 1) \times (\frac{(W-w)}{d_w} + 1)$, d_h and d_w is the vertical and horizontal stride, respectively.

After that, the conventional denoising process is independently applied to these local latent representations via $p_\theta(\mathbf{z}_{n,t-1}|\mathbf{z}_{n,t})$. And then Z_{t-1}^{local} is reconstructed to the original size with the overlapped parts averaged as $\mathbf{z}_{t-1} = \mathcal{R}_{local}(Z_{t-1}^{local})$, where $\mathcal{R}_{local}$ denotes the reconstruction process. Eventually, a higher-resolution panoramic image can be obtained by directly decoding $\mathbf{z}_0$ into image $\hat{\mathbf{x}}$.

MultiDiffusion provides effective panorama generation, thanks to smoothing the edge effects between generated patches. However, as discussed by [42], and illustrated in Fig. 2, it struggles with generating coherent semantic content for specific objects. The fundamental reason for this



(a) Progressive upscaling with skip residual. (b) Shifted crop sampling with dilated sampling.

Figure 3. **The proposed DemoFusion framework.** (a) Starting with conventional resolution generation, DemoFusion engages an "upsample-diffuse-denoise" loop, taking the low-resolution generated results as the initialization for the higher resolution through noise inversion. Within the "upsample-diffuse-denoise" loop, a noise-inverted representation from the corresponding time-step in the preceding diffusion process serves as skip-residual as global guidance. (b) To improve the local denoising paths of MultiDiffusion, we introduce dilated sampling to establish global denoising paths, promoting more globally coherent content generation.

is that each patch/diffusion path is constrained only by the text condition and lacks awareness of the global context of the other patches.

We introduce three modifications to the inference procedure of SDXL that enable a patch-wise high-resolution image generation strategy to achieve both global semantic coherence and rich local details. These are: *Progressive Upscaling* (see Sec. 3.2), *Skip Residual* (see Sec. 3.3) and *Dilated Sampling* (see Sec. 3.4). The overall flow of DemoFusion is summarised in Appendix A.

3.2. Progressive Upscaling

Progressively generating images from low to high resolution is a well-established concept [15]. By initially synthesizing a semantically coherent overall structure at low resolution, and subsequently increasing resolution to add detailed local features, models can produce coherent yet rich images. In this paper, we present a novel *progressive upscaling* generation process tailored for LDMs (Fig 3 (a)).

Consider a pre-trained latent diffusion model with parameters θ, operating on the latent space $\mathbb{R}^{c\times h\times w}$ to produce images with a resolution magnified by a factor of K. The scaling factor for the side length should be $S=\sqrt{K}$. And the target latent space is $\mathbb{R}^{c\times H\times W}$ where $H=Sh$ and $W=Sw$. Instead of directly synthesizing $\mathbf{z}_t\in\mathbb{R}^{c\times H\times W}$, we break the generation process into S distinct phases, each consisting of an "upsample-diffuse-denoise" loop, except for the first phase which follows an "initialise-denoise" scheme. Specifically, given diffusion and denoising process as $q(\mathbf{z}_T|\mathbf{z}_0)=\prod_{t=1}^{T}q(\mathbf{z}_t|\mathbf{z}_{t-1})$ and $p_\theta(\mathbf{z}_0|\mathbf{z}_T)=\prod_{t=T}^{1}p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$. Then, we can formulate the proposed progressive upscaling generation process as

$$p_\theta(\mathbf{z}_0^S|\mathbf{z}_T^1)=p_\theta(\mathbf{z}_0^1|\mathbf{z}_T^1)\prod_{s=2}^{S}(q(\mathbf{z'}_T^s|\mathbf{z'}_0^s)p_\theta(\mathbf{z}_0^s|\mathbf{z'}_T^s)), \tag{3}$$

where $\mathbf{z'}_0^s$ is obtained through explicit upsampling as $\mathbf{z'}_0^s=inter(\mathbf{z}_0^{s-1})$ and $inter(\cdot)$ is an arbitrary interpolation algorithm (*e.g.*, bicubic). In essence, we first run a regular LDM such as SDXL as $p_\theta(\mathbf{z}_0^1|\mathbf{z}_T^1)$. We then iteratively for each scale s: (i) upscale the low-resolution image $\mathbf{z}_0^{s-1}$ to $\mathbf{z'}_0^s$, (ii) reintroduce noise via the diffusion process to obtain $\mathbf{z'}_T^s$, and (iii) denoise to obtain $\mathbf{z}_0^s$. By repeating this process, we can compensate for the artificial interpolation-based upsampling and gradually fill in more and more local details.

3.3. Skip Residual

The "diffuse-denoise" process has parallels in some image editing works – people attempt to find the initial noise of an image using specialized noise inversion techniques, ensuring that the unedited parts remain consistent with the original image during the denoising editing process [9, 20]. However, these inversion techniques are less practical to DemoFusion's denoising process. Therefore, we instead simply use a conventional diffusion process by adding random Gaussian noise.

However, directly diffusing $\mathbf{z}_0^s$ to $\mathbf{z'}_T^s$ as initialization would result in most information loss. In contrast, diffusing to an intermediate t and then starting denoise from $\mathbf{z'}_t^s$ might be better. However, it is challenging to determine the optimal intersection time-step t of the "upsample-diffuse-denoise" loop – the larger the t, the more information is lost, which weakens the global perception; the smaller the t, the stronger the noise introduced by upsampling (refer to Appendix C). It is a difficult trade-off and

could be example-specific. Therefore, we introduce the skip residual as a general solution, which can be informally considered as a weighted fusion of multiple "upsample-diffuse-denoise" loops with a series of different intersection timesteps t (Fig. 3 (a)).

For each generation phase s, we have already obtained a series of noise-inversed versions of $\mathbf{z'}_0^s$ as $\mathbf{z'}_t^s$ with $t \in [1, T]$. During the denoising process, we introduce the corresponding noise-inversed versions as *skip residuals*. In other words, we modify $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$ to $p_\theta(\mathbf{z}_{t-1}|\hat{\mathbf{z}}_t)$ with

$$\hat{\mathbf{z}}_t^s = c_1 \times \mathbf{z'}_t^s + (1 - c_1) \times \mathbf{z}_t^s, \tag{4}$$

where $c_1 = ((1 + \cos\left(\frac{T-t}{T} \times \pi\right))/2)^{\alpha_1}$ is a scaled cosine decay factor with a scaling factor α_1. This essentially utilizes the results from the previous phase to guide the generated image's global structure during the initial steps of the denoising process. Meanwhile, we gradually reduce the impact of the noise residual, allowing the local denoising paths to optimize the finer details more effectively in the later steps.

3.4. Dilated Sampling

Beyond the explicit integration of global information as a residual, we introduce *dilated sampling* to give each denoising path more global context. The technique of dilating convolutional kernels to expand their receptive field is conventional in various dense prediction tasks [39]. The concurrent tuning-free method, SCALECRAFTER [7], similarly uses dilated convolutional kernels for adapting trained latent diffusion models to higher-resolution image generation. However, our approach diverges here: rather than dilating the convolutional kernel, we directly dilate the sampling within the latent representation. After that, the global denoising paths, derived through dilated sampling, are processed analogously to local denoising paths in MultiDiffusion.

As depicted in Fig. 3 (b), we applied shifted dilated sampling to obtain a series of global latent representation, *i.e.*, $Z_t^{global} = [\mathbf{z}_{0,t}, \cdots, \mathbf{z}_{m,t}, \cdots, \mathbf{z}_{M,t}] = \mathcal{S}_{global}(\mathbf{z}_t)$, $\mathbf{z}_{m,t} \in \mathbb{R}^{c \times h \times w}$. To sample from the whole latent representation, the dilation factor is set to be s and $M = s^2$. Similarly, we apply the general denosing process on these global latent representations as $p_\theta(\mathbf{z}_{m,t-1}|\mathbf{z}_{m,t})$. Then, the reconstructed global representations are fused with the reconstructed local representations to form the final latent representation:

$$\mathbf{z}_{t-1} = c_2 \times \mathcal{R}_{global}(Z_{t-1}^{global}) + (1 - c_2) \times \mathcal{R}_{local}(Z_{t-1}^{local}), \tag{5}$$

where $c_2 = ((1 + \cos\left(\frac{T-t}{T} \times \pi\right))/2)^{\alpha_2}$ is a scaled cosine decay factor with a scaling factor α_2, also chosen based on the characteristic of the diffusion model where earlier steps mainly reconstruct the overall structure, while later steps focus on refining the details.

It is noteworthy that directly using dilated sampling can lead to grainy images. This is because, unlike the local denoising paths, which have overlaps, the global denoising paths operate independently of each other. To address this issue, we employ a straightforward yet intuitive approach – applying a Gaussian filter $\mathcal{G}(\cdot)$ to the latent representation before performing dilated sampling as $Z_t^{global} = \mathcal{S}_{global}(\mathcal{G}(\mathbf{z}_t))$. The kernel size of the Gaussian filter is set to be $4s - 3$, making it sufficient at every phase. Moreover, the standard deviation of the Gaussian filter will decrease from σ_1 to σ_2 as $c_3 \times (\sigma_1 - \sigma_2) + \sigma_2$, where $c_3 = ((1 + \cos\left(\frac{T-t}{T} \times \pi\right))/2)^{\alpha_3}$ is also a scaled cosine decay factor with a scaling factor α_3, ensuring that the effect of the filter gradually diminishes as the directions of global denoising paths become consistent, preventing the final image from becoming blurry.

4. Experiments

Here, we report qualitative and quantitative experiments and ablation studies. For more details and results, please refer to Appendix: implementation details in Appendix B, more discussions in Appendix C, more visualisations in Appendix D, more applications in Appendix E, and all prompts we use in Appendix F.

4.1. Comparison

We compared DemoFusion with the following methods (i) **SDXL** [24], which is designed to generate images of 1024^2. In the quantitative experiments, we also report the results of inferencing it at higher resolutions. (ii) **MultiDiffusion** [2], our baseline method based on overlapped local patch denoising. (iii) **SDXL+BSRGAN**. Using a super-resolution model is an intuitive solution to directly upscale SDXL results. Here, we choose BSRGAN [40], a representative SR method, for comparison. (iv) **SCALECRAFTER** [7], a concurrent training-free high-resolution generation method built on SDXL, which upscales by dilating convolutional kernels at specific layers.

Qualitative Results: As shown in Fig. 4, each model is asked to generate images at $4\times$ and $16\times$ resolutions (compared to SDXL). We chose three prompts about realistic content rather than showcasing DemoFusion's prowess in artistic creation, as such content is more objective and facilitates a fair comparison.

Firstly, as previously mentioned, MultiDiffusion tends to generate repetitive content lacking semantic coherence. For SDXL+BSRGAN, we observe that the SR model effectively eliminates the blurriness and jagged edges of up-sampling, resulting in sharp and pleasing outcomes. However, the goal of the SR model is to produce images consistent with the input, which limits its performance in high-resolution generation – needing more detail for true high-resolution visuals beyond simple smoothing.

Figure 4. **Qualitative comparison with other baselines.** Local details have already been zoomed in, but it's still recommended to **ZOOM IN** for a closer look.

Checking the zoomed-in results of 4096^2 – compared to SDXL+BSRGAN, DemoFusion generates much richer details in the fur of the teddy bear, gives much richer details to Hermoine's eyes, and adds much more detail to the forest vegetation. This comparison confirms that high-resolution generation cannot be substituted by simple image super-

Method	2048 × 2048						2048 × 4096						4096 × 4096					
	FID ↓	IS ↑	FID_{crop} ↓	IS_{crop} ↑	CLIP ↑	Time	FID ↓	IS ↑	FID_{crop} ↓	IS_{crop} ↑	CLIP ↑	Time	FID ↓	IS ↑	FID_{crop} ↓	IS_{crop} ↑	CLIP ↑	Time
SDXL Direct Inference [24]	79.66	13.47	73.91	17.38	28.12	1 min	97.08	14.12	96.41	18.01	27.29	3 min	105.65	14.01	98.59	19.47	25.64	8 min
MultiDiffusion [2]	75.93	14.56	70.93	17.85	28.97	3 min	89.38	14.17	82.78	18.87	28.66	6 min	97.98	13.84	79.45	19.73	28.62	15 min
SDXL + BSRGAN [40]	66.41	16.22	67.42	21.11	29.61	1 min	**68.70**	16.29	75.03	21.76	29.01	1 min	**66.44**	**16.21**	77.20	22.42	**29.63**	1 min
SCALECRAFTER [7]	69.91	15.72	68.36	19.44	29.51	1 min	80.16	15.29	83.08	19.56	28.87	6 min	87.50	15.20	84.36	20.32	29.04	19 min
DemoFusion (Ours)	**65.73**	**16.41**	**64.81**	**21.40**	**29.68**	3 min	73.15	**16.37**	**71.35**	**23.55**	**29.05**	11 min	74.11	16.11	**70.34**	**24.28**	29.57	25 min

Table 1. **Quantitative comparison results**. The best results are marked in **bold**, and the second best results are marked by underline.

Figure 5. **Ablation studies** on the three components of DemoFusion: Progressive Upscaling (PU), Skip Residual (SR), and Dilated Upsampling (DS). All images are generated at 3072^2 ($9\times$ resolutions). Best viewed **ZOOMED-IN**.

resolution. As for SCALECRAFTER, while it partially addresses the issue of MultiDiffusion's repetitive content, it still needs improvement in semantic coherence. *E.g.*, the teddy bear has multiple arms, eyes, or mouths. Additionally, directly dilating the convolutional kernels has somewhat affected the performance of the LDM, resulting in an overall image quality degradation, and local details exhibit many repetitive patterns (*e.g.*, the trunks of the trees). In summary, the proposed DemoFusion achieves both rich local detail and strong global semantic coherence by modifying MultiDiffusion style patch-wise denoising paths to maximise the global context available for each path.

Quantitative Results: For quantitative comparison, we adopt 3 widely-used metrics: FID (Fréchet Inception Distance) [10], IS (Inception Score) [31], and CLIP Score [28]. Considering that FID and IS require resizing images to 299^2, which is not very suitable for high-resolution image assessment, inspired by [4], we additionally crop local patches of $1\times$ resolution and then resize them to calculate these metrics, termed FID_{crop} and IS_{crop}. The CLIP Score assesses the entire image's semantics; thus, we do not consider evaluating local patches here. We evaluate on the LAION-5B dataset [32] with $1K$ randomly sampled captions. Note that the results of FID and IS are related to the number of samples; therefore, the scores of FID_{crop} and IS_{crop} might be better than FID and IS due to more samples. The inference time is evaluated on an RTX 3090 GPU.

As shown in Tab. 1, DemoFusion achieved the best overall performance – securing first or second place across all metrics. As the resolution increases, DemoFusion may score slightly lower than SDXL+BSRGAN on FID and IS because BSRGAN is designed to adhere strictly to low-resolution inputs, and these metrics also downsample images to low resolution for evaluation. However, DemoFusion significantly outperforms SDXL+BSRGAN on FID_{crop} and IS_{crop}, indicating that DemoFusion can provide high-resolution local details. Besides, we observed that MultiDiffusion surpassed SCALECRAFTER on crop-based metrics due to these metrics' lack of an assessment of the overall structure of the image. Herefore, we keep the general FID and IS metrics. Regarding efficiency, since DemoFusion is based on MultiDiffusion and operates progressively, it requires a longer inference time. We discuss this point further in Sec. 5.

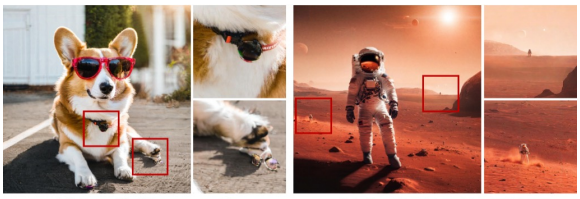
4.2. Ablation Study

The proposed DemoFusion consists of three components: (i) progressive upscaling, (ii) skip residual, and (iii) dilated sampling. To visually demonstrate the effectiveness of these three components, we conducted experiments on all possible combinations, as shown in Fig. 5. All images are generated at 3072^2 ($9\times$ resolutions). When all three components are removed, we generate at the original resolution first and then achieve higher resolutions via an "upsample-diffuse-denoise" loop. The results obtained under this setting are similar to naively generating via MultiDiffusion, with much repetitive content. However, this issue is gradually mitigated by incorporating the three proposed techniques, resulting in high-resolution images consistent with their original resolution counterparts.

(a) Stable Diffusion 1.5

(b) Stable Diffusion 2.1

Figure 6. **Results of DemoFusion on other LDMs**, *i.e.*, Stable Diffusion 1.5 (default resolution of 512^2) and Stable Diffusion 2.1 (default resolution of 768^2). All images are generated at $9\times$ resolutions. Best viewed **ZOOMED-IN**.



(a) Locally Unreasonable (b) Small Object Repetition

Figure 7. **Failure cases of DemoFusion**. (a) Irrational content appears locally in images with a sharp focus. (b) Small objects are repetitively present against a sparse background. All images are generated at $9\times$ resolutions. Best viewed **ZOOMED-IN**.

Specifically, we found that continuously introducing information from the low resolution via skip residual dramatically helps maintain the overall structure to obtain acceptable results. On this basis, dilated sampling can further introduce denoising paths with global perception during the denoising process, guiding local denoising paths towards the global optimal direction. However, these mutually independent global denoising paths introduce two drawbacks (even though we have introduced Gaussian filtering to alleviate this): (i) bringing grainy textures when generating from Gaussian noises and (ii) amplifying the artificial noises introduced during the upscaling process. The former can be alleviated by introducing skip residuals, while the latter can be addressed by progressive upscaling, which prevents the strong artificial noises brought by direct large-scale upscaling. Overall, the three proposed techniques are complementary and indispensable. It is fascinating to see how well they work together.

5. Limitations and Opportunities

DemoFusion exhibits limitations in the following aspects: (i) The nature of MultiDiffusion-style inference requires high computational load due to the overlapped denoising paths, and the progressive upscaling also prolongs inference times. (ii) As a tuning-free framework, DemoFusion's performance is directly correlated with the underlying LDM. In Fig. 6, we show the results based on other LDMs (Stable Diffusion 1.5 and Stable Diffusion 2.1), where DemoFusion is still effective, but the results are less astonishing than those on SDXL. (iii) DemoFusion entirely depends on the LDMs' prior knowledge of cropped images, and therefore, local irrational content may appear when generating sharp close-up images, as depicted in Fig. 7 (a). (iv) Although we have significantly mitigated the issue of repetitive content, the possibility of small repetitive content in background regions remains (see Fig. 7 (b)).

Behind these limitations, opportunities exist: (i) DemoFusion functions by fusing multiple denoising paths of the original size. This allows it to implement each denoising step in mini-batches, preventing the expected exponential increase in memory requirements. (ii) Although progressive upscaling requires more passes, users can acquire low-resolution intermediate results as "previews" within several seconds, facilitating rapid prompt iteration. (iii) The priors of current LDMs regarding image crops are solely derived from the general training scheme, which has already resulted in impressive performance. Training a bespoke LDM for a DemoFusion-like framework may be a promising direction to explore.

6. Conclusion

In this paper, we introduce DemoFusion, a tuning-free framework that integrates plug-and-play with open-source GenAI models to achieve higher-resolution image generation. DemoFusion is built upon MultiDiffusion and introduces *Progressive Upscaling*, *Skip Residual*, and *Dilated Sampling* techniques to enable generation with both global semantic coherence and rich local details. DemoFusion persuasively demonstrates the possibility of LDMs generating images at higher resolutions than those used for training and the untapped potential of existing open-source GenAI models. By advancing the frontier of high-resolution image generation without additional training or prohibitive memory requirements for inference, we hope that DemoFusion can help democratize high-resolution image generation.

Acknowledgement

This work was supported in part by Beijing Natural Science Foundation Project No. Z200002, in part by National Natural Science Foundation of China (NSFC) No. 62225601, U23B2052, in part by scholarships from China Scholarship Council (CSC) under Grant CSC No. 202206470055, in part by BUPT Excellent Ph.D. Students Foundation No. CX2022152, in part by the China Postdoctoral Science Foundation No. 2023M741961, and in part by the Program for Youth Innovative Research Team of BUPT No. 2023QNTD02.

References

[1] Stability AI. Stable diffusion: A latent text-to-image diffusion model. https://stability.ai/blog/stable-diffusion-public-release, 2022. 1, 2

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023. 2, 3, 5, 7

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3

[4] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *ECCV*, 2022. 7

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[6] Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K Wong. Headsculpt: Crafting 3d head avatars with text. In *NeurIPS*, 2023. 3

[7] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. *arXiv preprint arXiv:2310.07702*, 2023. 3, 5, 7

[8] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2023. 3

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2022. 3, 4

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 7

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[12] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 2022. 3

[13] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 3

[14] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *CVPR*, 2023. 3

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 4

[16] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *arXiv preprint arXiv:2306.05178*, 2023. 3

[17] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 3

[18] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 3

[19] MidJourney. Midjourney: An independent research lab. https://www.midjourney.com/, 2022. 1, 2

[20] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 3, 4

[21] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[22] Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Dreamcreature: Crafting photorealistic virtual creatures from imagination. *arXiv preprint arXiv:2311.15477*, 2023. 3

[23] OpenAI. Dall·e: Creating images from text. https://openai.com/blog/dall-e/, 2021. 1, 2

[24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 3, 5, 7

[25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2022. 3

[26] Zhiyu Qu, Tao Xiang, and Yi-Zhe Song. Sketchdreamer: Interactive text-augmented creative sketch ideation. In *BMVC*, 2023. 3

[27] Zhiyu Qu, Lan Yang, Honggang Zhang, Tao Xiang, Kaiyue Pang, and Yi-Zhe Song. Wired perspectives: Multi-view wire art embraces generative ai. In *CVPR*, 2024. 3

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 2016. 7

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 7

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3

[35] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 3

[36] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 3

[37] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *CVPR*, 2023. 3

[38] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *CVPR*, 2023. 3

[39] Fisher Yu and Koltun Vladlen. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 5

[40] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *CVPR*, 2021. 3, 5, 7

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[42] Qingping Zheng, Yuanfan Guo, Jiankang Deng, Jianhua Han, Ying Li, Songcen Xu, and Hang Xu. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. *arXiv preprint arXiv:2308.16582*, 2023. 2, 3