Supplementary Material

(a) Stable Diffusion 1.5 (1024 × 1024, 4×)

(b) Stable Diffusion 2.1 (1536 × 1536, 4×)

Fig. 1: Results of AccDiffusion on other stable diffusion variants: (a) Stable diffusion 1.5 (default resolution of 512^2) and (b) Stable diffusion 2.1 (default resolution of 768^2). All images are generated at 4× resolution. Prompts are provided in Sec. H.

A More Stable Diffusion Variants

We apply AccDiffusion on other LDMs, specifically Stable Diffusion 1.5 (SD 1.5) [5] and Stable Diffusion 2.1 [6] (SD 2.1). As shown in Fig. 1, AccDiffusion successfully generates higher-resolution images without repetition. It is important to note that the results of AccDiffusion depend on the prior knowledge of LDMs, and the performance of SD 1.5 and SD 2.1 is inferior to SDXL [4]. Therefore, the fidelity of their results are less astonishing than those on SDXL.

Fig. 2: More higher-resolution results of AccDiffusion on SDXL (default resolution of 1024^2). Best viewed with zooming in.

B More Visualization

We provide more results of AccDiffusion on SDXL. As shown in Fig. 2, our AccDiffusion can generate various higher-resolution images without object repetition. Prompts are provided in Sec. H.

C Default Setting of DemoFusion

To ensure a fair comparison with DemoFusion [2], we conduct our experiments using its default settings listed in Table 1. For a more comprehensive understanding of DemoFusion, please refer to the original paper [2].

Table 1: The default setting of DemoFusion [2].

Parameters	Explanation	Values
T	DDIM Steps	50
s	Guidance Scale	7.5
h	Latent Height	128
w	Latent Width	128
d_h	Height Stride	$\frac{h}{2}$
d_w	Width Stride	$\frac{w}{2}$
α_1	Scale factor 1	3
α_2	Scale factor 2	1
α_3	Scale factor 3	1

D More Qualitative Comparison Results

We provide more qualitative comparison results in Fig. 3 and Fig. 4. More qualitative results provide stronger evidence that our method can generate high-resolution images without repetition.

E Details on any aspect ratio generation.

First, we initialize a latent noise with the expected ratio and set the longer side to training resolution (*e.g.*, 1024×512 for 2:1). Then we use the same pipeline as the 1:1 aspect ratio to progressively generate higher-resolution images, as shifted window sampling and dilated sampling are compatible with any aspect ratio. More details can be found in DemoFusion [2].

F Indefinite extrapolation.

Following the recent works, we provide results in main paper within 4K for comparisons. Ideally, both AccDiffusion and patch-wise methods can extrapolate indefinitely. However, we find that AccDiffusion faces detail degradation when the resolution is beyond 6K ($36\times$), as shown in Fig. 5.

G Pseudo Code of AccDiffusion

AccDiffusion follows the pipeline of DemoFusion [2] and uses the patch-content-aware prompts during the progress of higher-resolution image generation. Additionally, AccDiffusion enhances dilated sampling with window interaction. Algorithm 1 illustrates the process of higher-resolution generation using AccDiffusion. We use red color to highlight two core modules proposed by AccDiffusion.

H Prompts Used in Supplement Material

Fig. 1:

1. A butterfly landing on a sunflower.
2. A fox peeking out from behind a bush.
3. A picturesque mountain scene with a clear lake reflecting the surrounding peaks.
4. A cute panda on a tree trunk.
5. A corgi wearing cool sunglasses.
6. Primitive forest, towering trees, sunlight falling, vivid colors.

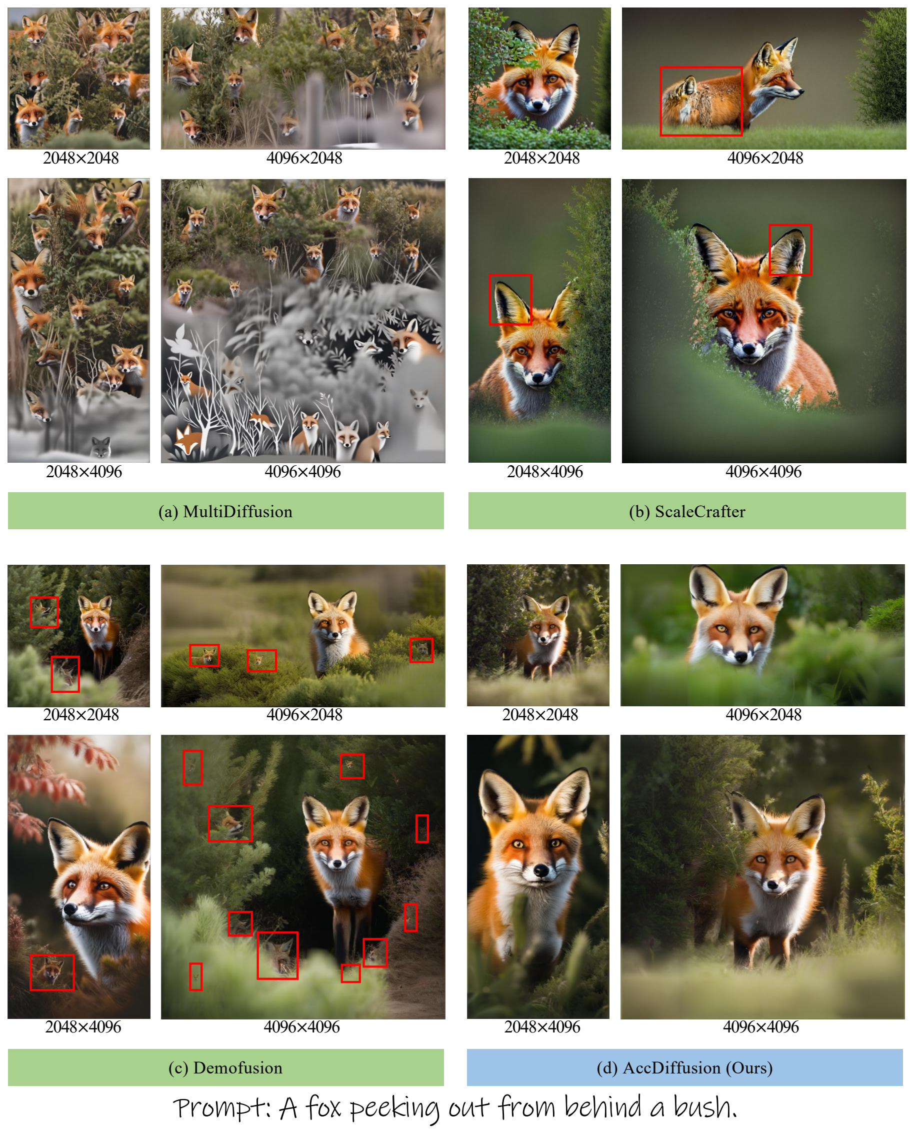

Fig. 3: Qualitative comparison of our AccDiffusion with existing training-free image generation extrapolation methods [1–3]. We draw a red box upon the generated images to highlight the repeated objects. Best viewed zoomed in.

Fig. 2:

1. A close-up of a fire spitting dragon, cinematic shot.
2. Cute adorable little goat, unreal engine, cozy interior lighting, art station, detailed' digital painting, cinematic, octane rendering.

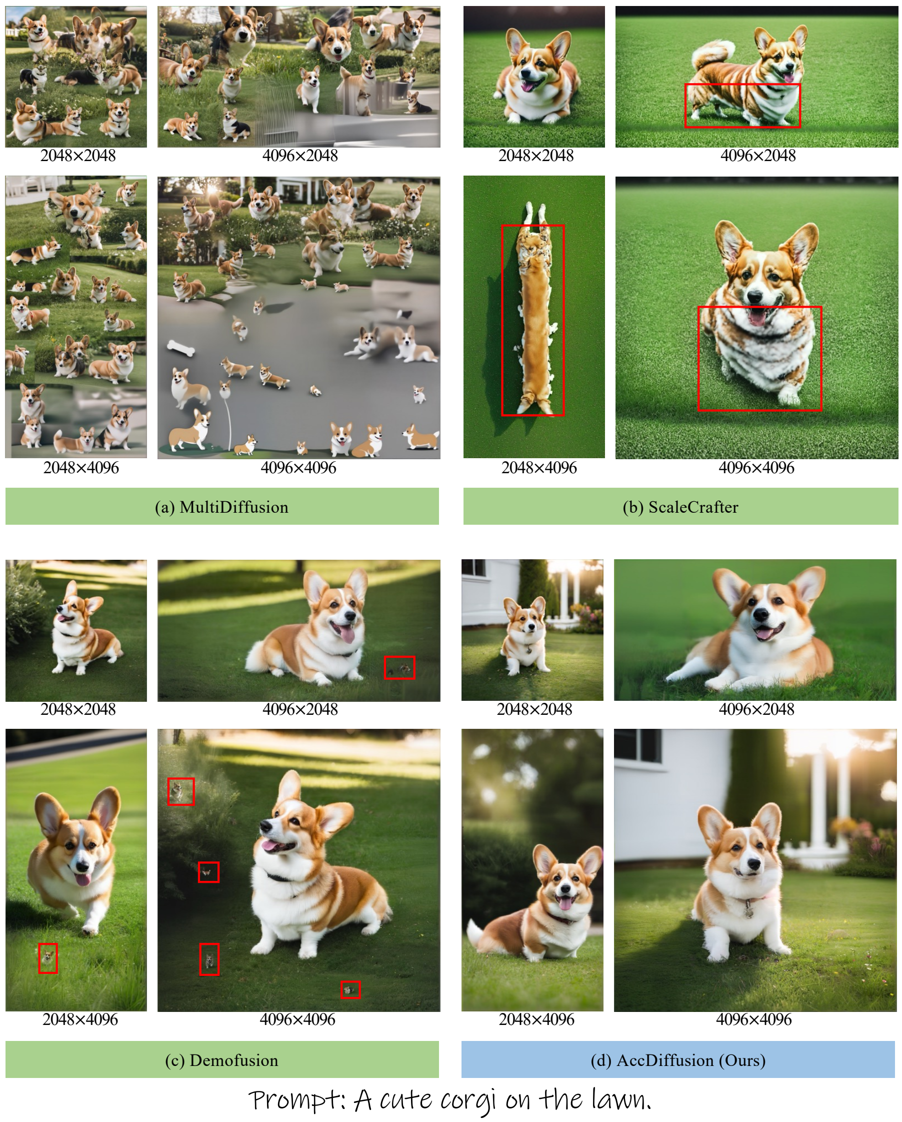

Fig. 4: Qualitative comparison of our AccDiffusion with existing training-free image generation extrapolation methods [1–3]. We draw a red box upon the generated images to highlight the repeated objects. Best viewed zoomed in.

3. A propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese.
4. A cute panda on a tree trunk.
5. a photograph of a red ball on a blue cube.

Prompt : "Summer landscape, vivid colors, a work of art, grotesque, Mysterious."

Fig. 5: Failure case of indefinite extrapolation.

6. a baby penguin wearing a blue hat, red gloves, green shirt, and yellow pants.
7. a cat drinking a pint of beer.
8. A young badger delicately sniffing a yellow rose, richly textured oil painting.
9. A cute cat on the lawn.

Algorithm 1 The process of higher-resolution generation using AccDiffusion

Input: h', w' ▷ Latent Size of Desired Image
$\mathcal{E}_\theta, h, w$ ▷ Pre-trained Stable diffusion and Pre-trained Latent Size
$y, \mathcal{D}$ ▷ Prompt and Decoder
η_1, η_2 ▷ Decreasing From 1 to 0 Using a Cosine Schedule

`############## Phase 1: Low resolution image generation ##############`
$\mathbf{z}_T \sim \mathcal{N}(0, I)$ ▷ Random Initialization
for $t = T$ to 1 **do**
    $\mathbf{z}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\mathbf{z}_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta\big(\mathbf{z}_t, t, \tau_\theta(y)\big).$
    ▷ Denoising with Image-content-aware Prompt and Save Cross-Attenion Map $\mathcal{M}$
end for
$\mathcal{Z}_0 = \mathbf{z}_0$
$S = \frac{h'}{h} \times \frac{w'}{w}$ ▷ Progressive Upscaling Times
`############ Phase 2: Higher-resolution image generation ############`
for $s = 2$ to S **do** ▷ Progressive Upscaling
    $\mathcal{Z}_0 = inter(\mathcal{Z}_0, (h \times s, w \times s))$ ▷ Interpolation Upsampling
    for $t = 1$ to T **do**
        $\mathcal{Z}'_t = \sqrt{\bar{\alpha}_t}\mathcal{Z}_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{I})$
        ▷ Getting Noise-inversed Representations
    end for
    $\mathcal{Z}_T = \mathcal{Z}'_T$
    for $t = T$ to 1 **do**
        $\hat{\mathcal{Z}}_t = \eta_1 \times \mathcal{Z}'_t + (1 - \eta_1) \times \mathcal{Z}_t$ ▷ Skip Residual
        $\{\mathbf{z}_t^i\}_{i=1}^{P_1} = \text{Sampling}_1(\hat{\mathcal{Z}}_t)$ ▷ Shift Window Sampling From MultiDiffusion
        $\mathcal{M} \to \{\gamma^i\}_{i=1}^{P_1}$ ▷ Calculating Patch-Content-Aware Prompt
        $\{\mathcal{D}_t^i\}_{i=1}^{P_2} = \text{Sampling}_2(\hat{\mathcal{Z}}_t)$ ▷ Dilated Sampling From DemoFusion
        for $\mathbf{z}_t^i$ in $\{\mathbf{z}_t^i\}_{i=1}^{P_1}$ **do**
            $\mathbf{z}_{t-1}^i = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\mathbf{z}_t^i + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta\big(\mathbf{z}_t^i, t, \tau_\theta(\gamma^i)\big).$
            ▷ Denoising with Patch-Content-Aware Prompt
        end for
        for $\mathcal{D}_t^i$ in $\{\mathcal{D}_t^i\}_{i=1}^{P_2}$ **do**
            ${\mathcal{D}_t}^{k,h,w} = {\mathcal{D}_t}^{f_t^{h,w}(k),h,w}$ ▷ Window Interaction with Bijective Function
            $\mathcal{D}_{t-1}^i = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\mathcal{D}_t^i + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta\big(\mathcal{D}_t^i, t, \tau_\theta(y)\big)$
            ▷ Denoising with Image-Content-Aware Prompt
            ${\mathcal{D}_{t-1}}^{k,h,w} = {\mathcal{D}_{t-1}}^{(f_t^{h,w})^{-1}(k),h,w}$ ▷ Recover

        end for
        $\mathcal{Z}_{t-1} = \eta_2 \times \text{Fuse}(\{\mathcal{D}_t^i\}_{i=1}^{P_2}) + (1 - \eta_2) \times \text{Fuse}(\{\mathbf{z}_t^i\}_{i=1}^{P_1})$
        ▷ Fusing Shift Window Sampling Patches and Dilated Sampling Patches
    end for
end for
Output: $\mathbf{x}_0 = \mathcal{D}(\mathcal{Z}_0)$ ▷ Decoding to Image

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. In: ICML (2023)
2. Du, R., Chang, D., Hospedales, T., Song, Y.Z., Ma, Z.: Demofusion: Democratising high-resolution image generation with no $$$. In: CVPR (2024)
3. 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)
4. 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)
5. Robin Rombach, P.E.: Stable diffusion v1-5 model card, `https://huggingface.co/runwayml/stable-diffusion-v1-5`
6. Robin Rombach, P.E.: Stable diffusion v2-1 model card, `https://huggingface.co/stabilityai/stable-diffusion-2-1`