Supplementary Materials for
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

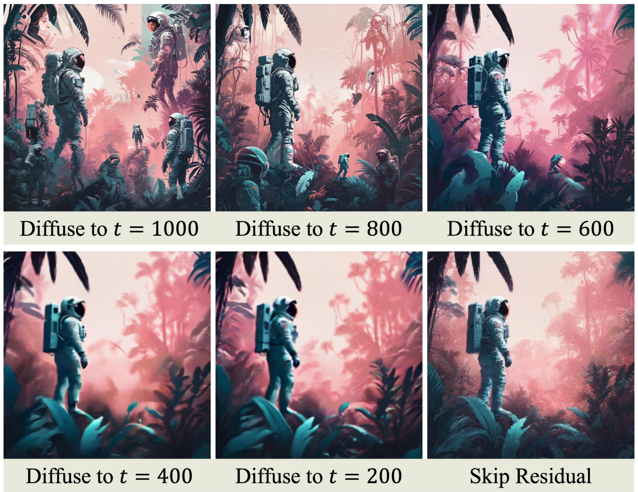

Figure 1. During the progressive upscaling process, we diffuse $\mathbf{z}_0^s$ to different time-steps t, and then denoise it back to obtain the results. The number of training time-step is 1000.

PU	SR	DS	FID	IS	FID$_{crop}$	IS$_{crop}$	CLIP
×	×	×	95.28	13.92	80.11	19.61	28.13
✓	×	×	90.77	13.95	79.23	20.08	28.52
×	✓	×	79.93	15.19	74.17	22.48	29.40
×	×	✓	94.35	14.89	82.32	19.64	28.85
✓	✓	×	75.92	15.66	72.98	23.20	29.50
✓	×	✓	89.26	15.02	80.04	21.86	28.87
×	✓	✓	76.53	15.71	73.22	23.09	29.48
✓	✓	✓	**74.11**	**16.11**	**70.34**	**24.28**	**29.57**

Table 1. Quantitative results of the ablation study. The best results are marked in **bold**. Impact of components: Progressive Upscaling (PU), Skip Residual (SR), and Dilated Upsampling (DS).

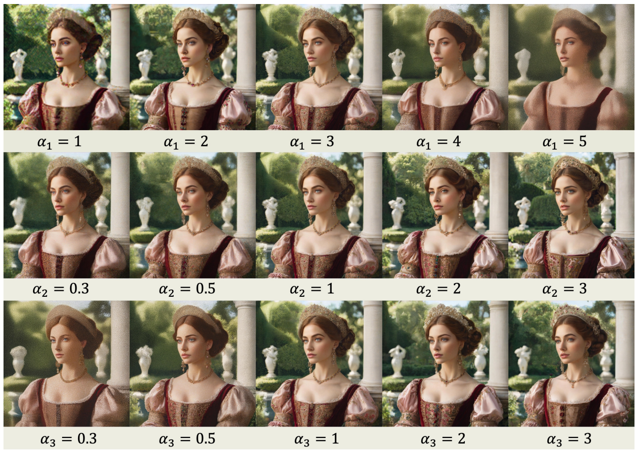

Figure 2. **Results with different α_1, α_2, and α_3.** All images are generated at 3072^2 ($9\times$ resolutions). Best viewed **ZOOMED-IN**.

A. Pseudo Code

We further illustrate the image synthesis process of DemoFusion in Algorithm 1.

B. Implementation Details

In cases where it is not explicitly stated, all the results in this paper are obtained based on SDXL with a DDIM scheduler of 50 steps. The guidance scale for all denoising paths is set to 7.5. The crop size of MultiDiffusion is set to be aligned with the maximum training size of pre-trained LDMs, *e.g.*, $h = w = 128$ for SDXL, and the stride is set to be $d_h = \frac{h}{2}$ and $d_w = \frac{w}{2}$. Each crop's position is subjected to a slight random perturbation, with maximum offsets of $\frac{h}{16}$ and $\frac{w}{16}$ in vertical and horizontal directions, respectively, further preventing the occurrence of seam issues.

When generating images with varying aspect ratios, we ensure that the longer side aligns with the maximum training size. Three scale factors α_1, α_2, α_3 were set to 3, 1, and 1 respectively. The Gaussian filter's standard deviation decreases from $\sigma_1 = 1$ to $\sigma_2 = 0.01$. To decode high-resolution images, we also employed a tiled decoder

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

[*] Corresponding Author

Algorithm 1 Image Synthesis Process of DemoFusion

```
########################## Phase 1 ############################
$\mathbf{z}_T^0 \sim \mathcal{N}(0, I)$                                        ▷ Random Initialization
for $t = T$ to $1$ do
    $p_\theta(\mathbf{z}_{t-i}^1|\mathbf{z}_t^1)$                                  ▷ Denoising Step
end for
####################### Phase 2 to $S$ #######################
for $s = 2$ to $S$ do
    $inter(\mathbf{z'}_0^s|\mathbf{z}_0^{s-1})$                                      ▷ Upsampling
    for $t = 1$ to $T$ do
        $q(\mathbf{z'}_t^s|\mathbf{z'}_{t-1}^s)$                                   ▷ Diffusion Step
    end for
    for $t = T$ to $1$ do
        $\hat{\mathbf{z}}_t^s = c_1 \times \mathbf{z'}_t^s + (1 - c_1) \times \mathbf{z}_t^s$                ▷ Skip Residual
        $\mathcal{S}_{local}(\hat{\mathbf{z}}_t^s) \rightarrow Z_t^{local}$                 ▷ Crop Sampling (MultiDiffusion)
        $\mathcal{S}_{global}(\hat{\mathbf{z}}_t^s) \rightarrow Z_t^{global}$                          ▷ Dilated Sampling
        for $\hat{\mathbf{z}}_{n,t}^s$ in $Z_t^{local}$ do
            $p_\theta(\mathbf{z}_{n,t-i}^s|\hat{\mathbf{z}}_{n,t}^s)$          ▷ Local Path Denoising Step (MultiDiffusion)
        end for
        for $\hat{\mathbf{z}}_{m,t}^s$ in $Z_t^{global}$ do
            $p_\theta(\mathbf{z}_{m,t-i}^s|\hat{\mathbf{z}}_{m,t}^s)$                     ▷ Global Path Denoising Step
        end for
        $\mathcal{R}_{local}(Z_{t-1}^{local}) \times (1 - c_2) + \mathcal{R}_{global}(Z_{t-1}^{global}) \times c_2 \rightarrow \mathbf{z}_t^s$   ▷ Fusing Local and Global Paths
    end for
end for
return $\mathbf{x}_0^S = \mathcal{D}(\mathbf{z}_0^S)$                                ▷ Decoding to Image
```

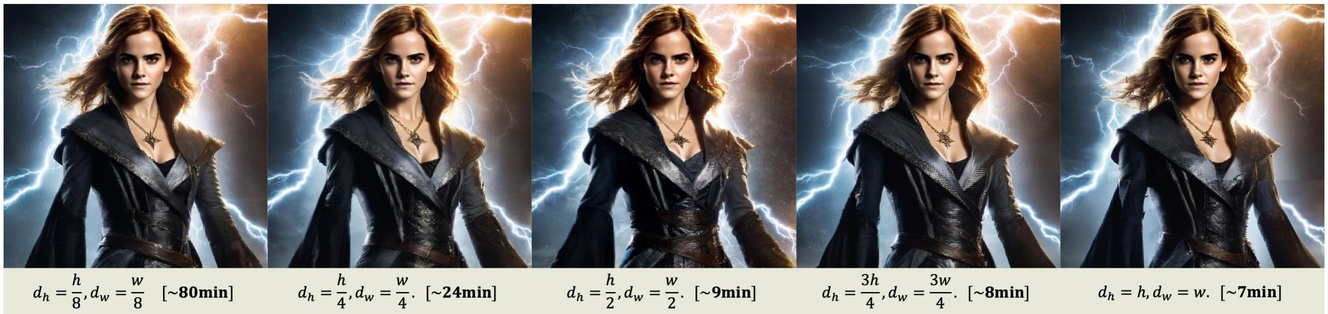
$d_h = \frac{h}{8}, d_w = \frac{w}{8}$ [~80min]
$d_h = \frac{h}{4}, d_w = \frac{w}{4}$. [~24min]
$d_h = \frac{h}{2}, d_w = \frac{w}{2}$. [~9min]
$d_h = \frac{3h}{4}, d_w = \frac{3w}{4}$. [~8min]
$d_h = h, d_w = w$. [~7min]

Figure 3. **Results with different strides d_h and d_w.** All images are generated at 3072^2 (9× resolutions). Best viewed **ZOOMED-IN**.

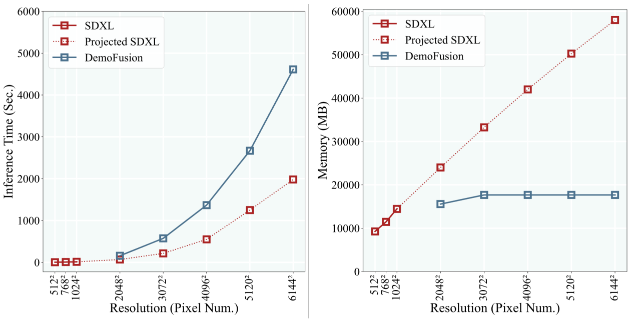

Figure 4. Inference time of SDXL *versus* DemoFusion (**Left**). Memory demands of SDXL *versus* DemoFusion (**Right**).

strategy as [1] and some open-source projects. To eliminate seams between tiles, we sample a larger range of features around each tile during the decoding process.

Note that SDXL permits the input of a coarse cropping condition [3], *i.e.*, the coordinates of the top-left corner of the cropping area. Therefore, when utilizing SDXL, we additionally input the coordinates of the top-left corner of the corresponding patch as a condition for local denoising paths. However, we observed that the presence or absence of this condition does not significantly impact the results. Besides, DemoFusion initiates the generation process from

https://github.com/pkuliyi2015/multidiffusion-upscaler-for-automatic1111

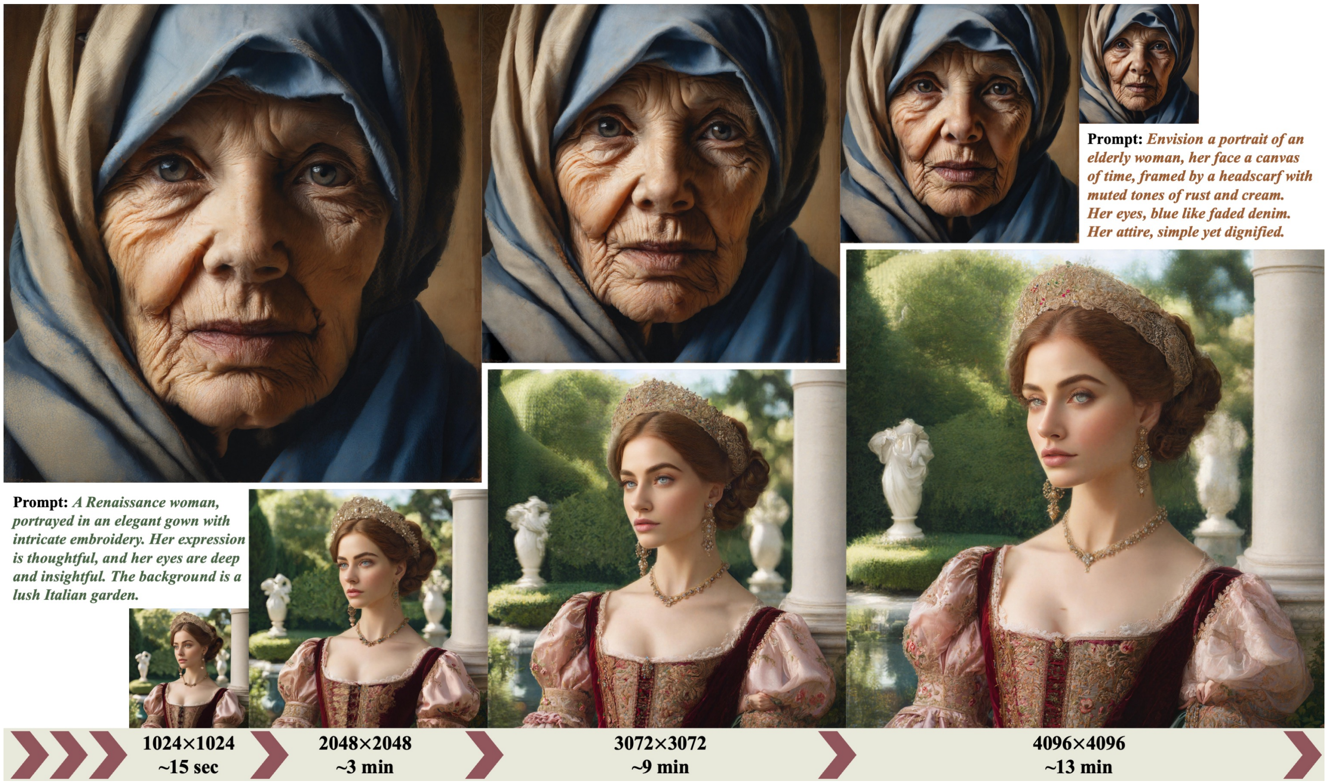
Prompt: Envision a portrait of an elderly woman, her face a canvas of time, framed by a headscarf with muted tones of rust and cream. Her eyes, blue like faded denim. Her attire, simple yet dignified.
Prompt: A Renaissance woman, portrayed in an elegant gown with intricate embroidery. Her expression is thoughtful, and her eyes are deep and insightful. The background is a lush Italian garden.
1024×1024
~15 sec
2048×2048
~3 min
3072×3072
~9 min
4096×4096
~13 min

Figure 5. **Illustration of the progressive upscaling process.** The time required for each phase is indicated. Best viewed **ZOOMED-IN**.

the highest resolution of the LDM during the first phase. When generating images with varying aspect ratios, we ensure that the longer side aligns with the highest resolution.

C. More Experimental Results

C.1. Diffusing to Different Time-step t

In Fig. 1, we illustrate that, when skip residual is removed, the effects of different time-steps we denoise to within the "upsample-diffuse-denoise" loop. This provides evidence for our discussion in the main text – 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.

C.2. Quantitative Results of Ablation Study

The quantitative results of the ablation study are shown in Tab. 1. Here, we only experiment with the resolution of 4096^2.

C.3. Effects of Scale Factors α_1, α_2, and α_3

A shared understanding of the DM's denoising process is that the DM first determines the coarse details and then gradually refines the local details. In line with this understanding, we adopt a unified strategy: utilizing cosine descending weights, we assign greater weights to skip residuals, dilated sampling, and accompanying Gaussian filtering in the early stages of the denoising process, gradually decreasing the weights as denoising progresses. Despite this unified approach, the three components still need distinct scale factors to control the descent rate.

Through grid search, we obtained the globally optimal parameter combination. In Fig. 2, we varied only one parameter at a time while keeping the others at their optimal values to demonstrate the impact of each parameter on the results. Note that a larger scale factor means a faster decline, which weakens the effect of this item, and vice versa.

According to the experimental results, when the skip residual effect is too strong (*i.e.*, $\alpha_1 = 1$), we observe significant artificial noise caused by upsampling. Because these factors interact with each other, when $\alpha_1 = 5$, we observe that the results are close to the one when $\alpha_3 = 1$ – Gaussian filtering leads to excessive smoothing of latent representation. The trade-off of dilated sampling is – too large a weight (*i.e.*, $\alpha_2 = 1$) can result in grainy images, while too small a weight (*i.e.*, $\alpha_2 = 5$) fails to provide sufficient global perception, leading to noticeable issues of content repetition. Regarding Gaussian filtering, excessive strength can lead to over-smoothing of the latent representation, while too small a strength can weaken the global denoising paths due to lack of interaction, resulting in content

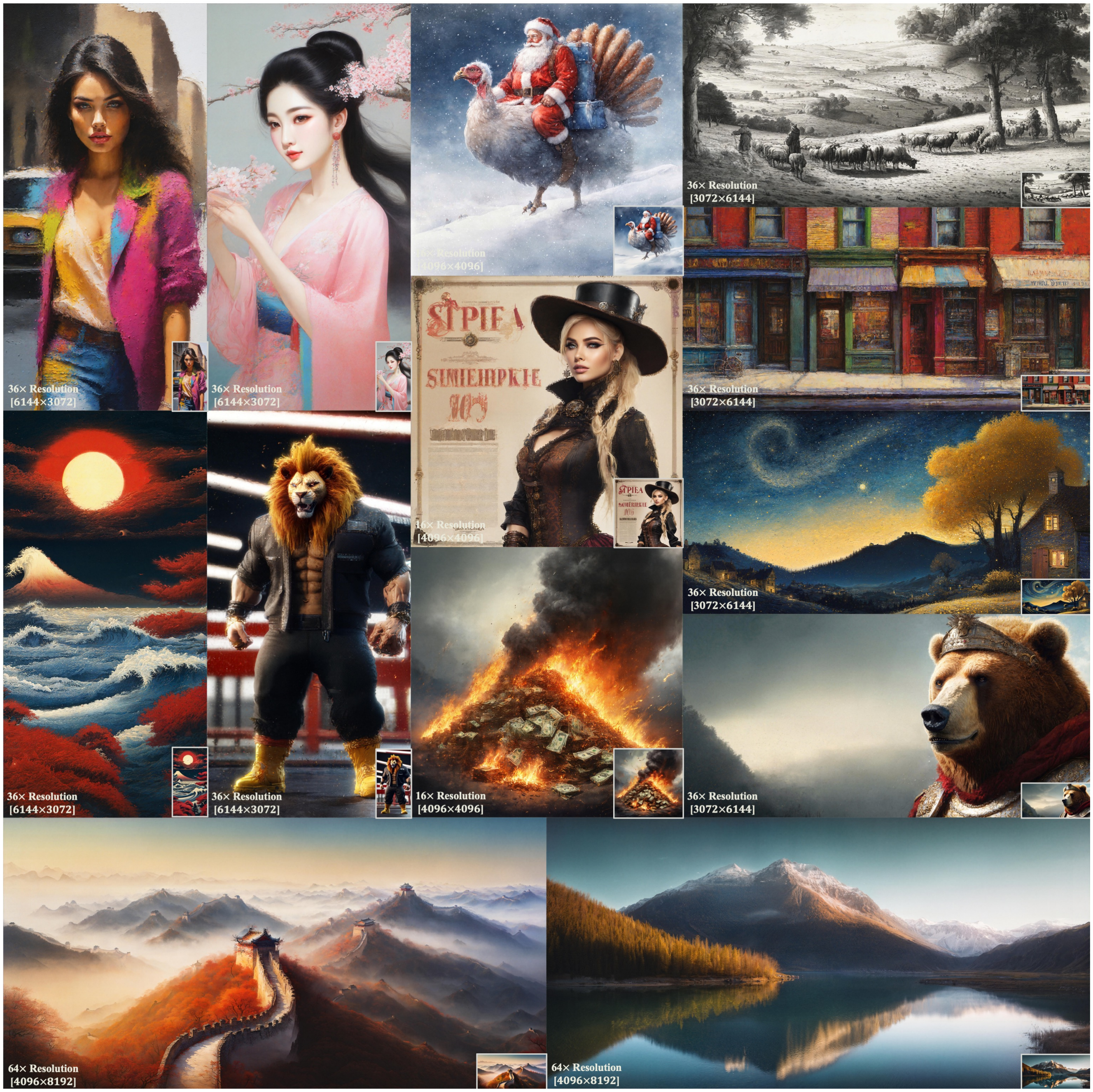

Figure 6. **More selected landscape samples of DemoFusion *versus* SDXL** [3] (all images in the figure are presented at their actual sizes). All generated images are produced using a single RTX 3090 GPU. Best viewed **ZOOMED-IN**.

repetition and grainy appearance.

C.4. Effect of Stride Sizes d_h and d_w

In general, the stride size d_h and d_w in MultiDiffusion determines the extent of the seam issue of images – a smaller stride means more seamless images, but at the same time, brings more overlapping computation. For DemoFusion, due to the proposed *Progressive Upscaling*, *Skip Residual*, and *Dilated Sampling techniques*, there is a better consistency between patches, relaxing the stride size requirement. In Fig. 3, we showcase the performance of DemoFusion and the corresponding inference time for different stride sizes. It can be observed that even in the case of $d_h = h$ and $d_w = w$, *i.e.*, no overlap between patches, we still achieve good global semantic coherence, while when $d_h < h$ and $d_w < w$, all the generated images have no noticeable seams;

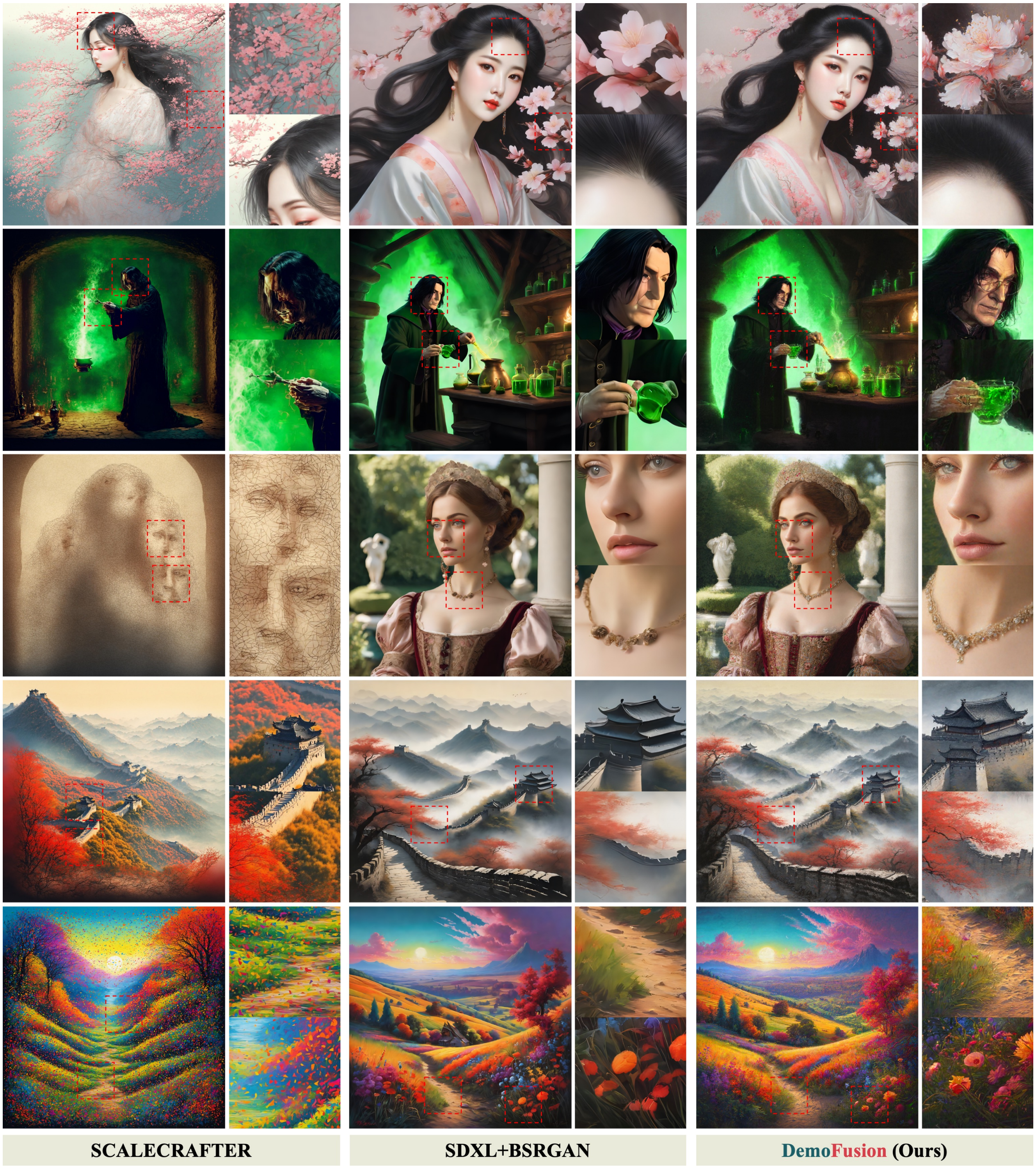

Figure 7. **More Qualitative comparison results.** All images are generated at 4096^2 ($16\times$ resolutions). Local details have already been zoomed in, but it's still recommended to **ZOOM IN** for a closer look.

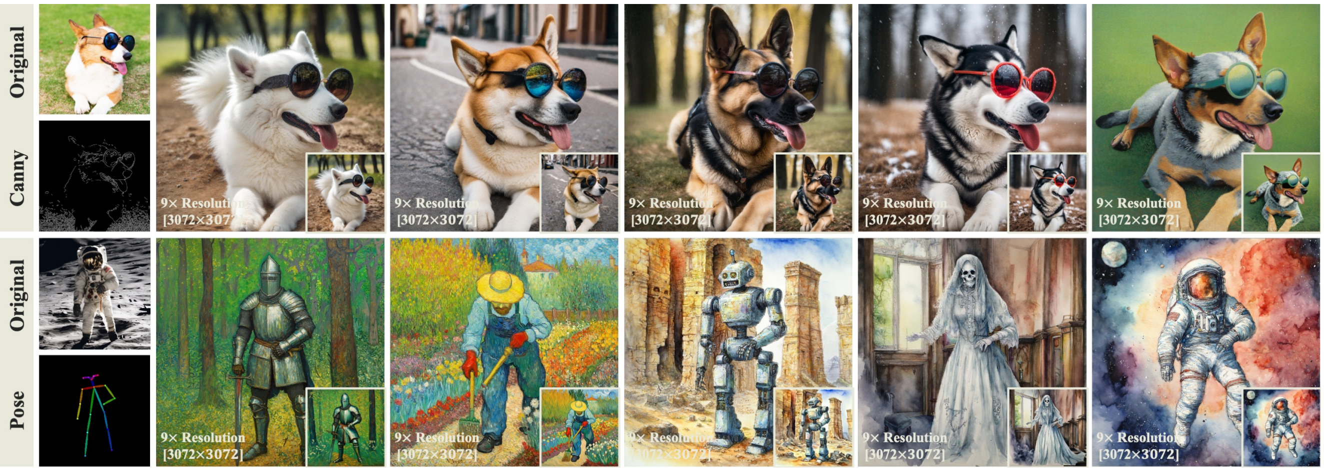

Figure 8. **Results of DemoFusion combining with ControlNet** [5]. All images are generated at 3072^2 ($9\times$ resolutions). Best viewed **ZOOMED-IN**.

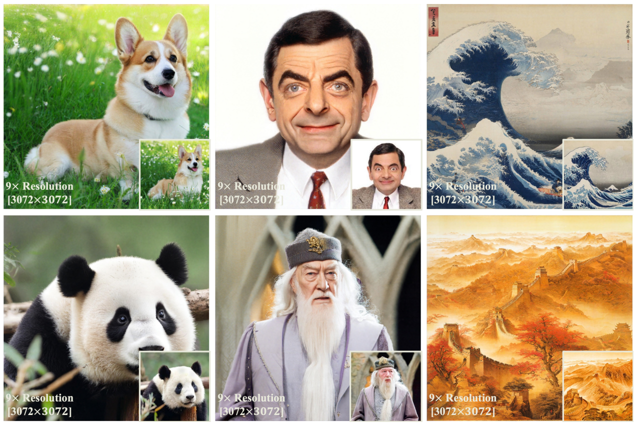

Figure 9. **Results of upscaling real images**. All images are upsacled to 3072^2. Best viewed **ZOOMED-IN**.

ultimately, in order to balance the performance and efficiency, we chose $d_h = \frac{h}{2}$ and $d_w = \frac{w}{2}$.

C.5. Resource Demands of DemoFusion

In Fig. 4, we illustrate the resource demand comparison of DemoFusion and the original SDXL [3]. Note that SDXL cannot generate valid content at resolutions higher than 1024^2. We just calculate the expected resource demands by assuming we have a high-resolution SDXL under the current framework. It can be seen that DemoFusion achieves high-resolution image generation on limited computational resources while paying a little bit more time cost.

D. More Visualizations

D.1. The Progressive Upscaling Process

To better demonstrate how the model progressively generates images with different resolutions, in Fig. 5, we show the model outputs at each phase of the generation process. We can observe that DemoFusion does an excellent job of achieving global consistency under different resolutions, indicating the reason for its success in resolving content repetition.

D.2. More Landscape Samples

In Fig. 6, we have supplemented more samples to show the performance of DemoFusion, and in particular, we further show results at the resolution of 8192×4096 ($64\times$ upscaling compared to the initial resolution of 512×1024).

D.3. More Comparison Results

In Fig. 7, we have supplemented more comprison results with **SDXL+BSRGAN** [4] and **SCALECRAFTER** [1] to demonstrate the effectiveness of DemoFusion. The results are consistent with those in the main text. Compared to SDXL+BSRGAN, DemoFusion provides better local details; while compared to SCALECRAFTER, DemoFusion better preserves the performance of SDXL during upscale.

E. More Applications

E.1. Combining with ControlNet

The tuning-free characteristic of DemoFusion enables seamless integration with many LDM-based applications. *E.g.*, DemoFusion combined with ControlNet [5] can achieve controllable high-resolution generation. In Fig. 8, we showcase examples using Canny edge and human pose as conditions.

E.2. Upscaling Real Images

Since DemoFusion works in a progressive manner, we can replace the output of phase 1 with representations obtained by encoding real images, thereby achieving upscaling of real images. However, we carefully avoid using the term "super resolution", as the outputs tend to lean towards the latent data distribution of the base LDM, making this process more akin to image generation based on a real image. The results are shown in the Fig. 9.

F. Prompts Used in This Paper

All prompts used in this paper are taken from the internet or generated by ChatGPT [2]. They are summarised here.

Fig. 1 in the main text:

- *Steampunk makeup, in the style of vray tracing, colorful impasto, uhd image, indonesian art, fine feather details with bright red and yellow and green and pink and orange colours, intricate patterns and details, dark cyan and amber makeup. Rich colourful plumes. Victorian style.*
- *Stunning feminine body, commercial image, beautiful girl from Spain, holographic photography shoots, large body of water sprayed, liquid splashing all over the places, street pop, luminous palette, close up, realistic impressionism, shiny/glossy, extreme colorsplash, behind that a universe of vortex of fire waves and ice waves, around fire splashes and ice splashes and floral, bonsais, roots, smoke swirls, dust swirls, tentacles of fire and ice, s-curve composition, leading lines, cinematic, style of hokusai, unreal engine, octane render, asymetric, golden ratio, style of hokusai, liquid splashes, merging, melting, splashing, droplets, mixing, fading away, exploding, swirling, intricate detail, modelshoot style, dreamlikeart, dramatic lighting. 8k, highly detailed, trending artstation.*
- *The beautiful scenery of Seattle, painting by Al Capp.*
- *By Tang Yau Hoong, ultra hd, realistic, vivid colors, highly detailed, UHD drawing, pen and ink, perfect composition, beautiful detailed intricate insanely detailed octane render trending on artstation, 8k artistic photography, photorealistic concept art, soft natural volumetric cinematic perfect light, ultra hd, realistic, vivid colors, highly detailed, UHD drawing, pen and ink, perfect composition, beautiful detailed intricate insanely detailed octane render trending on artstation, 8k artistic photography, photorealistic concept art, soft natural volumetric cinematic perfect light.*
- *A cute and adorable fluffy puppy wearing a witch hat in a halloween autumn evening forest, falling autumn leaves, brown acorns on the ground, halloween pumpkins spiderwebs, bats, a witch's broom.*
- *A robot standing in the rain reading newspaper, rusty and worn down, in a dystopian cyberpunk street, photo-realistic, urbanpunk.*
- *Einstein, a bronze statue, with a fresh red apple on his head, by Bruno Catalano.*
- *A woman in a pink dress walking down a street, cyberpunk art, inspired by Victor Mosquera, conceptual art, style of raymond swanland, yume nikki, restrained, robot girl, ghost in the shell.*
- *Photo of a rhino dressed suit and tie sitting at a table in a bar with a bar stools, award winning photography, Elke vogelsang.*
- *An astronaut riding a horse on the moon, oil painting by Van Gogh.*
- *Classic traditional cornucopia at the fall harvest festival, farm in the background, high quality masterful still-life painting, American pastoral, oil painting, festive spirit, vibrant cultural tradition, Autumnal atmosphere, vibrant rich colors.*

Fig. 2 in the main text:

- *An astronaut riding a horse on the moon, oil painting by Van Gogh.*

Fig. 3 in the main text:

- *An astronaut riding a horse on the moon, oil painting by Van Gogh.*

Fig. 4 in the main text:

- *A cute teddy bear in front of a plain white wall, warm and brown fur, soft and fluffy.*
- *Emma Watson as a powerful mysterious sorceress, casting lightning magic, detailed clothing.*
- *Primitive forest, towering trees, sunlight falling, vivid colors.*

Fig. 5 in the main text:

- *Astronaut in a jungle, cold color palette, muted colors, detailed, 8k.*
- *Emma Watson as a powerful mysterious sorceress, casting lightning magic, detailed clothing.*

Fig. 6 in the main text:

- *A panda wearing sunglasses.*
- *Astronaut on Mars During sunset.*
- *A serene lakeside during autumn, with trees displaying a palette of fiery colors.*
- *A hamster piloting a tiny hot air balloon.*
- *An astronaut riding a horse.*
- *A deep forest clearing with a mirrored pond reflecting a galaxy-filled night sky.*

Fig. 7 in the main text:

- *A corgi wearing cool sunglasses.*
- *Astronaut on Mars During sunset.*

Fig. 1 in Appendix:

- *Astronaut in a jungle, cold color palette, muted colors, detailed, 8k.*

Fig. 2 in Appendix:

- *A Renaissance noblewoman, portrayed in an elegant gown with intricate embroidery. Her expression is thoughtful, and her eyes are deep and insightful. The background is a lush Italian garden, reflecting the artistic style of the High Renaissance.*

Fig. 3 in Appendix:

- *Emma Watson as a powerful mysterious sorceress, casting lightning magic, detailed clothing.*

Fig. 5 in Appendix:

- *Envision a portrait of an elderly woman, her face a canvas of time, framed by a headscarf with muted tones of rust and cream. Her eyes, blue like faded denim. Her attire, simple yet dignified.*
- *A Renaissance noblewoman, portrayed in an elegant gown with intricate embroidery. Her expression is thoughtful, and her eyes are deep and insightful. The background is a lush Italian garden, reflecting the artistic style of the High Renaissance.*

Fig. 6 in Appendix:

- *Realistic oil painting of a stunning model merged in multicolor splash made of finely torn paper, eye contact, walking with class in a street.*
- *A painting of a beautiful graceful woman with long hair, a fine art painting, by Qiu Ying, no gradients, flowing sakura silk, beautiful oil painting.*
- *Katsushika Hokusai's Japanese depiction of a very turbulent sea with massive waves. The background s shows a beautiful dark night over a illuminated village. The colors are red and yellow, mood lighting Imagine a dreamlike scene blending the swirling cosmic colors of Vincent van Gogh's Starry Night with the surreal celestial precision of Salvador Dalí.*
- *Character of lion in style of saiyan, mafia, gangsta, citylights background, Hyper detailed, hyper realistic, unreal engine ue5, cgi 3d, cinematic shot, 8k.*
- *Santa Claus riding on top of a turkey, with very large bag of gifts, snow, ice, very cold place, realistic digital art, blurred background, expansive lighting, 4k, light gray and blue color palette, sharp and fine intricate details defined.*
- *Best Quality, Masterpiece, steampunk theme, centered, front cover of fashion magazine, concept art, design, magazine design, 1girl, cute, blonde ponytail hair, gothic steampunk dress, model pose, (epic composition, epic proportion), vibrant color, text, diagrams, advertisements, magazine title, typography.*
- *Burning pile of money, epic composition, digital painting, emotionally profound, thought-provoking, intense and brooding tones, high quality, masterpiece.*
- *A pastoral scene with shepherds, flocks, and rolling hills, in the tradition of a Jean-François Millet landscape.*
- *A painting of brooklyn new york 1940 storefronts, by John Kay, highly textured, rich colour and detail, ballard, deep colourś, style of raymond swanland, trio, oill painting, h 768, well worn, displayed, detailed 4 k oil painting, glenn barr, textured oil on canvas, looking cute.*
- *A swirling night sky filled with bright stars and a small village below, inspired by Vincent van Gogh's Starry Night.*
- *Portrait of a bear as a roman general, with a helmet, decorative, fantasy environment, oil painting, masterpiece, detailed, sharp, clear, cinematic lights.*
- *The Great Wall of China winding through mist-covered mountains, captured in the delicate brushwork and harmonious colors of a traditional Chinese landscape painting.*
- *RAW photo of a mountain lake landscape, clean water, 8k, UHD.*

Fig. 7 in Appendix:

- *A painting of a beautiful graceful woman with long hair, a fine art painting, by Qiu Ying, no gradients, flowing sakura silk, beautiful oil painting.*
- *Professor Snape brewing a potion in the dungeon, the room illuminated by the green glow of the cauldron.*
- *A Renaissance noblewoman, portrayed in an elegant gown with intricate embroidery. Her expression is thoughtful, and her eyes are deep and insightful. The background is a lush Italian garden, reflecting the artistic style of the High Renaissance.*
- *The Great Wall of China winding through mist-covered mountains, captured in the delicate brushwork and harmonious colors of a traditional Chinese landscape painting.*
- *Summer landscape, vivid colors, a work of art, grotesque, Mysterious.*

Fig. 8 in Appendix:

- *A Samoyed wearing a sunglasses, sticking out its tongue, dslr image, 8k.*
- *A Corgi wearing a sunglasses, sticking out its tongue, dslr image, 8k.*
- *A German Shepherd wearing a sunglasses, sticking out its tongue, on the grass, dslr image, 8k.*
- *A Husky wearing a sunglasses, sticking out its tongue, dslr image, 8k.*
- *An Australian Cattle Dog wearing a sunglasses, sticking out its tongue, style of Gian Lorenzo Bernini.*
- *A medieval knight standing in a lush forest, oil painting by Van Gogh.*
- *A gardener tending to a colorful, blooming garden, oil painting by Van Gogh.*
- *A robot exploring the ruins of an ancient civilization, watercolor by MORILAND.*
- *A ghost haunting an abandoned Victorian mansion, watercolor by MORILAND.*
- *An astronaut floating in space, watercolor by MORILAND.*

Fig. 9 in Appendix:

- *A cute corgi on the lawn.*
- *A portrait of Mr. Bean (Rowan Atkinson).*
- *Japanese Ukiyo-e, Kanagawa Surfing Sato.*
- *A cute panda on a tree trunk.*
- *A Portrait of Albus Dumbledore.*

- *A Chinese Painting of the Great Wall.*

References

[1] 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. 2, 6

[2] OpenAI. Chatgpt: Large-scale language models. https://www.openai.com/blog/chatgpt, 2022. 7

[3] 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. 2, 4, 6

[4] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *CVPR*, 2021. 6

[5] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 6