# Supplementary Material
# Customizing 360-Degree Panoramas through Text-to-Image Diffusion Models

Hai Wang[1]* Xiaoyu Xiang[2] Yuchen Fan[2] Jing-Hao Xue[1]
[1]University College London [2]Meta Reality Labs
{hai.wang.22, jinghao.xue}@ucl.ac.uk, {xiangxiaoyu, ycfan}@meta.com

## 1. Supplementary Content

This supplementary material begins by presenting additional ablation studies. Next, we showcase sample images from various scenes in our collected *360PanoI* dataset and generation process of their text prompts. Finally, more images synthesized using different methods are shown.

### 1.1. Additional Ablation Studies

**Order of Additional Denoising Operations Ablation.** In our proposed *StitchDiffusion* method, we additionally perform pre-denoising operations twice on the stitch block at each time step $t$ during the denoising process. In this experiment, we investigate the effect of the order in which these additional denoising operations are conducted, specifically at the beginning and at the end of denoising time step $t$. The corresponding results are shown in Figure 1. We can observe that: (1) if the additional denoising operations are performed twice on the stitch block at the end of denoising time step $t$, the synthesized image does not form a seamless 360-degree panorama; (2) however, when the additional denoising operations are conducted twice on the stitch block at the beginning of denoising time step $t$, the customized diffusion model effectively generates a seamless 360-degree panoramic image.

**Horizontal Sliding Distance Ablation.** To study the effect of horizontal sliding distance $\omega$ between adjacent cropped patches on the synthesized images, a comparison was carried out using various sliding distances. The visual results are presented in Figure 2. There is a noticeable seam in the middle of the region indicated by the red solid box when the sliding distance $\omega$ is set to 512. By reducing the sliding distance to 256, the noticeable seam is improved, but the local content inconsistency (ground and grass) still exists in the region now represented by the blue solid box. Finally, with a sliding distance of $\omega$ set to 128, the content within the region now marked by the green solid box exhibits seamless and consistent integration.



(a) additional denoising operations at the end of each time step



(b) additional denoising operations at the beginning of each time step

Figure 1. Ablation study on the order of additional denoising operations. The corresponding text prompt is '$V^*$, fantastical street, cinematic, anime style'. $J_0$ and $J_{syn}$ denote the clear denoised image and the final result, respectively. The leftmost and rightmost sides of $J_{syn}$ are not continuous when the additional denoising operations are performed twice at the end of denoising time step $t$.

**Poor Tags Ablation.** To assess the impact of poor tags within the input text prompt on the trigger word's effectiveness in controlling the image generation, we employed BLIP [4] to get the corresponding text prompts from real 360-degree panoramas. Subsequently, we conducted a comparison between images generated using these poor-tags-included text prompts with and without our trigger word. As shown in Figure 3, the presence of these poor tags in the text prompt hinders our trigger word's ability to control the generation of 360-degree panoramas by our method.

### 1.2. Sample Images and Their Text Prompts

The 360-degree panoramas in our collected *360PanoI* dataset have been sourced from Poly Haven [1]. To dis-

*Corresponding author. All experiments, data collection and processing were conducted in University College London.

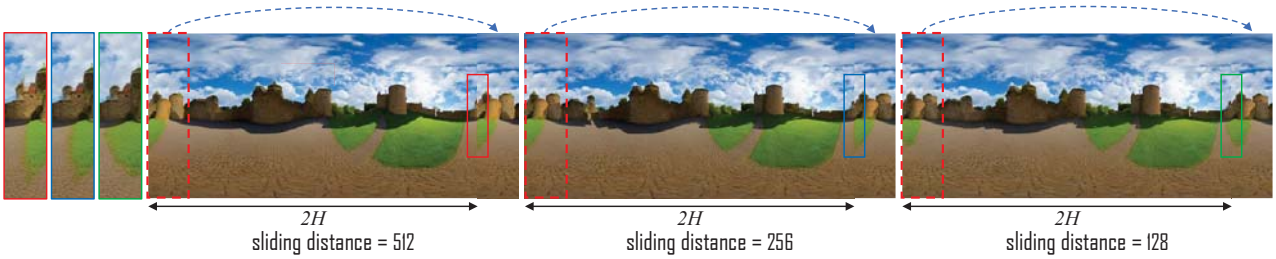


Figure 2. Ablation study on the horizontal sliding distance. The corresponding text prompt is '$V^*$, castle, a beautiful artwork illustration'. When the horizontal sliding distance $\omega$ in the *StitchDiffusion* is 128, the content in the green solid box is more consistent and seamless than those in the blue and red solid boxes for the sliding distances of 256 and 512.

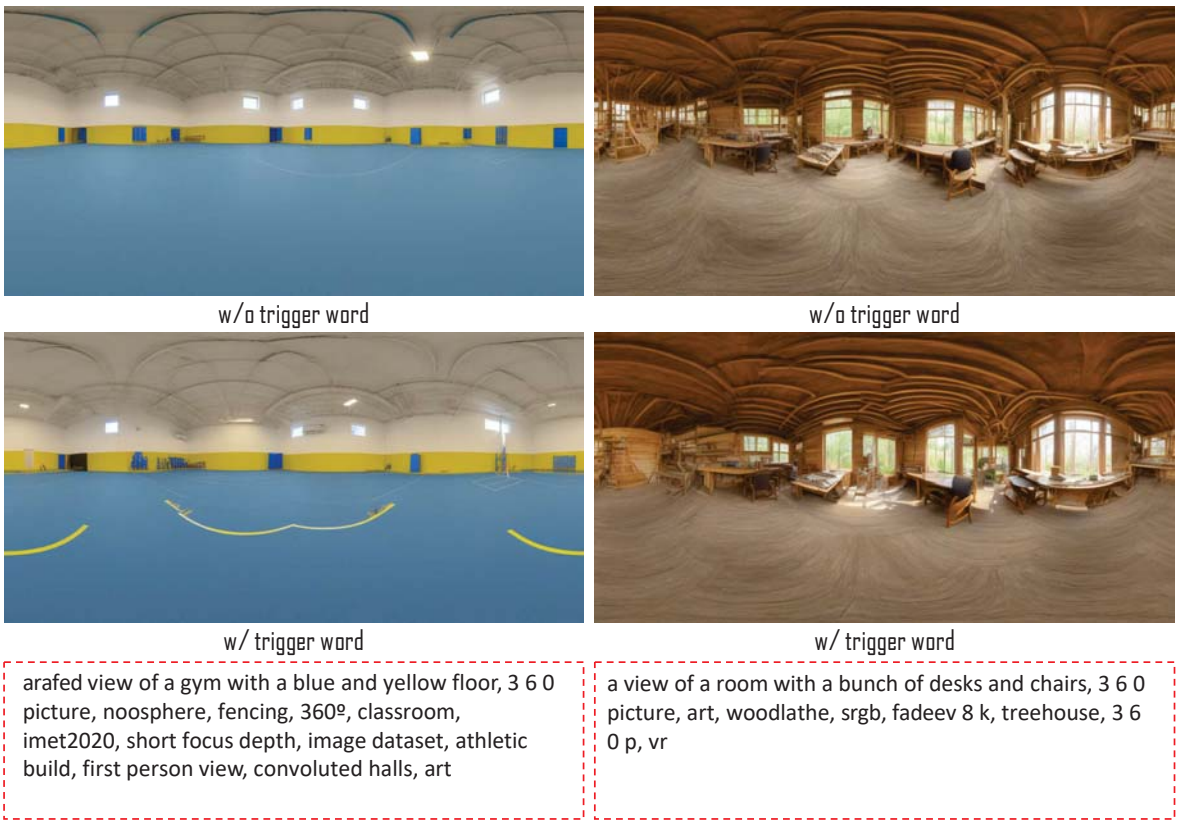


Figure 3. Ablation study on the effectiveness of the trigger word $V^*$ if poor tags in text prompts are not filtered out. We collected two real 360-degree panoramas from Poly Haven [1], which are independent of the *360PanoI* dataset, and utilized BLIP [4] to create corresponding text prompts denoted by the red dashed box. It is evident that without filtering out poor tags like '3 6 0 picture', the trigger word cannot effectively control our method's generation of 360-degree panoramas.

play the different scenes within our dataset, we randomly select one image from each scene. The corresponding sample images, with a resolution of $512 \times 1024$, are illustrated in Figure 4. Despite the *360PanoI* dataset only contains 8 scenes from the real world, it is important to note that the visual results presented in the main manuscript demonstrate the generalization capability of our customized diffusion model, utilizing the proposed *StitchDiffusion* technique, to generate 360-degree panoramas encompassing a wide variety of scenes unseen in the dataset. In addition, we provide a diagram in Figure 5 to demonstrate the generation process of the corresponding text prompts for these 360-degree panoramas within the *360PanoI* dataset.

### 1.3. More Visual Results

To highlight the distinctions between *MultiDiffusion* [2] and our proposed *StitchDiffusion*, we present a visual comparison between various schemes of *MultiDiffusion* and our *StitchDiffusion* in Figure 6. We can see that the combination of *MultiDiffusion* and our customized model in (a) fails

*indoor*

*nature*

*night*

*outdoor*

*skies*

*studio*

*sunrise-sunset*

*urban*

Figure 4. Sample images depicting the eight scenes contained within our collected *360PanoI* dataset are presented. In order to fine-tune a text-to-image diffusion model for customizing 360-degree panoramas, we utilize the entire set of 120 panoramic images from the dataset along with their corresponding text prompts acquired from BLIP [4].

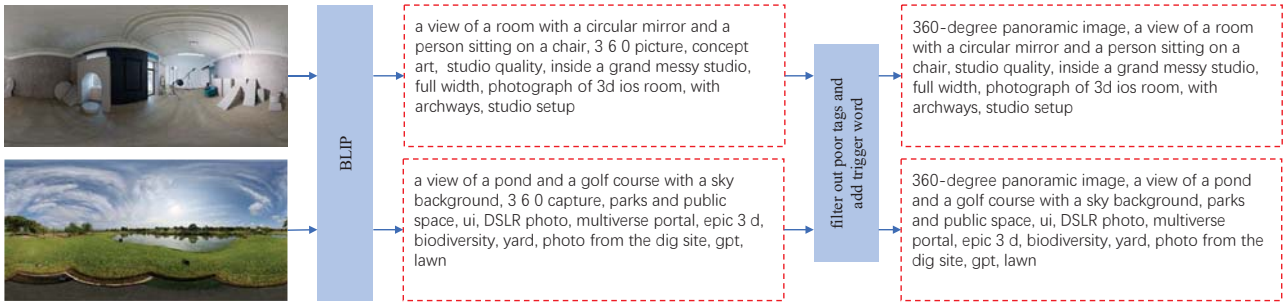


Figure 5. Diagram to show the generation process of text prompts in our *360PanoI* dataset using BLIP [4]. For the collected 120 360-degree panoramas, we employ BLIP to produce their text prompts. Then, we filter out poor tags such as '3 6 0 picture' and introduce a trigger word '360-degree panoramic image' into each text prompt, resulting in the final text prompts in our *360PanoI* dataset. Note that we only show 2 images from the 120 collected panoramas for illustration.

(a) *MultiDiffusion* w/ our customized model

(b) *MultiDiffusion* w/ our customized model + one time of additional denoising on stitch block

(c) *MultiDiffusion* w/ our customized model + two times of additional denoising on stitch block

(d) *StitchDiffusion* w/ our customized model

Figure 6. Visual results of different schemes. The corresponding text prompt for these generated images is '360-degree panoramic image, campus, unreal engine, studio quality, japanese anime style, anime'. Since the size of the cropped patch in *MultiDiffusion* [2] is $512 \times 512$, the stitch block here consists of the leftmost ($512 \times 256$) and rightmost ($512 \times 256$) regions in the image. We can see that the combination of *MultiDiffusion* and our customized model in (a) cannot generate a seamless 360-degree panorama. Even with one time of additional denoising applied to the stitch block in (b), or two times of additional denoising applied to the stitch block in (c), the results remains unseamless. In contrast, our method in (d) successfully synthesizes a seamless and plausible 360-degree panorama that aligns with the text prompt.

to generate a seamless 360-degree panorama. Even with one time of additional denoising applied to the stitch block in (b), or two times of additional denoising applied to the stitch block in (c), the results remains unseamless. In contrast, our method in (d) successfully synthesizes a seamless and plausible 360-degree panorama corresponding to the text prompt.

To further demonstrate the superiority of our method in generating 360-degree panoramas, we provide additional comparison results involving our method, latent diffusion combined with *MultiDiffusion* [2], and Text2Light [3], shown in Figure 7 and Figure 8. Our method outperforms the other two methods by producing seamless and plausible 360-degree panoramic images that correspond to the input text prompts. Moreover, to showcase the excellent generalizability of our proposed method, we present more synthesized 360-degree panoramas depicting various scenes in Figure 9 and Figure 10.

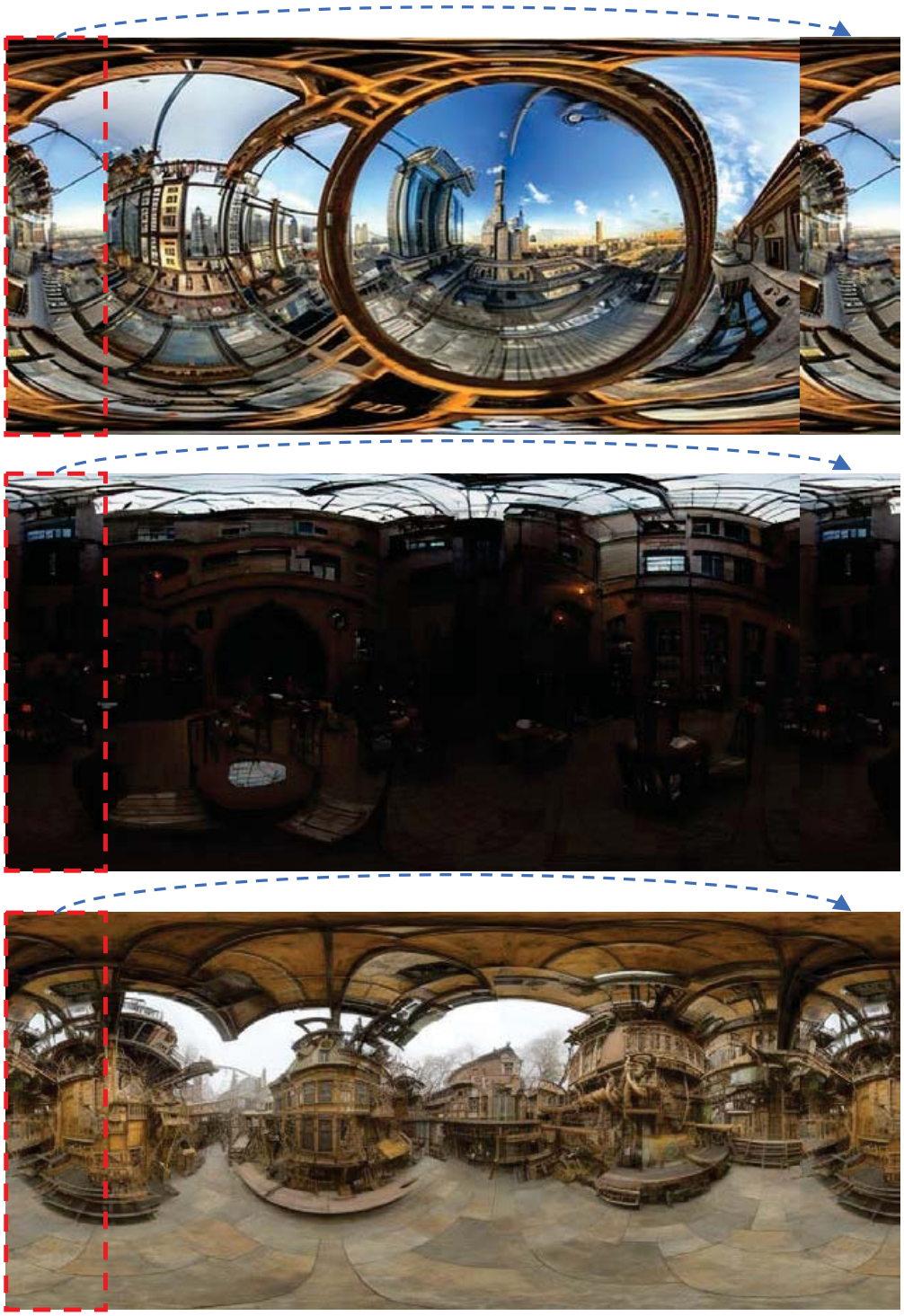


Figure 7. Visual comparison among *MultiDiffusion* [2] combined with latent diffusion (the first row), Text2Light [3] (the second row) and our method (the third row). To demonstrate the discontinuity or continuity between the leftmost and rightmost sides of the generated image, we copy the leftmost area ($512 \times 128$) represented by the red dashed box and paste it onto the rightmost side of the image. Our method generates a photorealistic and seamless 360-degree panoramic image compared to the other two methods.

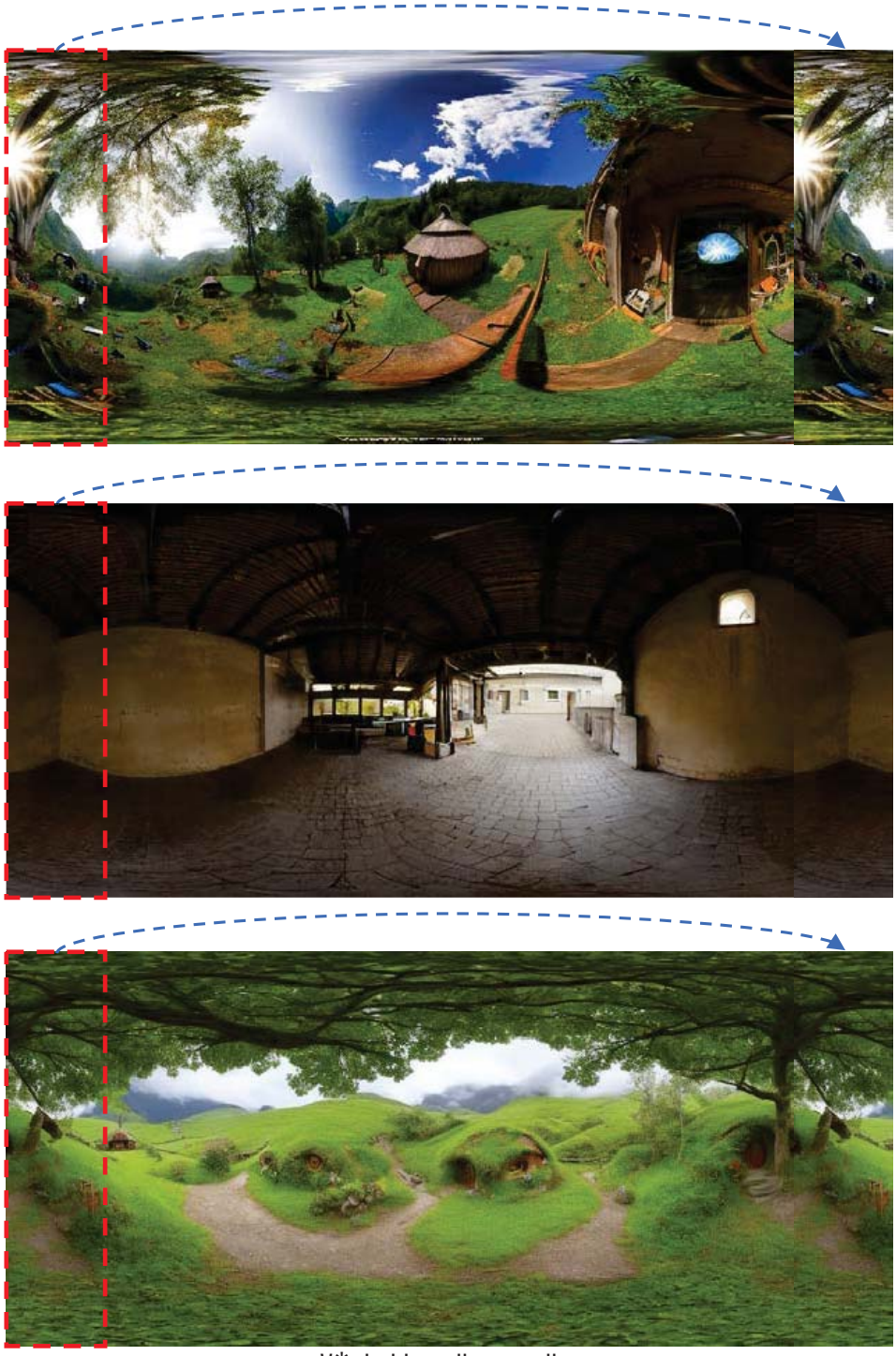


Figure 8. Visual comparison among *MultiDiffusion* [2] combined with latent diffusion (the first row), Text2Light [3] (the second row) and our method (the third row). To display the discontinuity or continuity between the leftmost and rightmost sides of the generated image, we copy the leftmost area ($512 \times 128$) represented by the red dashed box and paste it onto the rightmost side of the image. Compared to the two other approaches, our method excels in generating visual appealing and plausible 360-degree panoramas aligned with the input text prompts.

V*, cyberpunk building, mega structure, future

V*, outside view of a manor, digital art

Figure 9. The images generated by our method showcasing the themes of 'cyberpunk' and 'manor' are presented. To show the discontinuity or continuity between the leftmost and rightmost sides of the generated image, we copy the leftmost area ($512 \times 32$) represented by the red dashed box and paste it onto the rightmost side of the image. By carefully observing the illustrations, we can see that our method successfully captures the essence of the 'cyberpunk' and 'manor' themes, generating visually appealing 360-degree panoramic images.

V*, scotland indoor cabin, ancient style, hyper realistic

V*, library, japanese anime style, warm light

Figure 10. Illustration of 'cabin' and 'library' generated by our method. To display the discontinuity or continuity between the leftmost and rightmost sides of the generated image, we copy the leftmost area ($512 \times 32$) represented by the red dashed box and paste it onto the rightmost side of the image. The continuity between the leftmost and rightmost sides of the synthesized images is effectively maintained, ensuring a seamless transition and enhancing the overall immersive experience for viewers.

## References


[1] https://polyhaven.com/hdris. 1, 2

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 2, 4, 5, 6

[3] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 4, 5, 6

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 2, 3, 4