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

Abstract

Personalized text-to-image (T2I) synthesis based on diffusion models has attracted significant attention in recent research. However, existing methods primarily concentrate on customizing subjects or styles, neglecting the exploration of global geometry. In this study, we propose an approach that focuses on the customization of 360-degree panoramas, which inherently possess global geometric properties, using a T2I diffusion model. To achieve this, we curate a paired image-text dataset specifically designed for the task and subsequently employ it to fine-tune a pre-trained T2I diffusion model with LoRA. Nevertheless, the fine-tuned model alone does not ensure the continuity between the leftmost and rightmost sides of the synthesized images, a crucial characteristic of 360-degree panoramas. To address this issue, we propose a method called StitchDiffusion. Specifically, we perform pre-denoising operations twice at each time step of the denoising process on the stitch block consisting of the leftmost and rightmost image regions. Furthermore, a global cropping is adopted to synthesize seamless 360-degree panoramas. Experimental results demonstrate the effectiveness of our customized model combined with the proposed StitchDiffusion in generating high-quality 360-degree panoramic images. Moreover, our customized model exhibits exceptional generalization ability in producing scenes unseen in the fine-tuning dataset. Code is available at https://github.com/littlewhitesea/StitchDiffusion.

1. Introduction

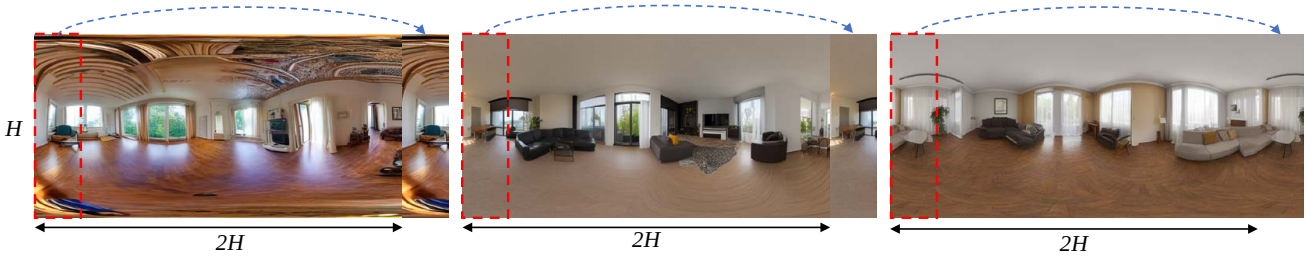
360-degree panoramic images [11, 33, 36, 54] are extensively utilized in virtual reality (VR) devices, such as head mount displays [3]. Unlike ordinary two-dimensional (2D) images, which have a limited viewing range, 360-degree panoramas encompass the entire $360° \times 180°$ field of view. This allows viewers to explore a scene from any angles, providing them with an immersive experience. The inherent globally geometric properties of 360-degree panoramas stem from this unique characteristic. There are various types of projections [50] used to represent 360-degree panoramas. In this paper, we specifically focus on the equirectangular projection (ERP), which represents the 360-degree panoramic image on a 2D surface. In this context, two essential properties of a 360-degree panorama arise: (1) the width of a 360-degree panoramic image is twice its height, and (2) the leftmost and rightmost sides of a 360-degree panorama are continuous.

Diffusion models [9, 10, 51] perform better in generating photorealistic and diverse images compared with generative adversarial networks (GANs) [8, 14], leading to increasing attention over the past two years. Thanks to their excellent generation quality and controllability, diffusion models have been widely explored for tackling numerous challenging tasks [4, 15, 18, 24, 30, 37, 45]. Notably, diffusion models applied to text-to-image (T2I) synthesis [29, 35, 37, 40] can produce high-quality images corresponding to descriptive text prompts, making them highly popular on social media. However, these models have limitations when it comes to synthesizing instances of customized concepts, such as a user's pet or personal item.

To handle this challenge, several personalized T2I generation algorithms [12, 17, 20, 38, 44] have been proposed. These algorithms enable the customization of T2I diffusion models by providing multiple images of a specific subject or concept, resulting in the synthesis of images containing the subject or concept in diverse contexts. Different from these existing personalized technologies [12, 20, 38] which focus on customizing specific subjects (e.g., dog, sunglasses) or styles (e.g., oil painting, pop art), our work aims to explore the customization of global geometry.

Specifically, we focus on customization of a T2I diffusion model for synthesizing 360-degree panoramas with inherent globally geometric properties. To begin, we build a paired image-text dataset called *360PanoI*. Due to limited computational resources and the need for fine-tuning efficiency, we employ the Low-Rank Adaptation (LoRA) [19, 39] technology to fine-tune a pre-trained T2I diffusion model using the collected *360PanoI* dataset. However, we encounter difficulties when generating 360-degree panora-

*Corresponding author



(a) *MultiDiffusion* w/ Stable Diffusion (b) *MultiDiffusion* w/ our customized model (c) *StitchDiffusion* w/ our customized model

Figure 1. Example results of three different methods with the text prompt 'V^*, *a living room with a couch and a table*', where V^* refers to the trigger word. To easily recognize the continuity or discontinuity between the leftmost and rightmost sides of the generated image, we copy the leftmost area indicated by the red dashed box and paste it onto the rightmost side of the image. The notation 'w/' denotes 'with'. Our customized model is achieved by fine-tuning Stable Diffusion [37]. Compared with *MultiDiffusion* [5] combined with Stable Diffusion [37] in (a), *MultiDiffusion* [5] with our customized model in (b) generates more visually appealing content. Moreover, in contrast to (b), our proposed *StitchDiffusion* with the customized model in (c) successfully synthesizes a seamless 360-degree panoramic image.

mas using the fine-tuned diffusion model, even with the use of *MultiDiffusion* [5], a recent method for producing traditional panoramas but disregarding the continuity between the leftmost and rightmost sides of the synthesized image (as shown in Figure 1). To address this issue, we put forward a tailored generation process named *StitchDiffusion* for synthesizing 360-degree panoramas. In the *StitchDiffusion* approach, we perform pre-denoising operations twice at each time step of the denoising process on the stitch block, which is constituted of the leftmost and rightmost image regions. After the denoising process is completed, we conduct a global cropping to produce the final image. This method guarantees that the fine-tuned T2I diffusion model generates seamless 360-degree panoramic images. Moreover, despite the limited number of scenes in our *360PanoI* dataset, the fine-tuned diffusion model demonstrates excellent generalization capabilities to unseen scenes. In other words, the fine-tuned diffusion model can successfully synthesize 360-degree panoramas of scenes not present in the fine-tuning dataset. This observation indicates that T2I diffusion models possess the potential to effectively capture and represent global geometry.

The contributions of this work can be summarized as follows: (1) We make the first attempt to explore the customization of 360-degree panoramas using T2I diffusion models, which is beneficial for employing T2I diffusion models in various application scenarios, such as indoor design and VR content creation. Our experimental results demonstrate that T2I diffusion models possess the capability to produce 360-degree panoramas with inherent geometric properties and generalize this ability to unseen scenes. (2) We propose a stitch method called *StitchDiffusion* as part of the generation process to synthesize seamless 360-degree panoramic images, which ensures the continuity between the leftmost and rightmost sides of the synthesized panoramas. (3) We curate a paired image-text dataset called *360PanoI* specifically for the synthesis of 360-degree panoramas. This dataset serves as a valuable resource for future studies and advancements in the field of 360-degree panoramic images.

2. Related Work

Text-to-Image Diffusion Models. Text-to-image (T2I) synthesis based on diffusion models [16, 29, 35, 37, 40] can generate images that align with the provided text prompts, which have showcased unprecedented levels of diversity and fidelity. We will only introduce several representative works here; for more comprehensive information, we refer readers to the survey paper [52]. GLIDE [29] stands out as a pioneering T2I diffusion model that uses classifier-free guidance in the T2I synthesis process. Different from GLIDE requiring to train its text encoder, Imagen [40] utilizes a pre-trained large transformer language model to encode textual input for image generation. Both GLIDE and Imagen operate in the pixel space, which demands substantial computational resources. To alleviate this requirement, Latent Diffusion Models (LDMs) [37] propose to train diffusion model in the latent space, significantly reducing the computational burden. In our work, we adopt Stable Diffusion [37], a variant of LDMs, as the foundational model due to its relatively lower demand for computing resources.

Personalized Text-to-Image Generation. Given one or multiple images of a specific subject or style provided by users, personalized text-to-image (T2I) generation [2, 12, 13, 17, 20, 38, 41, 42, 44, 46, 48] based on diffusion models aims to synthesize instances of the specific subject or style in diverse contexts. These personalized techniques can be broadly categorized into three groups. The first category is the *personalization-by-inversion* approach, initially explored in Textual Inversion [12]. This method optimizes an input vector in the textual embedding space to represent the desired subject or style. To enhance its expressive power, Extended Textual Inversion [46] and NeTI [2] propose optimizing multiple vectors and employing a neural mapper, respectively, resulting in stronger representations. DreamBooth [38], on the other hand, is a pioneering *personalization-by-fine-tuning* method. It introduces a class-specific prior preservation loss to mitigate language

drift [21, 28], and fine-tunes the entire T2I diffusion model for binding unique identifiers to user-provided subjects. In contrast, Custom Diffusion [20] and SVDiff [17] fine-tune only a small portion of parameters for improved efficiency. However, these approaches still require fine-tuning the diffusion model for each user-specific subject. Recognizing this limitation, *personalization-by-encoder* methods [13, 41, 48] have been proposed for rapid customization of T2I models. Specifically, these methods first train a mapping encoder, which is then used to directly map arbitrary input images into word embeddings representing the subject. Unlike the existing personalized approaches that concentrate on customizing specific subjects or artistic styles, our work in this paper explores the customization of global geometry, specifically 360-degree panoramic images. The successful customization of such complex geometries would demonstrate the inherent ability of T2I diffusion models to capture intricate spatial representations.

Panorama Generation. GAN-based panorama generation algorithms [7, 25, 26, 31, 43, 47, 49] have been extensively studied. In contrast, Text2Light [6] adopts a text-conditioned global sampler and structure-aware local sampler to generate panoramic images by sampling from a dual-codebook representation. Recently, diffusion models have also shown promising results in panorama synthesis [5, 22, 53]. DiffCollage [53] utilizes a semantic segmentation map as the condition for the diffusion model and generates 360-degree panoramas based on a complex factor graph. On the other hand, PanoGen [22] employs Stable Diffusion [37] to synthesize new indoor panoramic images with a recursive image outpainting technology based on multiple text descriptions. Distinguishing itself from PanoGen [22], *MultiDiffusion* [5] simultaneously samples the panoramic image through blending diffusion paths of all overlapped cropped patches to synthesize high-quality images. However, *MultiDiffusion* [5] does not guarantee the continuity between the leftmost and rightmost sides of the generated image, which is a natural property of the 360-degree panorama. To deal with this problem, we propose in this paper a method called *StitchDiffusion*. This approach leverages our customized diffusion model to synthesize panoramas that exhibit continuity between the leftmost and rightmost sides, resulting in a seamless viewing experience. Moreover, we demonstrate in this paper that our customized diffusion model possesses strong generalization capabilities, allowing it to generate a wide range of 360-degree panoramas in various contexts, even for scenes not present in the fine-tuning dataset.

3. Methodology

In this section, we first briefly review the use of *MultiDiffusion* [5] to synthesize panoramic images. Then, we describe the process of customizing a pre-trained T2I diffusion model for 360-degree panoramas. Finally, we introduce our proposed method, *StitchDiffusion*, which is able to handle the issue of discontinuity between the leftmost and rightmost sides of the generated 360-degree panoramic image when using *MultiDiffusion* [5].

3.1. Preliminaries

Given a pre-trained T2I diffusion model Γ, the sequential denoising process of this model, gradually from a noisy image I_T to the final clean image I_0, could be expressed as

$$I_{t-1} = \Gamma(I_t, \delta)\,, t \in \{T, T-1, \cdots, 1\}\ , \tag{1}$$

where I is in the image space $\mathcal{I} = \mathbb{R}^{H\times W\times C}$, and δ is the textual embedding of a text prompt. A target of *MultiDiffusion* [5] is to generate a panoramic image aligning with the given text prompt without the need for any training or fine-tuning of the diffusion model Γ, which serves as a reference model. To this end, *MultiDiffusion* [5] defines a different T2I diffusion model Ω called *MultiDiffuser*, which operates in an image space $\mathcal{J}$. Its sequential denoising process is

$$J_{t-1} = \Omega(J_t, \delta)\,, t \in \{T, T-1, \cdots, 1\}\ , \tag{2}$$

where J is in the image space $\mathcal{J} = \mathbb{R}^{H\times W'\times C}$, and W' is greater than or equal to W.

To establish a connection between the target image space $\mathcal{J}$ and reference image space $\mathcal{I}$, *MultiDiffusion* [5] further defines n mappings $F_i : \mathcal{J} \to \mathcal{I}$, where $i \in \{1, 2, \cdots, n\}$. At time step t of the denoising process, $F_i(J_t) \in \mathcal{I}$ is the i-th $H \times W$ cropped patch of image J_t. These n overlapped cropped patches cover the whole image J_t, illustrated in Figure 2(c). The value of n is determined by

$$n = \frac{W' - W}{\omega} + 1\,, \tag{3}$$

where ω denotes the horizontal sliding distance between adjacent cropped patches. Using these mappings, the new denoising process of diffusion model Ω at time step t is achieved by solving the following optimization problem:

$$\Omega(J_t, \delta) = \arg\min_{J\in\mathcal{J}} \sum_{i=1}^{n} \|F_i(J) - \Gamma(F_i(J_t), \delta)\|^2\,. \tag{4}$$

In fact, this is a least-square problem and the corresponding closed-form solution is

$$\Omega(J_t, \delta) = \sum_{i=1}^{n} \frac{F_i^{-1}(\mathbf{1})}{\sum_{j=1}^{n} F_j^{-1}(\mathbf{1})} \otimes F_i^{-1}(\Gamma(F_i(J_t), \delta))\,, \tag{5}$$

where $\mathbf{1}$ is in the image space $\mathcal{I} = \mathbb{R}^{H\times W\times C}$, and its all pixel values are equal to 1. By leveraging this new denoising process, *MultiDiffuser* Ω can directly utilize the pre-trained T2I diffusion model Γ without any training or fine-tuning steps to generate panoramic images aligned with the given text prompt.

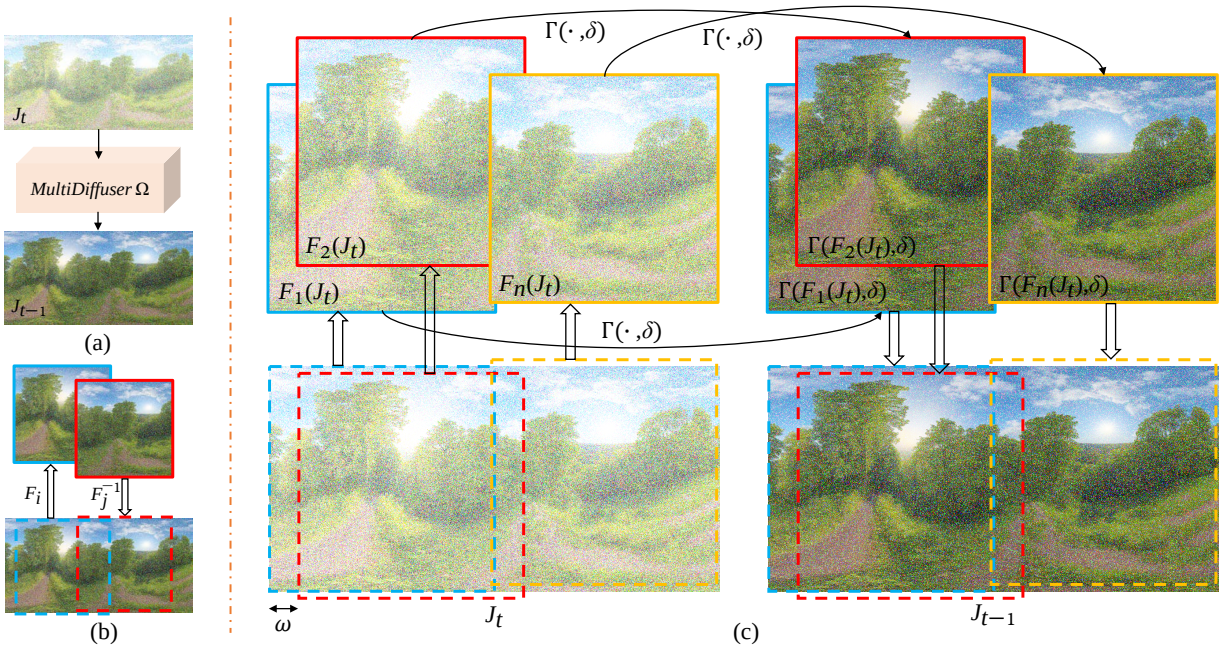


Figure 2. Overview of *MultiDiffusion* [5] for panorama generation. (a) The input J_t and output J_{t-1} of the *MultiDiffuser* Ω at time step t during the denoising process. (b) Illustration of the mapping F_i (directly cropping a patch from an image) and inverse mapping F_j^{-1}, where $i, j \in \{1, 2, \cdots, n\}$. (c) Detailed inner process of the *MultiDiffuser* Ω at time step t during the denoising process. Here, $\Gamma(\cdot, \delta)$ denotes the pre-trained T2I diffusion model with the textual embedding δ from a given text prompt, and ω is the horizontal sliding distance between adjacent cropped patches. Note that *MultiDiffusion* cannot guarantee the continuity between the leftmost and rightmost sides of generated panoramic images.

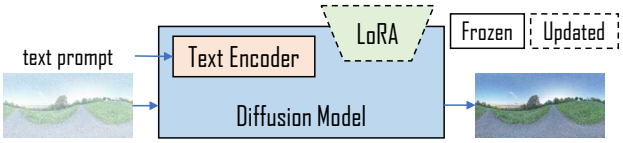


Figure 3. Illustration of customizing a T2I diffusion model with LoRA [19] for synthesizing 360-degree panoramic images. The paired image-text data used during the fine-tuning process are from our collected *360PanoI* dataset.

3.2. Customizing Models with LoRA

To customize a pre-trained T2I diffusion model for 360-degree panorama synthesis, we start by collecting a dataset of 360-degree panoramic images. Then, these images are tagged using BLIP [23], creating a paired image-text dataset called *360PanoI*. More detailed information about the collected dataset can be found in Section 4.1. For the fine-tuning process, we employ Low-Rank Adaptation (LoRA) [19] technology, which was initially proposed for fine-tuning large language models. Specifically, LoRA introduces trainable rank decomposition matrices into the pre-trained model, allowing for faster adaptation to downstream tasks with lower computational requirements compared to full fine-tuning. Recent work [39] has validated the effectiveness of LoRA in pre-trained T2I diffusion models. Considering its efficiency and low demand for computational resources, we employ LoRA to fine-tune the pre-trained diffusion model for generating 360-degree panoramic images using the *360PanoI* dataset, as shown in Figure 3.

Given the ground-truth image I^{gt} and its corresponding textual embedding δ, the preliminary customized model Γ_θ with LoRA is fine-tuned by using the loss function L_{pano} to denoise a variably-noised image $\alpha_t I^{gt} + \sigma_t \boldsymbol{\epsilon}$ as follows:

$$L_{pano} = \mathbb{E}_{I^{gt},\delta,\boldsymbol{\epsilon},t}\left[\gamma_t \|\Gamma_\theta(\alpha_t I^{gt} + \sigma_t \boldsymbol{\epsilon}, \delta) - I^{gt}\|^2\right], \quad (6)$$

where θ refers to the trainable matrices of LoRA, $\boldsymbol{\epsilon}$ and γ_t represent the noise following a Gaussian distribution and the functions of diffusion process time t, respectively, and α_t and σ_t are terms used to manage the noise schedule and the sample quality, respectively. Upon completing the fine-tuning process, we obtain the final customized diffusion model denoted as $\Gamma_{\hat{\theta}}$, where $\hat{\theta}$ denotes the updated parameters of LoRA.

3.3. StitchDiffusion for 360-degree Panoramas

Firstly, let us review two natural properties of a 360-degree panorama represented by the equirectangular projection [50]: (1) the width W of the 360-degree panoramic image is twice its height H, resulting in a final generated panorama size of $H \times 2H$; (2) there should be continuity between the leftmost and rightmost sides of the 360-degree panorama. However, as shown in Figure 2, the *MultiDiffusion* method fails to ensure this continuity. To solve this

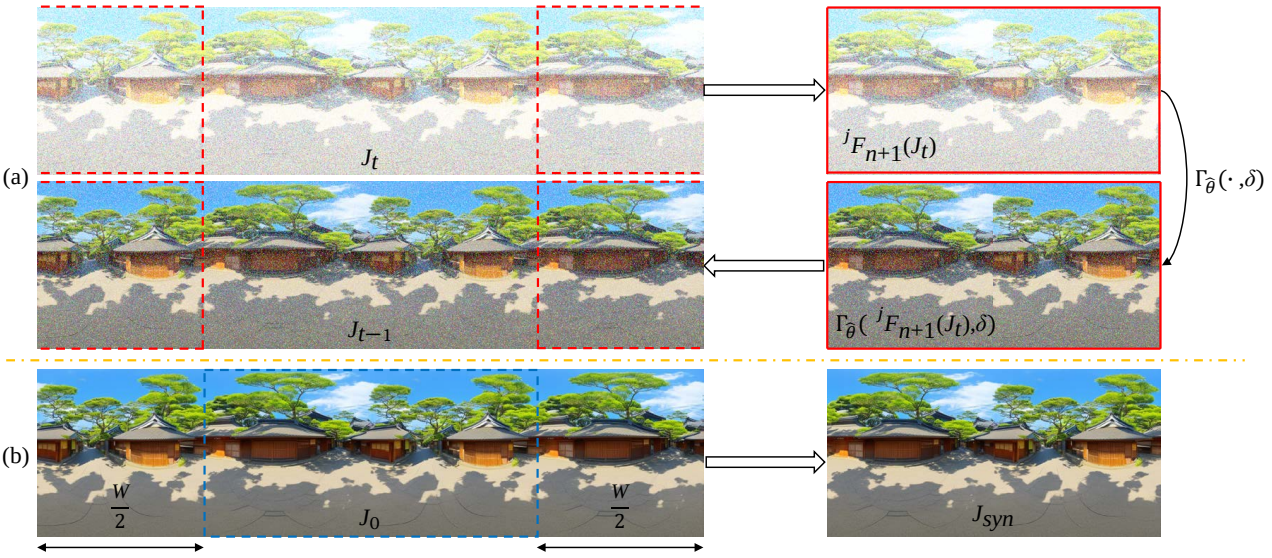


Figure 4. Overview of our proposed *StitchDiffusion* for generating 360-degree panoramas. (a) At each time step t of the denoising process, the $H \times W$ stitch block undergoes pre-denoising operations twice, which is constituted by the leftmost ($H \times \frac{W}{2}$) and rightmost ($H \times \frac{W}{2}$) regions of the image J_t. Here, the value of W is twice that of H. (b) The global cropping denoted by the blue dashed box of the final clear result J_0 is taken to achieve the 360-degree panorama J_{syn}. Note that if image J_0 ($H \times 4H$) is divided horizontally along the middle into two equal parts, then the left half ($H \times 2H$) of J_0 is identical to the right half ($H \times 2H$) of J_0, which could ensure the continuity between the leftmost and rightmost sides of the J_{syn} obtained from global cropping.

problem and guarantee seamless 360-degree panoramas, we put forward a new generation process called *StitchDiffusion*.

Specifically, we utilize the *MultiDiffuser* Ω to generate a panoramic image J with a resolution of $H \times (2H + 2H)$ using the customized diffusion model $\Gamma_{\hat{\theta}}$, that is, W' in Equation 3 is equal to $4H$. In addition, for *MultiDiffuser* Ω at time step t of denoising process in Figure 2, we additionally employ the customized diffusion model $\Gamma_{\hat{\theta}}$ to perform pre-denoising operations twice on a stitch block, which consists of the leftmost ($H \times \frac{W}{2}$) and rightmost ($H \times \frac{W}{2}$) regions of the current noisy image J_t, as illustrated in Figure 4(a). Here, W is twice the value of H. In this situation, the corresponding denoising process at time step t of our proposed *StitchDiffusion* can be expressed as

$$J_{t-1} = \sum_{j=1}^{2} \frac{{}^jF_{n+1}^{-1}(\mathbf{1})}{\Lambda} \otimes {}^jF_{n+1}^{-1}(\Gamma_{\hat{\theta}}({}^jF_{n+1}(J_t), \delta)) + \sum_{i=1}^{n} \frac{F_i^{-1}(\mathbf{1})}{\Lambda} \otimes F_i^{-1}(\Gamma_{\hat{\theta}}(F_i(J_t), \delta)) , \tag{7}$$

where ${}^jF_{n+1}(\cdot)$ and ${}^jF_{n+1}^{-1}(\cdot)$ denote the j-th additional mapping and inverse mapping of the stitch block, respectively, Λ denotes ${}^1F_{n+1}^{-1}(\mathbf{1}) + {}^2F_{n+1}^{-1}(\mathbf{1}) + \sum_{j=1}^{n} F_j^{-1}(\mathbf{1})$. Using the denoising process of our *StitchDiffusion*, we can get a clear image J_0 with a resolution of $H \times (2H + W)$ at the end of the entire denoising process. To obtain the final 360-degree panoramic image J_{syn} with a resolution of $H \times 2H$, we perform a global cropping operation on J_0:

$$J_{syn} = J_0[\frac{W}{2} : -\frac{W}{2}] , \tag{8}$$

illustrated in Figure 4(b). This operation ensures that the leftmost and rightmost sides of the panoramic image J_{syn} are continuous, as desired for a 360-degree panorama.

4. Experiments

4.1. Dataset and Implementation Details

Dataset. We collected 120 360-degree panoramic images in the real world from Poly Haven [1]. The images were sourced from a variety of scenes, including *indoor*, *nature*, *night*, *outdoor*, *skies*, *studio*, *sunrise-sunset*, and *urban* settings. Each scene consists of 15 panoramas. Due to limited computational resources, we performed an 8x rescale operation on these images using bilinear interpolation to obtain 360-degree panoramas with a resolution of 512×1024 pixels. Subsequently, we utilized BLIP [23] to tag these processed images. However, the generated text prompts contained poor tags such as '3 6 0 picture', which might potentially impact the fine-tuning process. Therefore, we removed these tags. Additionally, we introduced a trigger word '360-degree panoramic image', denoted as V^* in this paper, into each text prompt. Finally, our *360PanoI* dataset is constituted of 120 360-degree panoramas with a resolution of 512×1024 pixels, along with their corresponding text prompts. We present one sample image for each scene in the supplementary material.

Implementation Details. To customize a T2I diffusion model for 360-degree panorama synthesis using the *360PanoI* dataset, we employed Stable Diffusion 2-1 [37] and LoRA [19]. In detail, the LoRA architecture consists

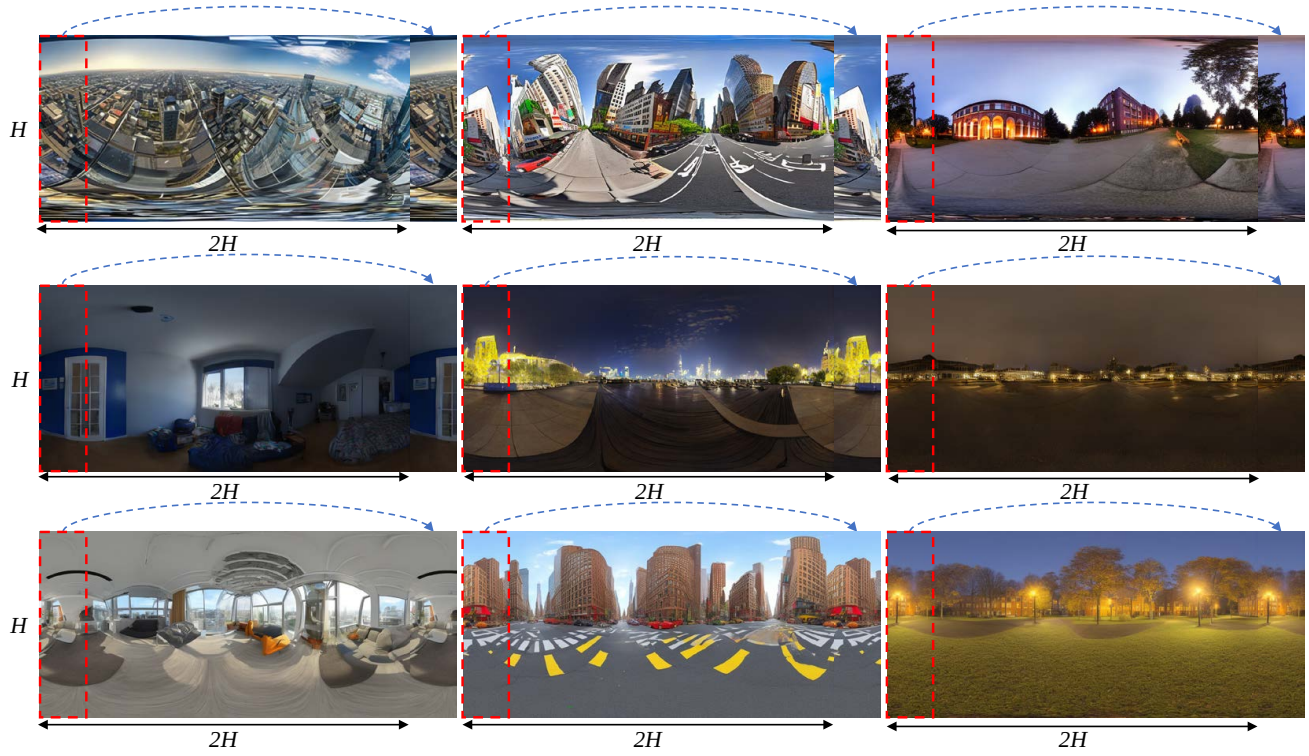


Figure 5. Visual comparison among *MultiDiffusion* [5] combined with Stable Diffusion [37] (the first row), Text2Light [6] (the second row) and our method (the third row). The corresponding text prompts of these images from left to right are 'V^*, futuristic flat, concept art', 'V^*, cartoon new york street', and 'V^*, university campus, foggy night', respectively. To demonstrate the discontinuity or continuity between the leftmost and rightmost sides of the generated image, we copy the left area denoted by the red dashed box and paste it onto the rightmost side of the image. We can see that our method generates seamless and plausible 360-degree panoramas corresponding to the input text prompts. For more comparison results, please refer to the supplementary material.

Table 1. Quantitative comparison between our method and Stable Diffusion (SD) [37] combined with *MultiDiffusion* [5]. Our method consisting of the customized model and *StitchDiffusion* is superior to the baseline in terms of both CLIP-score and FID.

Method	CLIP-score↑	FID↓
SD+*MultiDiffusion*	0.752±0.023	177.886±6.478
Ours	0.768±0.005	160.960±6.431

of two linear layers with an intermediate dimension of 32. During fine-tuning, we used a batch size of 2 and set the learning rate for the T2I diffusion model to 1e-4. The fine-tuning process was performed for 10 epochs using AdamW [27], which took approximately 40 minutes to complete. In the inference stage, we set the values of H and W to 512 and 1024, respectively, while the horizontal sliding distance ω between adjacent cropped patches was set to 128 in image space. It is important to highlight that the practical implementation of our *StitchDiffusion* process operates on J and I within latent space. That means the values of H, W and ω in latent space are 64, 128 and 16, respectively. All experiments were conducted on a single Tesla T4 GPU.

4.2. Comparisons

Due to the absence of a direct approach based on diffusion model for producing 360-degree panoramic images from an input text prompt, we adopt *MultiDiffusion* [5], a state-of-the-art method to generate normal panoramas, in combination with Stable Diffusion [37] as a baseline. Furthermore, we compare our approach with Text2Light [6], a state-of-the-art non-diffusion-based technique. The visual results are presented in Figure 5. We can see that the baseline method yields images with unsatisfactory content. Notably, the leftmost and rightmost sides of the images generated by the baseline are not continuous, indicating its inability to synthesize 360-degree panoramas. While Text2Light achieves improved continuity in synthesized images, it fails to capture the essence of 'futuristic' and 'cartoon' themes in the text prompts. In contrast, our proposed method consisting of the customized diffusion model and *StitchDiffusion* produces seamless and plausible 360-degree panoramic images corresponding to the text prompts.

To further quantitatively assess the plausibility of images generated by different methods, we first collected additional 20 real 360-degree panoramas from Poly Haven [1] as our ground truth, and then applied the same processing methodology outlined in Section 4.1 to acquire their text prompts. With these text prompts in hand, we attempted to generate the corresponding images using Text2Light. However, we found that some text prompts exceeded the token limit

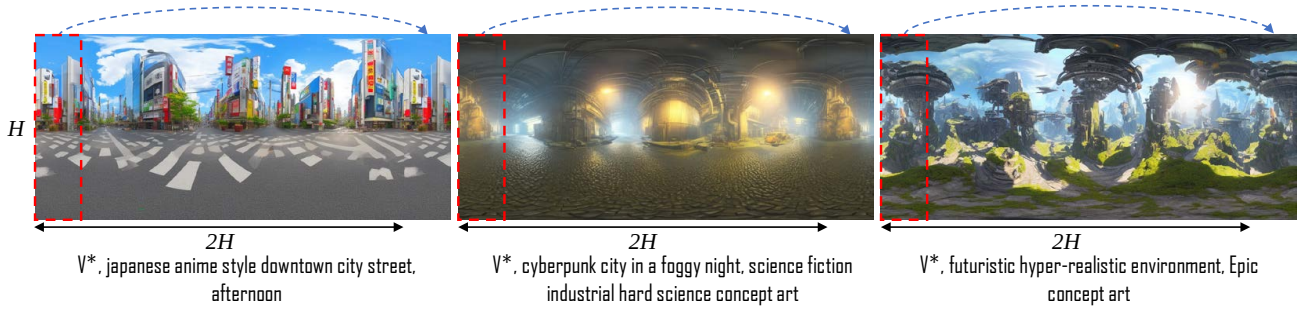
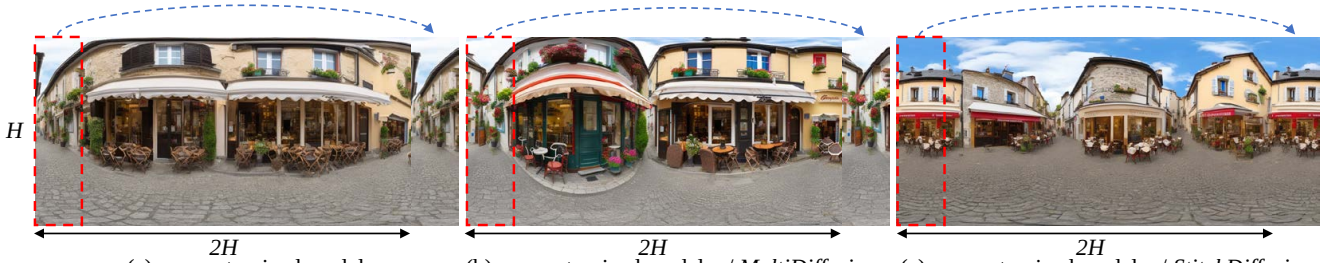


Figure 6. Visual results of unseen scenes synthesized by our method. Despite the fact that the collected *360PanoI* dataset only contains 8 scenes from the real world, our customized model with the proposed *StitchDiffusion* effectively produces 360-degree panoramas of diverse unseen scenes, showcasing its excellent generalization ability. More generated images can be seen in the supplementary material.



(a) our customized model (b) our customized model w/ *MultiDiffusion* (c) our customized model w/ *StitchDiffusion*

Figure 7. Ablation study on *StitchDiffusion*. The corresponding text prompt of these images is 'V^*, traditional french cafe in the street, small village'. The customized model cannot ensure continuity between the leftmost and rightmost sides of the generated image, even with *MultiDiffusion* [5]. In contrast, *StitchDiffusion* enables the customized model to generate a 360-degree panorama.

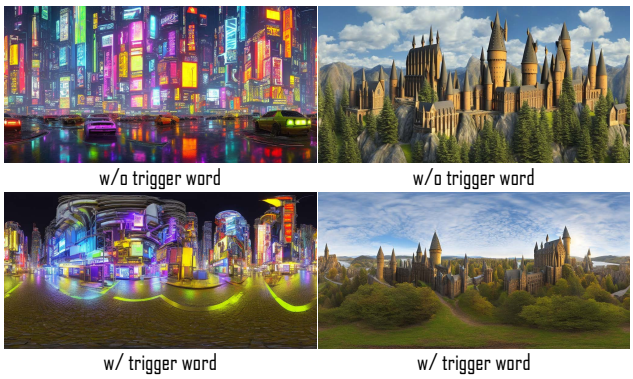


Figure 8. Ablation study on the trigger word V^*. The corresponding text prompts of left and right images in the second row are 'V^*, cyberpunk city, neon lights, science fiction' and 'V^*, Hogwarts campus, hyper realistic', respectively. With the trigger word included in the text prompt, the generated images cover the entire $360^\circ \times 180^\circ$ field of view. Here, 'w/o' and 'w/' refer to 'without' and 'with', respectively.

of Text2Light, preventing Text2Light from synthesizing images with them as inputs. In contrast, both our method and Stable Diffusion [37] combined with *MultiDiffusion* [5] can handle all text prompts of the ground truth panoramas. For a fair comparison, we only synthesized the corresponding images using our method and Stable Diffusion [37] combined with *MultiDiffusion* [5].

Now, with these generated images and their corresponding ground truth, we can calculate the quantitative results for the two methods. Specifically, we randomly cropped 1000 patches of size 512×512 from the 20 ground truth images and recorded the locations of each patch. Using these recorded locations, we similarly cropped corresponding 1000 patches from the generated images. Next, we employed the image encoder of CLIP [34] to extract embeddings of these patches. We then calculated the average cosine similarity between the embeddings of the generated patches and the real patches, denoted as CLIP-score. Additionally, we utilized the Frechet Inception Distance (FID) [32] to quantify the distance between the distribution of generated patches and the distribution of real patches. To further verify the effectiveness and robustness of our method in generating plausible images, we repeated the generation process 10 times, and then calculated the corresponding mean and standard deviation of the two metrics. The quantitative results, as shown in Table 1, indicate that our method outperforms the baseline in terms of the two metrics.

4.3. Generalizability

To evaluate the generalization ability of our customized diffusion model to unseen scenes, we fed a variety of text prompts describing scenes not included in the *360PanoI* dataset into the model. The corresponding generated images are presented in Figure 6. It is evident that our customized diffusion model using the proposed *StitchDiffusion* produces visually appealing 360-degree panoramas of di-

Figure 9. Ablation study on the number of pre-denoising times. The corresponding text prompt is 'V^*, an old city close up, sharp focus'. Although leftmost and rightmost sides of the generated image are continuous when conducting the pre-denoising operation on the stitch block only once at time step t of our *StitchDiffusion*, local content inconsistency in the red solid box appears in the image. Increasing the number of pre-denoising operations can improve the local content consistency (see the blue and green solid boxes). Considering a trade-off between computational efficiency and image quality, we choose to perform the pre-denoising operations twice on the stitch block.

verse unseen scenes, such as Japanese anime style, cyberpunk, and hyper-realistic environment. Remarkably, these results are achieved even though the collected *360PanoI* dataset only contains 8 scenes from the real world, which highlights the excellent generalizability of our customized T2I diffusion model.

4.4. Ablation Studies

We only present a subset of our ablation studies here; other ablation studies are in the supplementary material.

***StitchDiffusion* Ablation.** To study the effect of our proposed *StitchDiffusion* method on the generated results, we conducted a comparative analysis. Specifically, we compared the images synthesized by our customized diffusion model with and without *StitchDiffusion*. In addition, we introduced the results produced by combining the customized diffusion model with *MultiDiffusion* [5] for a more comprehensive comparison. The generated images are displayed in Figure 7. We can observe that the customized diffusion model alone is unable to synthesize 360-degree panoramas, primarily due to the limited capability of the diffusion model to capture and represent the continuous properties of these images. While *MultiDiffusion* demonstrates effectiveness in generating ordinary panoramic images, it also encounters difficulties in ensuring continuity between the leftmost and rightmost sides of the generated images. However, by incorporating our designed *StitchDiffusion* method, the customized model accurately synthesizes seamless 360-degree panoramic images.

Trigger Word Ablation. To investigate the impact of the trigger word V^* on the synthesis process, we conducted a comparison between images generated by our method with and without the trigger word included in the input text prompts. The visual results are shown in Figure 8. We can see that the trigger word V^* plays a crucial role in the generated image. When the trigger word is omitted from the text prompt, the resulting image fails to encompass the entire field of view spanning 360 degrees horizontally and 180 degrees vertically. Conversely, when the text prompt includes the trigger word, our customized model with *StitchDiffusion* successfully produces 360-degree panoramas.

Number of Pre-Denoising Times Ablation. In our *StitchDiffusion* method, we perform pre-denoising operations twice on the stitch block at each time step t of the denoising process, as depicted in Figure 4. To explore the impact of the number of pre-denoising operations on the synthesized images, we conducted an experiment comparing the results obtained with different numbers of pre-denoising operations at each time step t. The results are presented in Figure 9. We can see that when we only conduct one pre-denoising operation, despite the leftmost and rightmost sides of the generated image are continuous, there is a local content inconsistency exampled by the red solid box in the image. However, by conducting two or three pre-denoising operations on the stitch block, our customized model effectively generates high-quality 360-degree panoramic images without noticeable local content inconsistency. For a higher computational efficiency, we adopt the approach of performing pre-denoising operations twice in our method.

5. Conclusion

In this study, we have explored the customization of a T2I diffusion model for generating 360-degree panoramas. Our approach involved the establishment of a paired image-text dataset called *360PanoI*, followed by fine-tuning Stable Diffusion using LoRA. However, the fine-tuned diffusion model alone falls short in ensuring the continuity between the leftmost and rightmost sides of the generated images. To address this limitation, we proposed a method called *StitchDiffusion*, which successfully enables the customized diffusion model to synthesize seamless 360-degree panoramas. Through extensive experiments, we have verified the effectiveness of the proposed method and demonstrated that our customized diffusion model exhibits exceptional generalization ability, producing diverse and high-quality 360-degree panoramic images even in previously unseen scenes. The applications of our work are vast, particularly in fields such as indoor design, game and VR content creation, where the utilization of 360-degree panoramas is prevalent. Moreover, the *360PanoI* dataset we collected will be beneficial for any future investigations into 360-degree panoramic images.

References

[1] https://polyhaven.com/hdris. 5, 6, 11, 12

[2] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *arXiv preprint arXiv:2305.15391*, 2023. 2

[3] Christoph Anthes, Rubén Jesús García-Hernández, Markus Wiedemann, and Dieter Kranzlmüller. State of the art of virtual reality technology. In *2016 IEEE aerospace conference*, pages 1–19. IEEE, 2016. 1

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1

[5] 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, 3, 4, 6, 7, 8, 11, 14, 15, 16

[6] 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. 3, 6, 14, 15, 16

[7] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11431–11440, 2022. 3

[8] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 1

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[11] Ricardo Eiris, Masoud Gheisari, and Behzad Esmaeili. Pars: Using augmented 360-degree panoramas of reality for construction safety training. *International journal of environmental research and public health*, 15(11):2452, 2018. 1

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2

[13] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 2, 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[15] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems*, 2022. 1

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[17] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 1, 2, 3

[18] 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. 1

[19] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 1, 4, 5

[20] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2, 3

[21] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4385–4395, 2019. 3

[22] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *arXiv preprint arXiv:2305.19195*, 2023. 3

[23] 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. 4, 5, 11, 12, 13, 14

[24] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 1

[25] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4512–4521, 2019. 3

[26] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. Infinitygan: Towards infinite-pixel image synthesis. *arXiv preprint arXiv:2104.03963*, 2021. 3

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[28] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift

with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR, 2020. 3

[29] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804, 2022. 1, 2

[30] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 1

[31] Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 352–371. Springer, 2022. 3

[32] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 7

[33] Hai Chien Pham, N Dao, Akeem Pedro, Quang Tuan Le, Rahat Hussain, Sungrae Cho, and CSIK Park. Virtual field trip for mobile construction safety education using 360-degree panoramic virtual reality. *International Journal of Engineering Education*, 34(4):1174–1191, 2018. 1

[34] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[36] KA Ritter III and Terrence L Chambers. Three-dimensional modeled environments versus 360 degree panoramas for mobile virtual reality training. *Virtual Reality*, 26(2):571–581, 2022. 1

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 5, 6, 7, 14, 15, 16

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 2

[39] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023. 1, 4

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[41] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2, 3

[42] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023. 2

[43] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10521–10530, 2019. 3

[44] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. *arXiv preprint arXiv:2305.01644*, 2023. 1, 2

[45] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 1

[46] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2

[47] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 477–492. Springer, 2022. 3

[48] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2, 3

[49] Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, 2022. 3

[50] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020. 1, 4

[51] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022. 1

[52] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 2

[53] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel genera-

tion of large content with diffusion models. *arXiv preprint arXiv:2303.17076*, 2023. 3

[54] Yun Zhang, Fang-Lue Zhang, Zhe Zhu, Lidong Wang, and Yao Jin. Fast edit propagation for 360 degree panoramas using function interpolation. *IEEE Access*, 10:43882–43894, 2022. 1

A. Supplementary Content

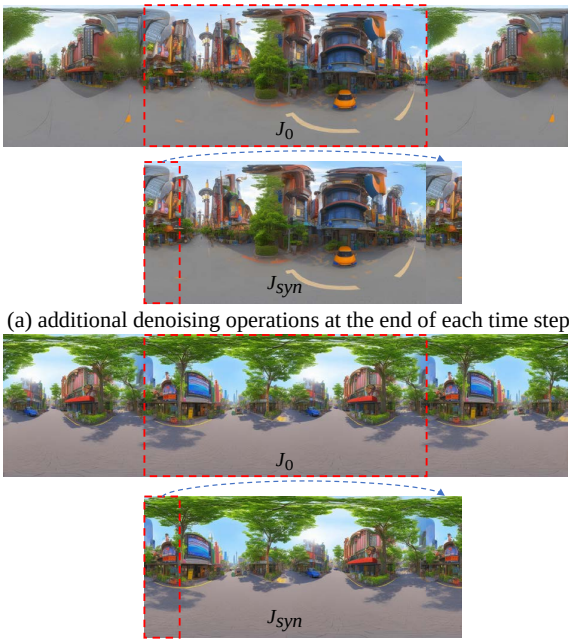
This supplementary material begins by presenting additional ablation studies. Next, we showcase sample images from various scenes in our collected *360PanoI* dataset and generation process of their text prompts. Finally, more images synthesized using different methods are shown.

A.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 10. 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 ω between adjacent cropped patches on the synthesized images, a comparison was carried out using various sliding distances. The visual results are presented in Figure 11. There is a noticeable seam in the middle of the region indicated by the red solid box when the sliding distance ω 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 ω set to 128, the content within the region now marked by the green solid box exhibits seamless and consistent integration.

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 [23] 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 12, 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.



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

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

Figure 10. Ablation study on the order of additional denoising operations. The corresponding text prompt is 'V^*, zootopia 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.

A.2. Sample Images and Their Text Prompts

The 360-degree panoramas in our collected *360PanoI* dataset have been sourced from Poly Haven [1]. To display the different scenes within our dataset, we randomly select one image from each scene. The corresponding sample images, with a resolution of 512×1024, are illustrated in Figure 13. 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 14 to demonstrate the generation process of the corresponding text prompts for these 360-degree panoramas within the *360PanoI* dataset.

A.3. More Visual Results

To highlight the distinctions between *MultiDiffusion* [5] and our proposed *StitchDiffusion*, we present a visual comparison between various schemes of *MultiDiffusion* and our *StitchDiffusion* in Figure 15. We can see that the combi-

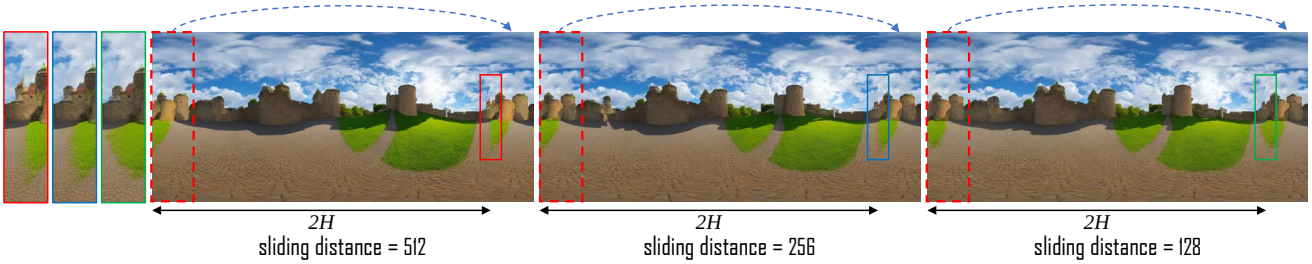


Figure 11. Ablation study on the horizontal sliding distance. The corresponding text prompt is 'V^*, castle, a beautiful artwork illustration'. When the horizontal sliding distance ω in the *StitchDiffusion* is 128, the content in the green solid box is more consistent and seamless than those in the blue and red solid boxes for the sliding distances of 256 and 512.

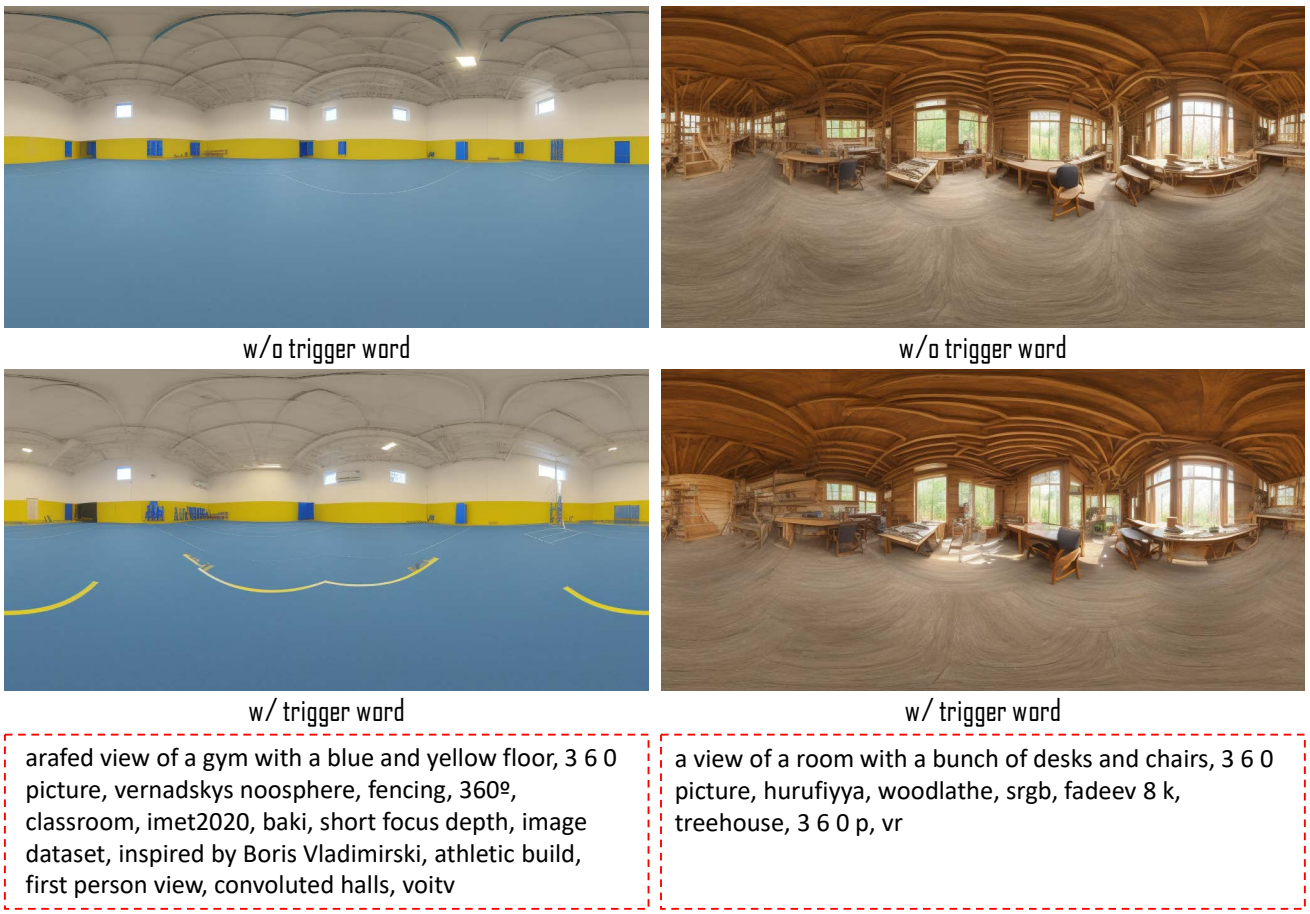


Figure 12. Ablation study on the effectiveness of the trigger word V^* if poor tags in text prompts are not filtered out. We collected two real 360-degree panoramas from Poly Haven [1], which are independent of the *360PanoI* dataset, and utilized BLIP [23] to create corresponding text prompts denoted 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.

nation of *MultiDiffusion* and our customized model in (a) fails 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

indoor

nature

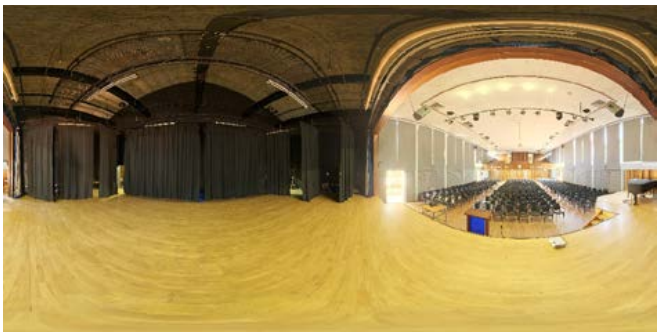
night

outdoor

skies

studio

sunrise-sunset



urban

Figure 13. 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 [23].

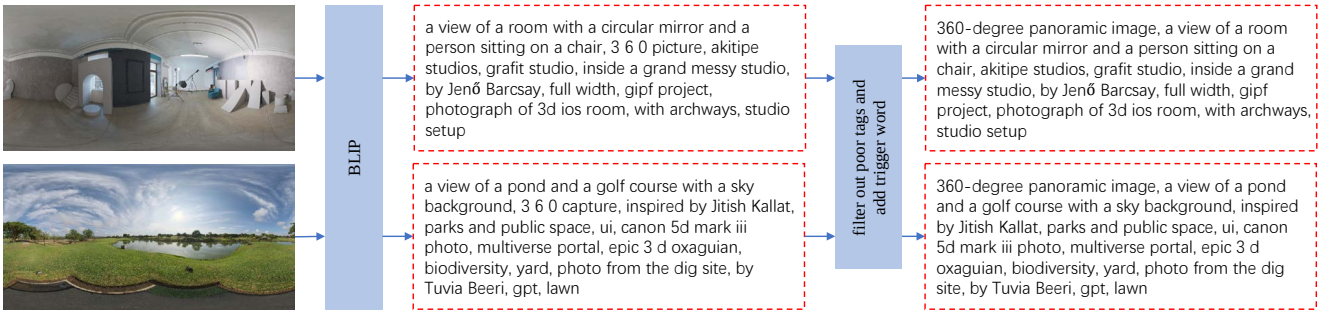


Figure 14. Diagram to show the generation process of text prompts in our *360PanoI* dataset using BLIP [23]. For the collected 120 360-degree panoramas, we employ BLIP to produce their text prompts. Then, we filter out poor tags such as '3 6 0 picture' and introduce a trigger word '360-degree panoramic 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 15. 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 by makoto shinkai'. Since the size of the cropped patch in *MultiDiffusion* [5] is 512×512, the stitch block here consists of the leftmost (512×256) and rightmost (512×256) 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.

comparison results involving our method, Stable Diffusion [37] combined with *MultiDiffusion* [5], and Text2Light [6], shown in Figure 16 and Figure 17. 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 18 and Figure 19.

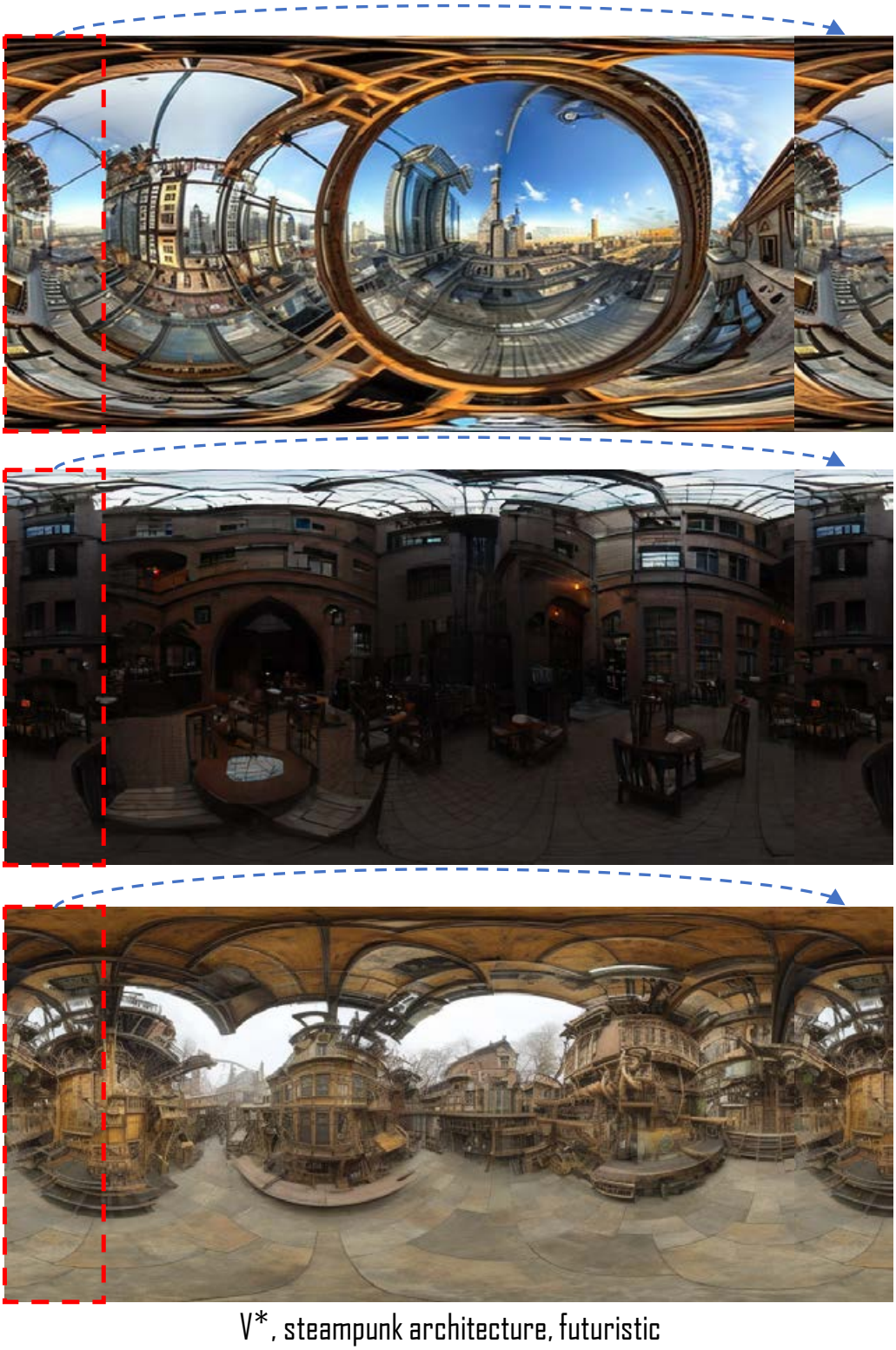


Figure 16. Visual comparison among *MultiDiffusion* [5] combined with Stable Diffusion [37] (the first row), Text2Light [6] (the second row) and our method (the third row). To demonstrate the discontinuity or continuity between the leftmost and rightmost sides of the generated image, we copy the leftmost area (512×128) 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.

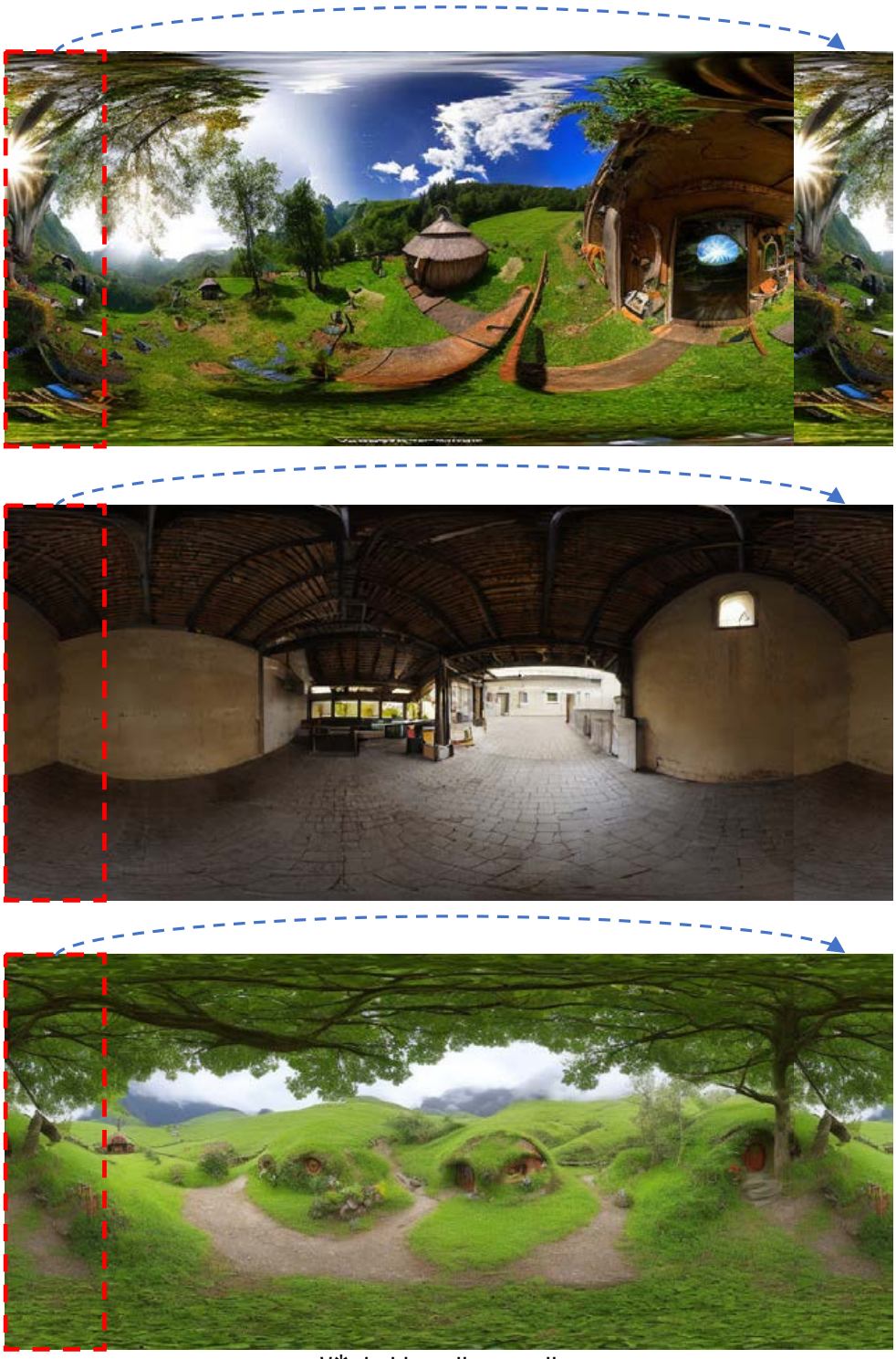
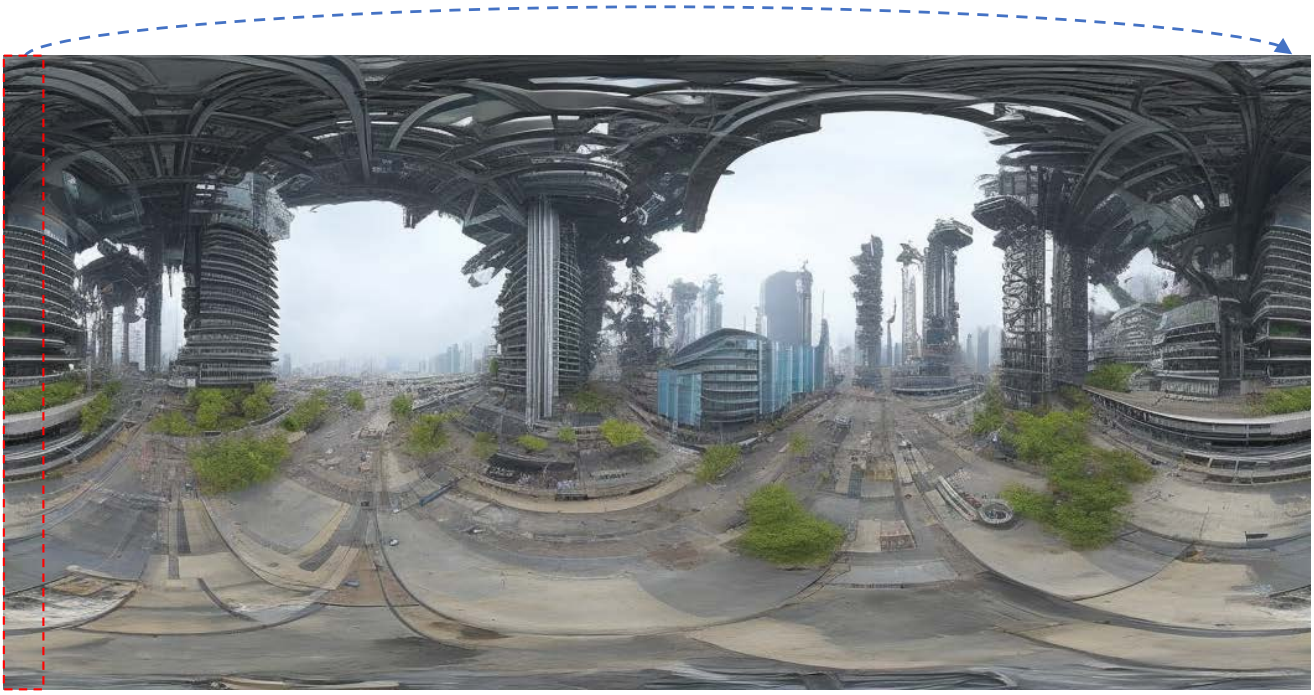


Figure 17. Visual comparison among *MultiDiffusion* [5] combined with Stable Diffusion [37] (the first row), Text2Light [6] (the second row) and our method (the third row). To display the discontinuity or continuity between the leftmost and rightmost sides of the generated image, we copy the leftmost area (512×128) 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 18. 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 (512×32) 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 19. 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 (512×32) 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.