

360PanT: Training-Free Text-Driven 360-Degree Panorama-to-Panorama Translation

Hai Wang* and Jing-Hao Xue
University College London

{hai.wang.22, jinghao.xue}@ucl.ac.uk

Abstract

Preserving boundary continuity in the translation of 360-degree panoramas remains a significant challenge for existing text-driven image-to-image translation methods. These methods often produce visually jarring discontinuities at the translated panorama’s boundaries, disrupting the immersive experience. To address this issue, we propose 360PanT, a training-free approach to text-based 360-degree panorama-to-panorama translation with boundary continuity. Our 360PanT achieves seamless translations through two key components: boundary continuity encoding and seamless tiling translation with spatial control. Firstly, the boundary continuity encoding embeds critical boundary continuity information of the input 360-degree panorama into the noisy latent representation by constructing an extended input image. Secondly, leveraging this embedded noisy latent representation and guided by a target prompt, the seamless tiling translation with spatial control enables the generation of a translated image with identical left and right halves while adhering to the extended input’s structure and semantic layout. This process ensures a final translated 360-degree panorama with seamless boundary continuity. Experimental results on both real-world and synthesized datasets demonstrate the effectiveness of our 360PanT in translating 360-degree panoramas. Code is available at <https://github.com/littlewhitesea/360PanT>.

1. Introduction

Text-driven image-to-image (I2I) translation seeks to generate a new image that reflects a given target prompt while following the structure and semantic layout of an input image. For text-driven I2I translation, recent training-free methods, such as Prompt-to-Prompt (P2P) [2], Plug-and-Play (PnP) [1] and FreeControl [3], are based on pre-trained latent diffusion models (LDMs) [7] and typically employ DDIM inversion [4] to obtain the corresponding

noisy latent representation of the input image. Subsequently, they leverage attention control [1,2] or spatial control [3] to guide the translation process during denoising. By harnessing the powerful generative capabilities of pre-trained LDMs [7], these methods demonstrate commendable performance in translating ordinary images.

However, directly applying these techniques to 360-degree panoramic images, which are commonly represented by using equirectangular projection [8], presents a unique and significant challenge. Unlike ordinary images, 360-degree panoramas possess inherent boundary continuity, where the leftmost and rightmost edges seamlessly connect. Existing I2I translation methods based on DDIM inversion fail to preserve this crucial characteristic, resulting in noticeable discontinuities at the boundaries of translated panoramas, as shown in Figure 1. To solve this problem, we propose 360PanT, a training-free method tailored for text-driven 360-degree panorama-to-panorama (Pan2Pan) translation. Our approach comprises two primary components: **boundary continuity encoding** and **seamless tiling translation with spatial control**.

Boundary continuity encoding aims to embed the boundary continuity information of the input 360-degree panorama into the noisy latent representation. This is achieved by first creating an extended input image obtained from splicing two copies of the original input panorama. This extended input is then processed by the encoder of a pre-trained LDM. Finally, DDIM inversion is applied to the resulting latent feature, yielding a noisy latent feature that intrinsically encodes the boundary continuity.

While one might consider directly applying existing state-of-the-art (SOTA) I2I translation techniques like PnP [1] or FreeControl [3] to this noisy latent feature, such an approach presents two significant drawbacks. Firstly, processing the entire noisy latent feature on a single high-end GPU (e.g., 24GB) leads to out-of-memory errors. Secondly, these SOTA methods cannot guarantee the preservation of identical left and right halves throughout the denoising process, potentially disrupting the 360-degree panoramic structure.

*Corresponding author.

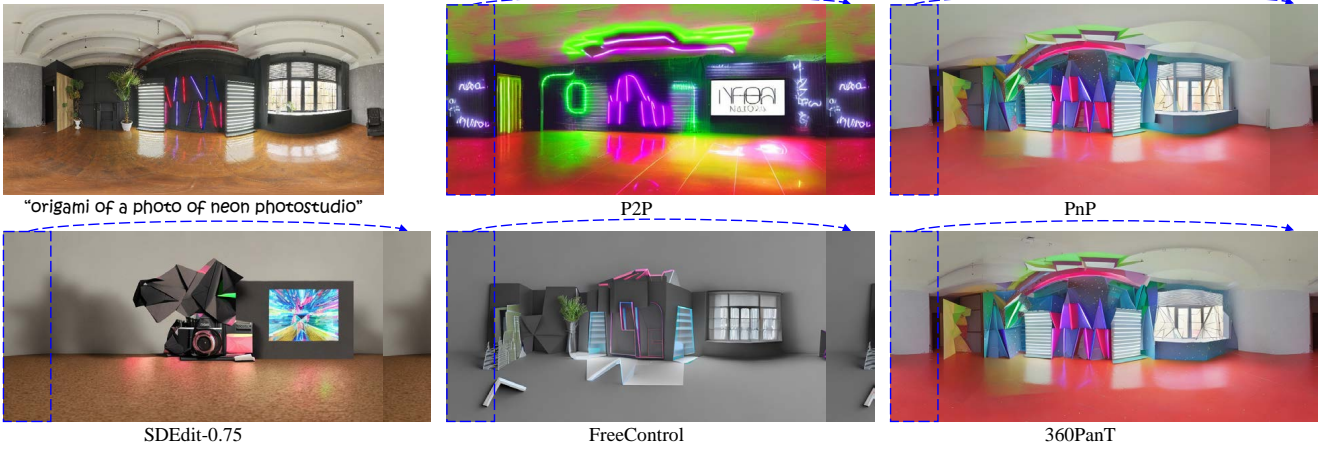


Figure 1. **Example of text-driven 360-degree panorama-to-panorama translation.** To easily identify visual continuity or discontinuity at the boundaries of the translated panoramic image, we copy the left area indicated by the blue dashed box and paste it onto the rightmost side of the image. Compared with other methods, our 360PanT performs best in maintaining boundary continuity and preserving the structure and semantic layout of the input 360-degree panorama in the translated result.

To address these issues, we put forward **seamless tiling translation with spatial control**. Specifically, we leverage a key property of StitchDiffusion [6], a method designed for generating 360-degree panoramas from using a customized latent diffusion model [7]. StitchDiffusion inherently produces images with identical left and right halves, ensuring panoramic continuity. Moreover, cropped patches of the noisy latent feature, instead of the entire noisy latent feature, are independently processed within StitchDiffusion during denoising. This **seamless tiling translation** strategy effectively addresses the memory constraints and guarantees the preservation of the 360-degree panoramic structure.

However, relying solely on the noisy latent feature and the target prompt leads to a translated image that deviates from the structure and semantic layout of the extended input. To solve this problem, we integrate **spatial control** into the seamless tiling translation process. Inspired by PnP [1], we inject spatial features and self-attention maps from the extended input image into the seamless tiling translation process. The spatial control mechanism enables the translated image to maintain the structure and semantic layout of the extended input, resulting in a finely translated 360-degree panorama.

Furthermore, an alternative to spatial feature and self-attention map injection is explored. Drawing inspiration from FreeControl [3], we introduce structure guidance and appearance guidance into the seamless tiling translation process. This approach allows our 360PanT to support a variety of 360-degree panoramic maps (e.g., segmentation masks and edge maps) as input conditions instead of a standard 360-degree panoramic image.

Novelties and Contributions. (1) We propose 360PanT, the first training-free method for text-driven 360-degree panorama-to-panorama translation, which consists of two

key components: boundary continuity encoding and seamless tiling translation with spatial control. (2) Beyond standard 360-degree panoramic images, 360PanT can expand its capacity to support various types of 360-degree panoramic maps (e.g., segmentation masks and edge maps) as input conditions. This flexibility extends its applications to various scenarios requiring diverse input formats. (3) Extensive experiments on both real-world and synthesized datasets demonstrate the effectiveness of our proposed method in translating 360-degree panoramas through text prompts.

2. Related Work

Text-Driven 360-Degree Panorama Generation. The objective of text-driven panorama generation [30, 34–36] is to produce panoramic images aligned with given textual descriptions. Unlike ordinary panoramic images, 360-degree panoramic images offer immersive experiences and find broad applications in virtual reality [38], autonomous driving [37], and indoor design [40].

For synthesizing 360-degree panoramas from text prompts, Text2Light [33] introduces a hierarchical framework comprising a dual-codebook discrete representation, a text-conditioned global sampler, and a structure-aware local sampler. In contrast, recent approaches [6, 31, 32, 39, 41] explore text-to-image latent diffusion models [7] for text-driven 360-degree panorama generation. Among these methods, StitchDiffusion [6] proposes additional denoising twice on the stitch patch based on MultiDiffusion [30], ensuring the generated image to have identical left and right halves. We leverage this crucial attribute of StitchDiffusion to achieve seamless tiling translation in our 360PanT.

Text-Guided Image-to-Image Translation. Image-to-image (I2I) translation aims to learn a mapping that transforms images between domains while maintaining the se-

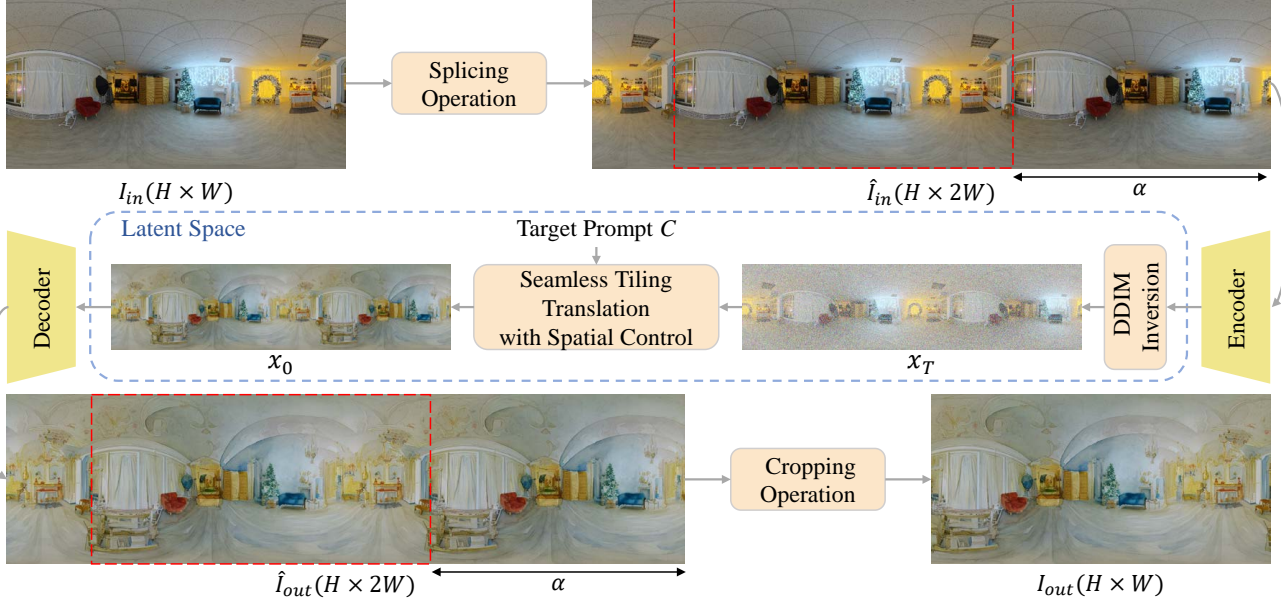


Figure 2. **Method overview.** Our 360PanT comprises two primary components: boundary continuity encoding and seamless tiling translation with spatial control. The boundary continuity encoding component embeds the boundary continuity information of I_{in} into the noisy latent feature x_T . Subsequently, guided by the target prompt C , x_T undergoes seamless tiling translation with spatial control to produce the denoised translated latent feature x_0 . Finally, the translated 360-degree panorama I_{out} , aligned with the target prompt C , is achieved by cropping from the translated image \hat{I}_{out} .

mantic layout and structure of the input image. Over the past few years, GAN-based I2I translation methods have been extensively investigated [10–19]. Recently, diffusion models [4, 20–23] have emerged as a powerful alternative to GANs [9], exhibiting superior performance in image synthesis. This shift has motivated research into exploring diffusion models for I2I translation [1–3, 5, 24–29].

Notably, training-free text-driven I2I translation methods [1–3, 5, 29], building upon pre-trained latent diffusion models (LDMs) [7], have gained significant attention. For example, Plug-and-Play (PnP) [1] proposes to inject spatial features and self-attention maps into the denoising process of the translated image for enhancing structure preservation. Different from PnP, FreeControl [3] introduces appearance guidance and structure guidance to achieve spatial control of the translated image. Leveraging the powerful generative capabilities of pre-trained LDMs, these text-driven I2I methods achieve impressive results on ordinary images. However, when applied to 360-degree panoramic images, they fail to maintain visual continuity at the boundaries of the translated images. To address this problem, we propose a training-free method called 360PanT. By using our designed boundary continuity encoding and seamless tiling translation with spatial control, 360PanT successfully achieves the translation of 360-degree panoramas.

3. Methodology

The framework of our 360PanT is illustrated in Figure 2, consisting of two key components: boundary continuity encoding and seamless tiling translation with spatial control. Details of each component are elaborated in the following.

3.1. Boundary Continuity Encoding

Recent training-free text-driven image-to-image (I2I) translation methods, such as Prompt-to-Prompt (P2P) [2], Plug-and-Play (PnP) [1] and FreeControl [3], face inherent limitations when applied to 360-degree panoramas. This limitation stems from the inability of the DDIM inversion process [4], a core component of these methods, to encode the continuous information between the leftmost and rightmost sides of a 360-degree panorama. DDIM inversion, primarily designed for ordinary images, converts a clear image into a noisy latent representation without accounting for the cyclical nature of 360-degree panoramas. Consequently, these training-free I2I translation methods [1–3, 5] relying on DDIM inversion fail to maintain visual continuity between the edges of the final translated panorama.

To address this challenge, we propose a straightforward yet effective method to encode this crucial continuous information. Our approach involves firstly splicing two identical copies of the input panorama, to create an extended image that serves as input for the DDIM inversion process [4]. Formally, given an input 360-degree panorama I_{in} with dimen-

sions $3 \times H \times W$, the extended input \hat{I}_{in} with dimensions $3 \times H \times 2W$ is constructed as follows:

$$\hat{I}_{in} = \text{Splice}(I_{in}[:, :, \alpha : W], I_{in}, I_{in}[:, :, 0 : \alpha]), \quad (1)$$

where α is a split constant controlling the splicing point, and *Splice* denotes the image splicing operation. Note that (1) setting α equal to W results in \hat{I}_{in} being a direct concatenation of two copies of I_{in} ; and (2) the extended input \hat{I}_{in} consistently maintains identical left and right halves regardless of the value of α . Subsequently, this extended image \hat{I}_{in} is encoded into the latent space, and DDIM inversion is applied to its corresponding latent feature, yielding a noisy latent feature x_T with dimensions $4 \times \frac{H}{8} \times \frac{2W}{8}$, which naturally embeds the boundary continuous information of the original 360-degree panorama.

3.2. Seamless Tiling Translation

At this stage, we have a noisy latent feature x_T including the continuous information of the original 360-degree panorama I_{in} . A direct approach to performing training-free text-driven panorama-to-panorama translation would be to apply existing I2I translation methods, such as PnP [1] or FreeControl [3], to x_T and then crop the translated image (with dimensions $3 \times H \times 2W$) to obtain the final 360-degree output. However, this approach has two significant drawbacks: (1) directly processing the entire x_T on a single high-end GPU (e.g., 24GB) results in out-of-memory errors; and (2) these methods cannot ensure that the translated image will still maintain identical left and right halves during the denoising process, potentially disrupting the panoramic structure.

To overcome these issues, we leverage a key property of StitchDiffusion [6], a method designed for generating 360-degree panoramas using a customized latent diffusion model [7]. StitchDiffusion inherently produces images with identical left and right halves by design, ensuring the preservation of the 360-degree panoramic structure. Furthermore, at denoising step t , where $t \in \{T, T-1, \dots, 1\}$, cropped patches of x_t , rather than the entire x_t , are independently processed within StitchDiffusion. Therefore, instead of directly applying existing I2I translation methods, we employ StitchDiffusion to translate the noisy latent feature x_T . This approach effectively addresses the aforementioned memory constraints and ensures the translated image maintaining identical left and right halves.

Specifically, at denoising step t , the noisy latent feature x_t is divided into n overlapping patches. Let $F_i(x_t)$ represent the i -th cropped patch of size $\frac{H}{8} \times \frac{W}{8}$, where $i \in \{1, 2, \dots, n\}$. Here, the mapping F_i denotes the cropping operation for the i -th patch, and its inverse mapping, F_i^{-1} , places the patch back into its original position. The number of patches, n , is determined by $\frac{W}{8\omega} + 1$, where ω indicates the sliding distance between adjacent patches

$F_i(x_t)$ and $F_{i+1}(x_t)$. In addition, let Φ and C denote a pre-trained latent diffusion model [7] and a target prompt, respectively. In this situation, the sequential denoising process of a training-free I2I translation using StitchDiffusion, termed seamless tiling translation process, can be represented as

$$x_{t-1} = \sum_{j=1}^2 \frac{{}^j F_{n+1}^{-1}(\mathbf{1})}{\Pi} \otimes {}^j F_{n+1}^{-1}(\Phi({}^j F_{n+1}(x_t), C)) \\ + \sum_{i=1}^n \frac{F_i^{-1}(\mathbf{1})}{\Pi} \otimes F_i^{-1}(\Phi(F_i(x_t), C)), \quad (2)$$

where ${}^j F_{n+1}(\cdot)$ and ${}^j F_{n+1}^{-1}(\cdot)$ are the j -th additional mapping and inverse mapping of the stitch patch, respectively; and Π denotes ${}^1 F_{n+1}^{-1}(\mathbf{1}) + {}^2 F_{n+1}^{-1}(\mathbf{1}) + \sum_{i=1}^n F_i^{-1}(\mathbf{1})$, where $\mathbf{1}$ refers to a latent feature with dimensions $4 \times \frac{H}{8} \times \frac{W}{8}$ with all values equal to 1. The stitch patch, a special cropped patch, is defined as $\text{Splice}(x_t[:, :, \frac{3W}{16} : \frac{2W}{8}], x_t[:, :, 0 : \frac{W}{16}])$, where, as in Eq. 1, *Splice* is the splicing operation.

Through the seamless tiling translation process (Eq. 2), we obtain the final denoised latent feature x_0 with dimensions $4 \times \frac{H}{8} \times \frac{2W}{8}$. Consequently, the corresponding translated image \hat{I}_{out} with dimensions $3 \times H \times 2W$ decoded from x_0 maintains identical left and right halves while corresponding to the target prompt C .

3.3. Seamless Tiling Translation with Spatial Control

Capitalizing on both boundary continuity encoding and seamless tiling translation, the diffusion model Φ can produce a translated image \hat{I}_{out} with identical left and right halves, aligned with the target prompt C . However, the seamless tiling translation relies solely on C and the initial noisy latent feature x_T . Consequently, the translated image \hat{I}_{out} may not fully adhere to the structure and semantic layout of the extended input \hat{I}_{in} . To address this issue, we propose to incorporate spatial control into the seamless tiling translation, enabling training-free text-based 360-degree panorama-to-panorama (Pan2Pan) translation.

Specifically, following the Plug-and-Play (PnP) method [1], we inject spatial features \mathbf{f}_t and self-attention maps \mathbf{A}_t from $x_{t-1}^o = \Phi(x_t^o, \emptyset)$ into the seamless tiling translation process, where $t \in \{T, T-1, \dots, 1\}$. Here, x_T^o is identical to x_T , and \emptyset represents a null text prompt. In this context, the seamless tiling translation process with spatial control is given by

$$x_{t-1} = \sum_{j=1}^2 \frac{{}^j F_{n+1}^{-1}(\mathbf{1})}{\Pi} \otimes {}^j F_{n+1}^{-1}(\Phi({}^j F_{n+1}(x_t), C; \mathbf{f}_t, \mathbf{A}_t)) \\ + \sum_{i=1}^n \frac{F_i^{-1}(\mathbf{1})}{\Pi} \otimes F_i^{-1}(\Phi(F_i(x_t), C; \mathbf{f}_t, \mathbf{A}_t)). \quad (3)$$

Utilizing this spatially controlled translation process, we decode the final denoised latent feature x_0 to get the translated image \hat{I}_{out} with dimensions $3 \times H \times 2W$. Subsequently, we extract the final translated 360-degree panorama I_{out} with dimensions $3 \times H \times W$ by cropping \hat{I}_{out} :

$$I_{out} = \hat{I}_{out}[:, :, W - \alpha : 2W - \alpha], \quad (4)$$

where, as in Eq. 1, α is the split constant.

To further enhance 360PanT’s versatility and enable support for diverse input conditions (e.g., segmentation masks and edge maps) beyond using standard 360-degree panoramic images, we explore an alternative to spatial feature and self-attention map injection. Inspired by FreeControl [3], we introduce structure guidance $g_s(t)$ and appearance guidance $g_a(t)$ into the seamless tiling translation process, where $t \in \{T, T-1, \dots, 1\}$. These guidance terms, $g_s(t)$ and $g_a(t)$, are derived from the denoising process of x_t and x_t^r , respectively. Here, x_T^r is a randomly initialized latent feature following a normal distribution, which is not equal to x_T . In this context, the seamless tiling translation process incorporating FreeControl’s spatial control is updated as

$$\begin{aligned} x_{t-1}^r = & \sum_{j=1}^2 \frac{j F_{n+1}^{-1}(\mathbf{1})}{\Pi} \otimes j F_{n+1}^{-1}(\Phi(j F_{n+1}(x_t^r), C; g_a(t), g_s(t))) \\ & + \sum_{i=1}^n \frac{F_i^{-1}(\mathbf{1})}{\Pi} \otimes F_i^{-1}(\Phi(F_i(x_t^r), C; g_a(t), g_s(t))). \end{aligned} \quad (5)$$

Note that this seamless tiling translation process is performed on the latent feature x_T^r instead of x_T to support diverse input conditions. Similarly, we obtain the final translated image \hat{I}_{out} with dimensions $3 \times H \times 2W$ decoded from x_0^r . A cropping operation is then carried out to achieve the corresponding translated 360-degree panorama I_{out} , as described in Eq. 4.

4. Experiments and Results

Implementation details. The values of H and W in this paper are 512 and 1024. We set the values of split constant α and sliding distance ω to 768 and 16, respectively. The version of the pre-trained latent diffusion model [7] is Stable Diffusion 2-1-base. For seamless tiling translation with spatial control, our 360PanT method primarily employs PnP’s spatial control mechanism [1]. To enable support for diverse input conditions, we introduce a variant denoted as 360PanT (F), which utilizes FreeControl’s spatial control [3] instead of PnP. The settings for the spatial control components and denoising steps T within 360PanT and 360PanT (F) are consistent with the default settings of PnP and FreeControl, respectively. All experiments were carried out using a single NVIDIA L4 GPU.

Datasets. Our 360PanT is capable of translating both real-world and synthesized 360-degree panoramas guided by text prompts. Due to the absence of a benchmark dataset for text-driven 360-degree panorama-to-panorama (Pan2Pan) translation, we established two datasets for this purpose. The first dataset, termed *360PanoI-Pan2Pan*, is derived from the *360PanoI* dataset [6], which contains 120 real-world 360-degree panoramas across eight scenes. Complementing this, we created *360syn-Pan2Pan*, a synthesized dataset comprising 120 360-degree panoramic images generated using the method outlined in [6]. To construct text-image pairs for 360-degree Pan2Pan translation, we defined 10 translation types (e.g., watercolor painting, anime artwork, and cartoon). The target prompt for each input 360-degree panorama was formed by randomly selecting a translation type and combining it with the original text prompt. For further details on the target prompts for the two datasets, please refer to the supplementary material.

Evaluation metrics. To quantitatively evaluate the effectiveness of various methods for text-driven 360-degree Pan2Pan translation, we employ metrics used in PnP [1]. Specifically, we utilize text and image encoders from CLIP [42] to extract textual embeddings of target prompts and image embeddings of corresponding translated panoramic images. We then calculate the average cosine similarity, which is referred to as *CLIP-score*, between these textual and image embeddings. In addition, we use the DINO-ViT self-similarity distance [43], denoted as *DINO-score*, to assess the preservation of structural integrity in the translated 360-degree panoramas compared to the input 360-degree panoramas. These two metrics are reported for the *360PanoI-Pan2Pan* and *360syn-Pan2Pan* datasets.

4.1. Comparisons with Other Methods

We compare our 360PanT with state-of-the-art (SOTA) text-driven image-to-image (I2I) translation approaches: SDEdit [5], Pix2Pix-zero [29], Prompt-to-Prompt (P2P) [2], Plug-and-Play (PnP) [1], FreeControl [3]. Visual results from the different methods on the translation of real-world and synthesized 360-degree panoramas are illustrated in Figure 3 and Figure 4, respectively. These figures demonstrate that these SOTA text-driven I2I translation methods fail to preserve the boundary continuity in the translated panoramas. In contrast, our 360PanT not only successfully maintains the visual continuity at the boundaries of the translated panoramas, but also ensures the translated results adhere to the structure and semantic layout of the input 360-degree panoramas. Note that due to space limitations, we only present part visual results here; additional visual results are in the supplementary material.

To further evaluate the performance, we analyze the *CLIP-score* and *DINO-score* metrics across the *360PanoI-Pan2Pan* and *360syn-Pan2Pan* datasets. The results, de-

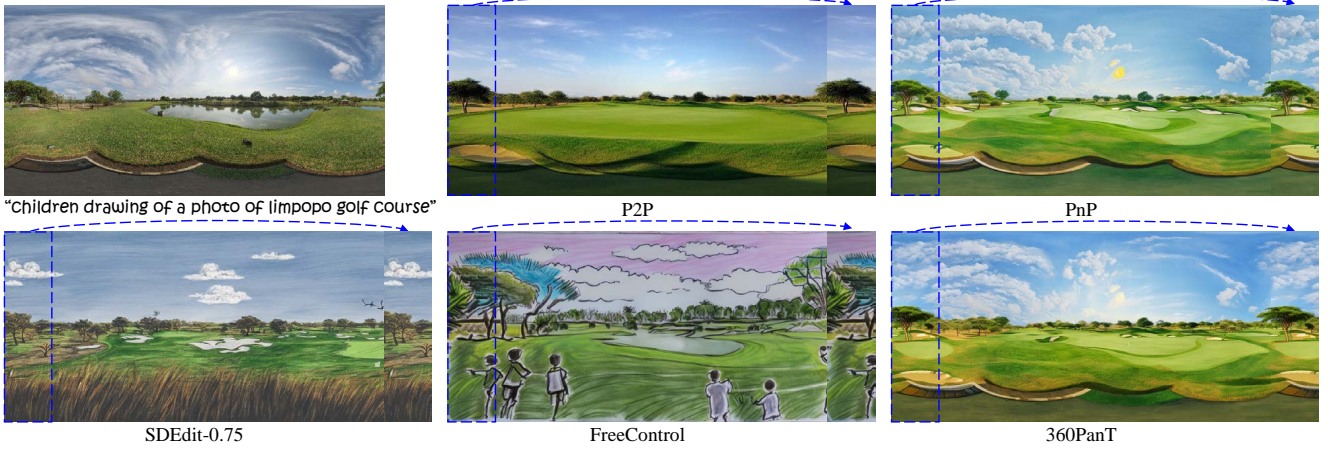


Figure 3. **Visual results on real-world 360-degree panorama.** To easily identify visual continuity or discontinuity at the boundaries, we copy the left area of the panorama indicated by the blue dashed box and paste it onto the rightmost side of the image. Current I2I translation methods fail to maintain visual continuity at the boundaries of the translated panoramas. In contrast, our 360PanT not only ensures boundary continuity but also preserves the guidance structure in the translated 360-degree output.



Figure 4. **Visual results on synthesized 360-degree panorama.** Compared to other text-driven I2I methods, our 360PanT performs better in maintaining the visual continuity at the boundaries while also adhering to the structure and semantic layout of the input 360-degree panoramic image. For more visual results, please refer to the supplementary material.

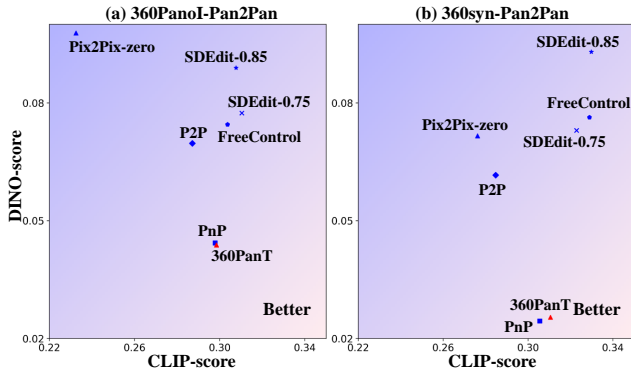


Figure 5. **Quantitative comparison.** DINO-score (lower is better) is to evaluate the structure preservation, while CLIP-score (higher is better) is to assess the prompt fidelity. Bottom-right is the best.

picted in Figure 5, reveal a close alignment between PnP and 360PanT in both metrics. This similarity is expected, given that 360PanT adopts the same spatial control as PnP. However, a key limitation of PnP is its inability to maintain visual continuity at panorama boundaries. Conversely, our 360PanT can produce translated panoramas with continuous boundaries.

4.2. Ablation Studies

Effect of seamless tiling translation. To demonstrate the effectiveness of seamless tiling translation, we conducted some simple I2I translation experiments. Specifically, with a 360-degree panorama denoted as I_{in} with dimensions $3 \times 512 \times 1024$ (indicated by the red dashed box in Figure 6), two identical copies were directly spliced to generate an extended image \hat{I}_{in} . Subsequently, DDIM inversion [4] was applied to the latent feature representation of \hat{I}_{in} . The re-

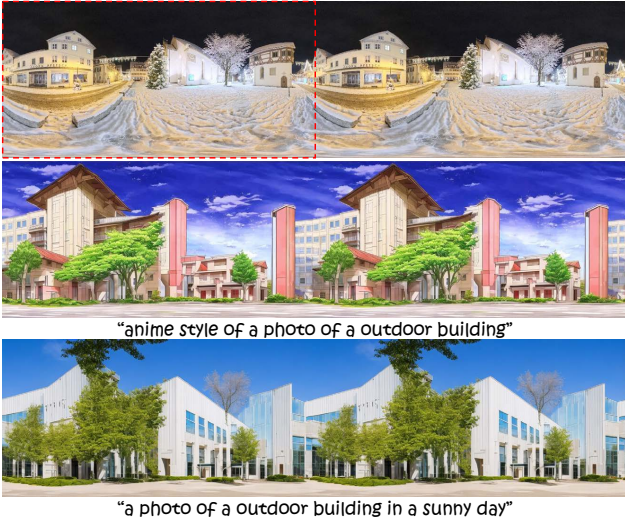


Figure 6. **Ablation on seamless tiling translation effect.** The image in the first row is the extended input, with the input 360-degree panorama highlighted within the **red dashed box**. The second and third rows display the translated images using seamless tiling translation with distinct target prompts. Notably, both translated images exhibit identical left and right halves, effectively demonstrating the seamless tiling effect, while simultaneously corresponding to their respective target prompts.

sulting noisy latent feature, x_T , underwent seamless tiling translation (Eq. 2) guided by two distinct text prompts, respectively. This process yielded two corresponding translated images. Figure 6 illustrates the successful generation of two translated images with dimensions $3 \times 512 \times 2048$. These images exhibit identical left and right halves while corresponding to their respective target prompts, highlighting the efficacy of the seamless tiling translation.

Seamless tiling translation with spatial control. To investigate the impact of spatial control mechanisms on seamless tiling translation, we carried out a comparative experimental study. In this study, we utilized an extended input image \hat{I}_{in} for translation guided by a target prompt. This image underwent three distinct translation processes: (1) seamless tiling translation alone, (2) seamless tiling translation with PnP’s spatial control, and (3) seamless tiling translation with FreeControl’s spatial control. The resulting translated images are displayed in Figure 7.

We observe that, firstly, incorporating FreeControl’s spatial control into seamless tiling translation improves the translated image’s adherence to the structure and semantic layout of the extended input image \hat{I}_{in} , compared with seamless tiling translation alone. Secondly, integrating PnP’s spatial control into seamless tiling translation preserves the structure and semantic layout of \hat{I}_{in} even more effectively than using FreeControl’s spatial control. Based on these findings, we adopt PnP’s spatial control in our 360PanT method. To distinguish between these variations, we refer to 360PanT incorporating FreeControl’s spatial control as 360PanT (F) throughout this paper. While

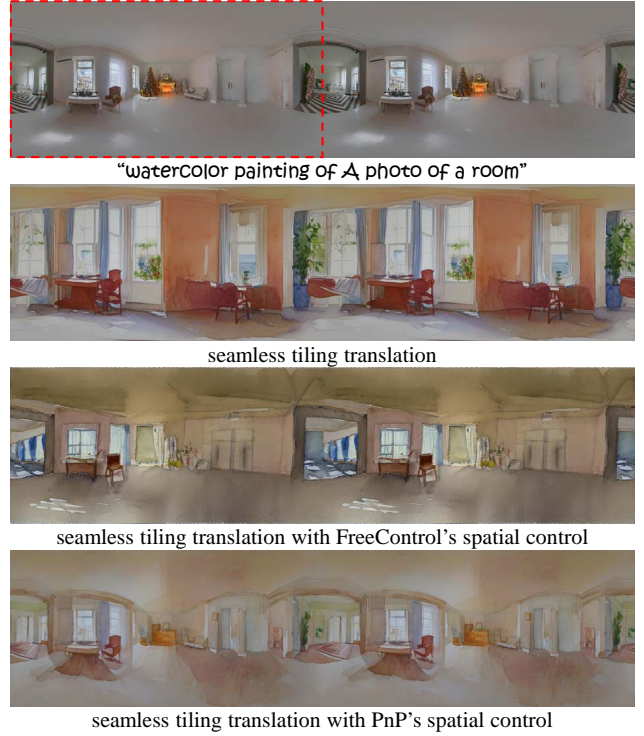


Figure 7. **Ablation on spatial control for seamless tiling translation.** The first row displays the extended input \hat{I}_{in} , with the original input 360-degree panorama highlighted within the **red dashed box**. Subsequent rows present the translated images generated by using the same target prompt but employing the following different methods: (2nd row) seamless tiling translation alone; (3rd row) seamless tiling translation with FreeControl’s spatial control; and (4th row) seamless tiling translation with PnP’s spatial control. Visual comparison discloses that integrating FreeControl’s spatial control enhances the preservation of structure and semantic layout from \hat{I}_{in} ; and incorporating PnP’s spatial control improves preservation even more than FreeControl’s approach.

360PanT (F) is not so effective as 360PanT in structure preservation, it enables support for various input conditions beyond using standard 360-degree panoramic images, as described in Section 4.3.

Choice of split constant α . To study the influence of parameter α on the quality of the final translated 360-degree panorama, we carried out experiments on our 360PanT with varying α values: W (1024) and $\frac{3W}{4}$ (768), where W denotes the width of the input 360-degree panorama. As shown in Figure 8, our 360PanT with $\alpha = W$ demonstrates significantly better boundary continuity than the PnP baseline. However, upon closer inspection of the zoomed-in region indicated by the **red solid box**, a minor crack artifact is noticeable in the stitched area. Conversely, employing 360PanT with $\alpha = \frac{3W}{4}$ yields a 360-degree panorama without visible cracks in the stitched region. We set α to $\frac{3W}{4}$ in this paper. An intuitive explanation of this parameter choice, supported by further analysis, is provided in the supplementary material.

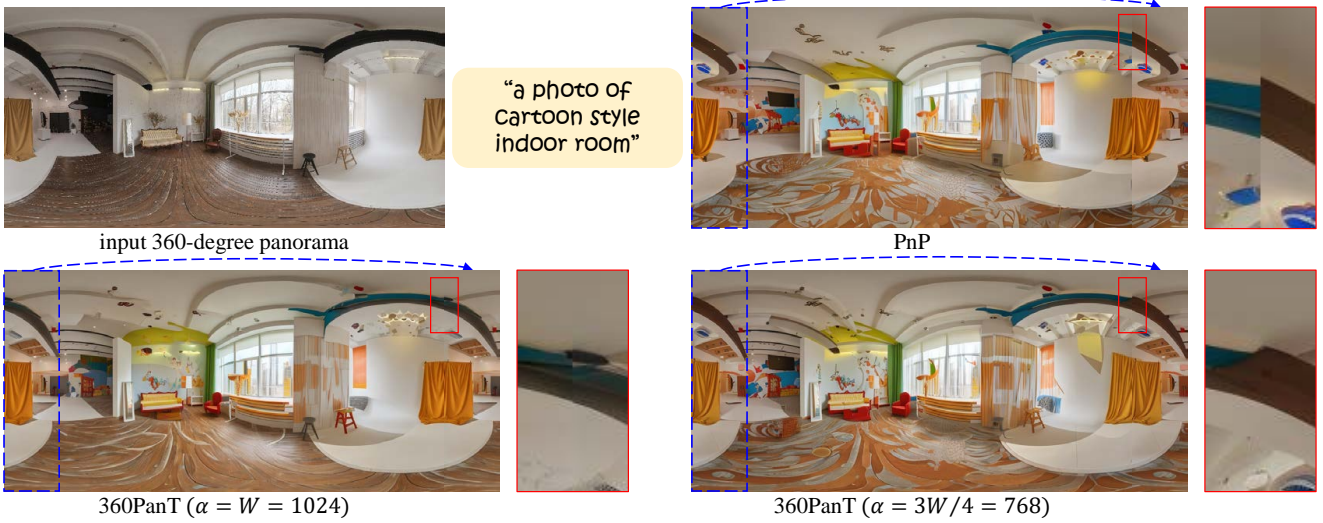


Figure 8. **Ablation on choice of split constant α .** While 360PanT with $\alpha = W$ significantly improves the boundary continuity of the translated panorama compared with PnP, a minor crack artifact is still noticeable in the stitched area upon closer inspection (see zoomed-in region highlighted by the red solid box). In contrast, setting α to $\frac{3W}{4}$ in 360PanT yields a panorama without visible crack artifacts in the stitched region. A further explanation of this parameter choice is available in the supplementary material.

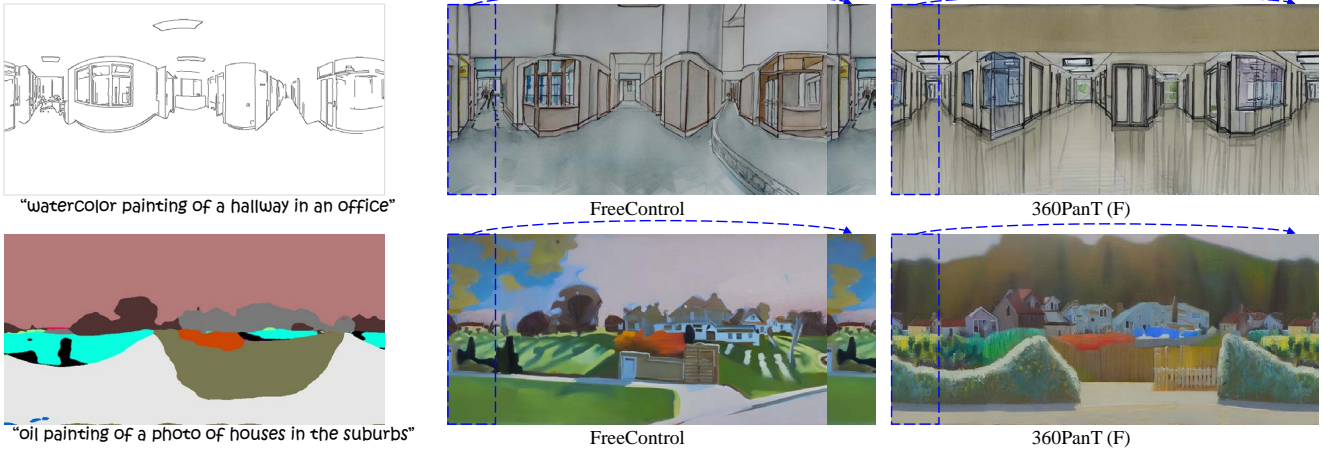


Figure 9. **Visual results using other control conditions.** FreeControl is unable to guarantee the boundary continuity of the translated panoramas. In contrast, our 360PanT (F) enables the translated 360-degree panoramas with continuous boundaries regardless of the input conditions. For more visual results, please refer to the supplementary material.

4.3. Translation using Other Control Conditions

To showcase the efficacy of our 360PanT (F) in handling diverse input conditions beyond 360-degree panoramic images, we present translated 360-degree panoramas generated from other control signals. Specifically, we consider a Canny edge map and a segmentation mask as the input control conditions, respectively, extracted from corresponding 360-degree panoramic images by using the same methods described in FreeControl [3]. Figure 9 demonstrates a comparative study, highlighting the limitations of FreeControl in preserving visual continuity under these conditions. In contrast, our 360PanT (F) effectively maintains boundary continuity in the translated 360-degree panoramas.

5. Conclusion

We propose 360PanT, a training-free method for text-driven 360-degree panorama-to-panorama translation. This method integrates boundary continuity encoding and seamless tiling translation with spatial control. By constructing an extended input image, the boundary continuity encoding embeds continuity information from the original 360-degree panorama into a noisy latent representation. Guided by a target prompt, the seamless tiling translation with spatial control leverages this latent representation to generate a translated image with identical left and right halves while following the structure and semantic layout of the extended input image. This process successfully results in a

final translated 360-degree panorama aligned with the target prompt. Extensive experiments on both real-world and synthesized 360-degree panoramas prove the effectiveness of our method in translating 360-degree panoramic images.

References

- [1] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, “Plug-and-play diffusion features for text-driven image-to-image translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1921–1930. 1, 2, 3, 4, 5, 12
- [2] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, “Prompt-to-prompt image editing with cross attention control,” *arXiv preprint arXiv:2208.01626*, 2022. 1, 3, 5, 12
- [3] S. Mo, F. Mu, K. H. Lin, Y. Liu, B. Guan, Y. Li, and B. Zhou, “Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition,” *arXiv preprint arXiv:2312.07536*, 2023. 1, 2, 3, 4, 5, 8, 11, 12
- [4] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020. 1, 3, 6
- [5] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, “SDEdit: Guided image synthesis and editing with stochastic differential equations,” in *International Conference on Learning Representations*, 2022. 3, 5, 12
- [6] H. Wang, X. Xiang, Y. Fan, and J.-H. Xue, “Customizing 360-degree panoramas through text-to-image diffusion models,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4933–4943. 2, 4, 5, 11
- [7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695. 1, 2, 3, 4, 5
- [8] M. Xu, C. Li, S. Zhang, and P. Le Callet, “State-of-the-art in 360 video/image processing: Perception, assessment and compression,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 5–26, 2020. 1
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020. 3
- [10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134. 3
- [11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 3
- [12] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” in *International conference on machine learning*. PMLR, 2017, pp. 1857–1865. 3
- [13] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” *Advances in neural information processing systems*, vol. 30, 2017. 3
- [14] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189. 3
- [15] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, “Tsit: A simple and versatile framework for image-to-image translation,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 206–222. 3
- [16] Y. Liu, M. De Nadai, D. Cai, H. Li, X. Alameda-Pineda, N. Sebe, and B. Lepri, “Describe what to change: A text-guided unsupervised image-to-image translation approach,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1357–1365. 3
- [17] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, “Vqgan-clip: Open domain image generation and editing with natural language guidance,” in *European Conference on Computer Vision*. Springer, 2022, pp. 88–105. 3
- [18] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, “Encoding in style: a stylegan encoder for image-to-image translation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2287–2296. 3
- [19] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797. 3
- [20] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265. 3
- [21] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. 3
- [22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *International Conference on Learning Representations*, 2020. 3
- [23] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021. 3
- [24] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, “Palette: Image-to-image diffusion models,” in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10. 3

- [25] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen, “Pretraining is all you need for image-to-image translation,” *arXiv preprint arXiv:2205.12952*, 2022. **3**
- [26] G. Kwon and J. C. Ye, “Diffusion-based image translation using disentangled style and content representation,” in *The Eleventh International Conference on Learning Representations*, 2022. **3**
- [27] G. Kim, T. Kwon, and J. C. Ye, “Diffusionclip: Text-guided diffusion models for robust image manipulation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435. **3**
- [28] B. Li, K. Xue, B. Liu, and Y.-K. Lai, “Bbmd: Image-to-image translation with brownian bridge diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1952–1961. **3**
- [29] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, “Zero-shot image-to-image translation,” in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11. **3, 5, 12**
- [30] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, “Multidiffusion: Fusing diffusion paths for controlled image generation,” *arXiv preprint arXiv:2302.08113*, 2023. **2**
- [31] M. Feng, J. Liu, M. Cui, and X. Xie, “Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models,” *arXiv preprint arXiv:2311.13141*, 2023. **2**
- [32] C. Zhang, Q. Wu, C. C. Gambardella, X. Huang, D. Phung, W. Ouyang, and J. Cai, “Taming stable diffusion for text to 360 panorama image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6347–6357. **2**
- [33] Z. Chen, G. Wang, and Z. Liu, “Text2light: Zero-shot text-driven hdr panorama generation,” *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–16, 2022. **2**
- [34] Q. Zhang, J. Song, X. Huang, Y. Chen, and M.-Y. Liu, “Diffcollage: Parallel generation of large content with diffusion models,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 10 188–10 198. **2**
- [35] Y. Lee, K. Kim, H. Kim, and M. Sung, “Syncdiffusion: Coherent montage via synchronized joint diffusions,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 50 648–50 660, 2023. **2**
- [36] J. Li and M. Bansal, “Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation,” *Advances in Neural Information Processing Systems*, vol. 36, 2024. **2**
- [37] J.-R. Xue, J.-W. Fang, and P. Zhang, “A survey of scene understanding by event reasoning in autonomous driving,” *International Journal of Automation and Computing*, vol. 15, no. 3, pp. 249–266, 2018. **2**
- [38] K. Ritter III and T. L. Chambers, “Three-dimensional modeled environments versus 360 degree panoramas for mobile virtual reality training,” *Virtual Reality*, vol. 26, no. 2, pp. 571–581, 2022. **2**
- [39] Z. Lu, K. Hu, C. Wang, L. Bai, and Z. Wang, “Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 14 211–14 219. **2**
- [40] K. C. Shum, H.-W. Pang, B.-S. Hua, D. T. Nguyen, and S.-K. Yeung, “Conditional 360-degree image synthesis for immersive indoor scene decoration,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4478–4488. **2**
- [41] S. Tang, F. Zhang, J. Chen, P. Wang, and F. Yasutaka, “Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion,” *arXiv preprint 2307.01097*, 2023. **2**
- [42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. **5**
- [43] N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel, “Splicing vit features for semantic appearance transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 748–10 757. **5**

A. Supplementary Content

This supplementary material begins by providing an intuitive explanation for the choice of α . Subsequently, we detail the process of producing target prompts for both real-world and synthesized datasets. Further visual results obtained under different control conditions are then presented. Finally, we showcase additional translated results from using different methods on real-world and synthesized 360-degree panoramic images.

Explanation for the choice of α . To intuitively explain the choice of the split constant α , Figure 10 visually depicts the cropping process in 360PanT at denoising step t (where $t \in \{T, T-1, \dots, 1\}$) for three distinct α values. The top row displays the input 360-degree panorama I_{in} and a diagram of the cropping operations based on the sliding window mechanism employed in the seamless tiling translation with spatial control. Each cropped patch, including the special cropped patch (stitch patch), then undergoes independent denoising guided by a target prompt. Subsequent rows highlight the cropped patches matching I_{in} during the sliding window process, indicated by red or yellow dashed boxes. Observe that when $\alpha = W$ or $\alpha = \frac{W}{2}$, two cropped patches matching I_{in} but in different locations are denoised at each step t . Conversely, when $\alpha = \frac{3W}{4}$, only a single cropped patch matching I_{in} undergoes denoising at each step. Crucially, the continuity of boundaries of these highlighted patches are not considered during denoising. Consequently, at each denoising step t , the fewer cropped patches matching I_{in} are denoised, the better the boundary conti-

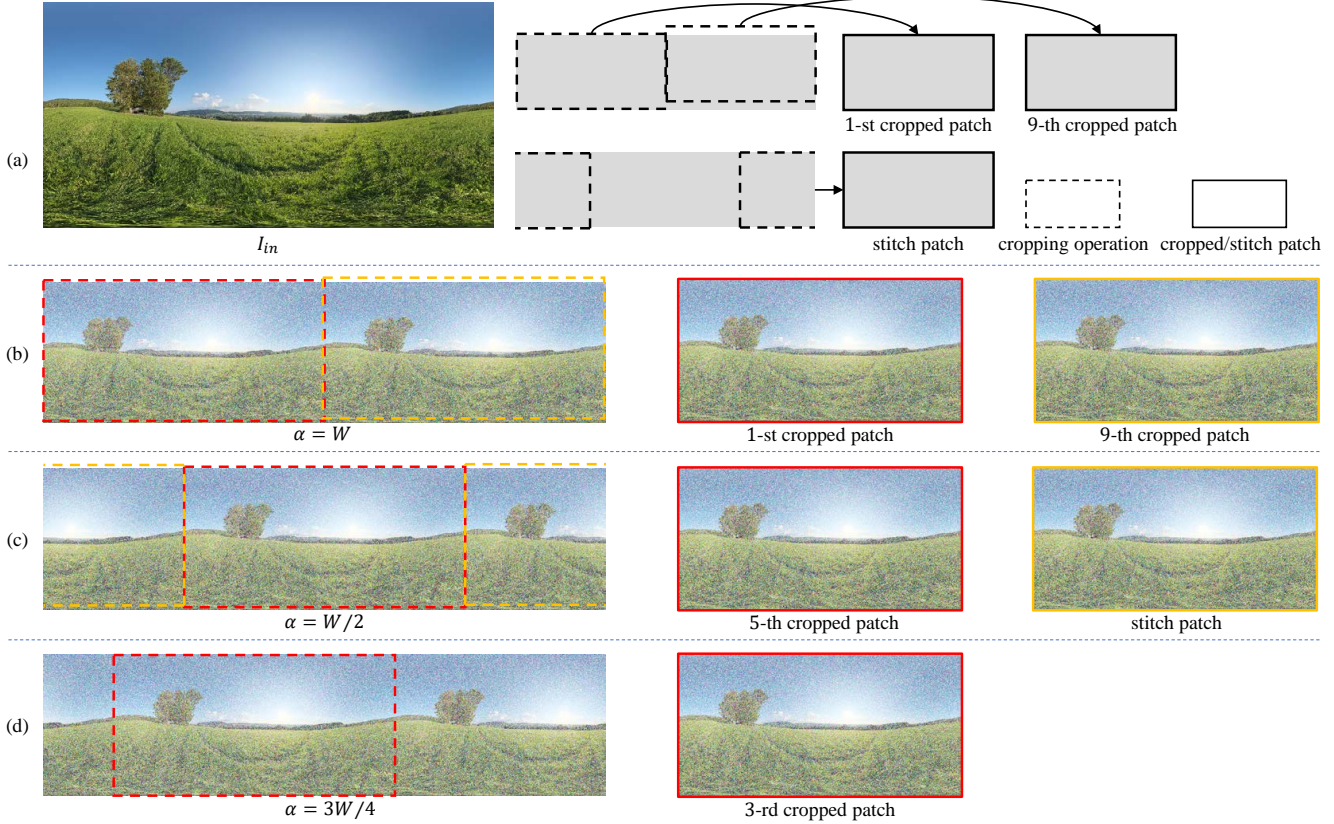


Figure 10. **Intuitive explanation for the choice of split constant α .** The cropped patches matching I_{in} during the sliding window process are highlighted by red or yellow dashed boxes. Note that the stitch patch is a special cropped patch. At each denoising step t , when $\alpha = W$ in (b) or $\alpha = \frac{W}{2}$ in (c), two cropped patches matching I_{in} but in different locations are denoised. Conversely, when α is set to $\frac{3W}{4}$ in (d), only one cropped patch matching I_{in} undergoes denoising. To ensure better boundary continuity in the final translated result, we choose to set α to $\frac{3W}{4}$.

nuity of the final translated 360-degree panorama. Therefore, we set α to $\frac{3W}{4}$ in this paper, which results in a final translated 360-degree panorama with seamlessly connected boundaries, effectively avoiding local visible cracks.

Generation process of target prompts. Figure 11 illustrates the target prompt generation process for each real-world 360-degree panorama within the *360PanorI-Pan2Pan* dataset. Utilizing a consistent template, “a photo of {image name}”, an original prompt is constructed for each 360-degree panoramic image. Subsequently, a target prompt is formulated by combining a randomly selected translation type with the original prompt. Figure 12 depicts the analogous process for the *360syn-Pan2Pan* dataset comprising synthesized 360-degree panoramas. Initially, 120 synthesized 360-degree panoramas are generated using a text-to-360-degree panorama model [6] guided by 120 original prompts. Similar to the real-world dataset, each target prompt consists of a randomly chosen translation type and its corresponding original prompt.

Translation using other control conditions. Diverse

control conditions are extracted from corresponding 360-degree panoramic images using the methods described in FreeControl [3]. If a control condition lacks continuous boundaries, the translated result by our 360PanT (F) will exhibit noticeable content inconsistency at the boundaries. For instance, Figure 13 illustrates how using an extracted depth map I_{in} with discontinuous boundaries as input leads to visible cracks in the extended input map \hat{I}_{in} . Consequently, the translated image by 360PanT (F) shows content inconsistency in the stitched area. In contrast, we observe that extracted Canny edge maps and segmentation masks effectively maintain continuous boundaries. As shown in Figure 14, when using them as control conditions, FreeControl fails to preserve boundary continuity, but our 360PanT (F) consistently produces translated 360-degree panoramas with continuous boundaries, regardless of the input conditions.

Visual results of various methods. To further demonstrate the efficacy of 360PanT for 360-degree panorama translation, we present additional visual comparisons with SDEdit

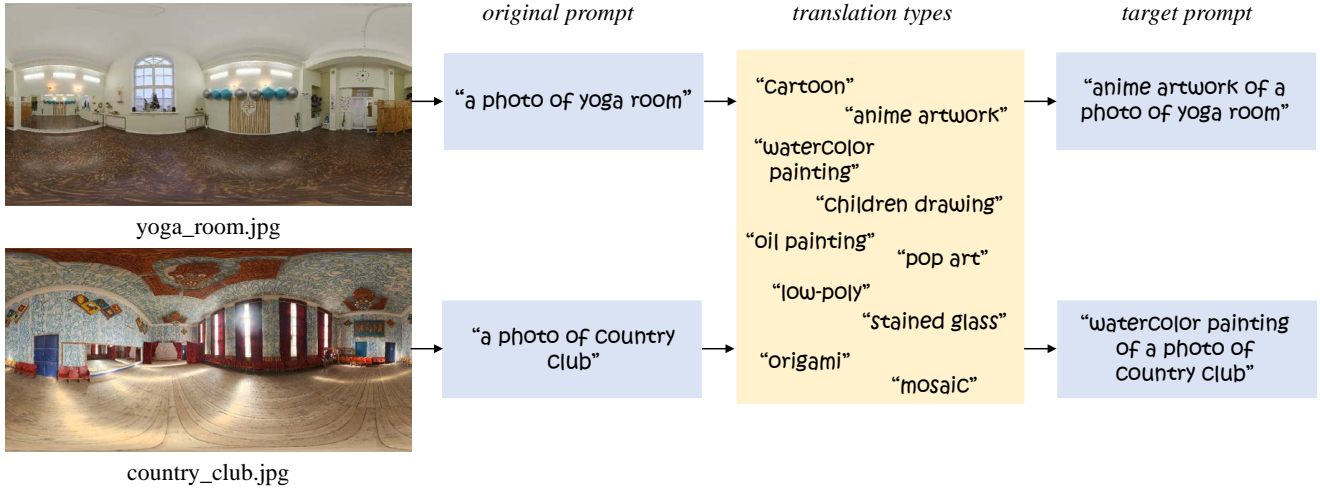


Figure 11. **Target prompt generation for real-world 360-degree panoramas within the 360Panol-Pan2Pan dataset.** Our 10 translation types are presented. A target prompt is formulated by combining a randomly selected translation type with the original prompt.



Figure 12. **Target prompt generation for synthesized 360-degree panoramas in the 360syn-Pan2Pan dataset.** Each target prompt consists of a randomly chosen translation type and its corresponding original prompt.

[5], Pix2Pix-zero [29], P2P [2], PnP [1] and FreeControl [3] on both real-world and synthesized 360-degree panoramas. As illustrated in Figures 15, 16, 17, 18, 19, and 20, 360PanT outperforms these methods in maintaining visual continuity at the boundaries while also adhering to the structure and semantic layout of the input 360-degree panoramic images.

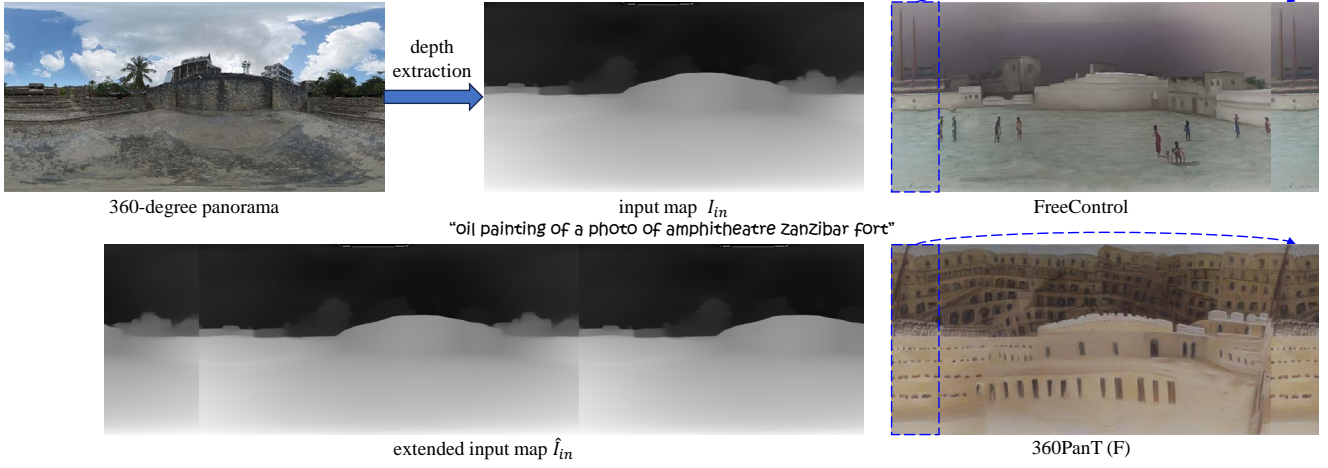


Figure 13. **Depth map with discontinuous boundaries as the control condition.** The boundaries of depth map I_{in} extracted from the 360-degree panorama are not continuous, resulting in visible cracks in the extended input map \hat{I}_{in} . In this situation, the translated panorama by our 360PanT (F) exhibits content inconsistency in the stitched area.

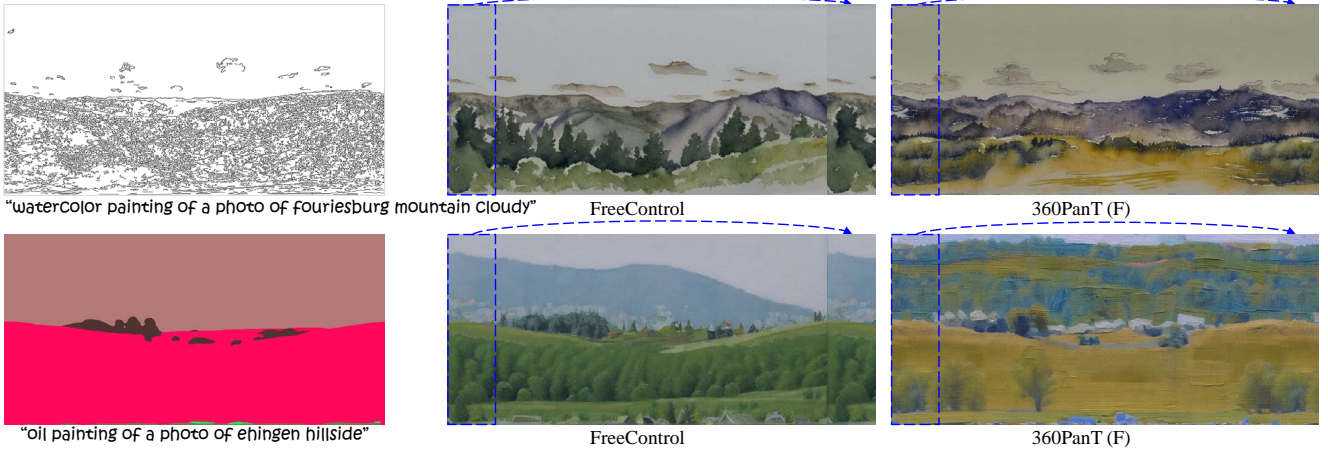


Figure 14. **Visual results using other control conditions.** The extracted Canny edge map and segmentation mask can both effectively maintain continuous boundaries. When using them as control conditions, respectively, FreeControl is unable to guarantee the boundary continuity of the translated panoramas. In contrast, our 360PanT (F) enables the translated 360-degree panoramas with continuous boundaries regardless of the input conditions.



Figure 15. **Visual results on real-world 360-degree panorama.** To easily identify visual continuity or discontinuity at the boundaries, we copy the left area of the panorama indicated by the blue dashed box and paste it onto the rightmost side of the image.

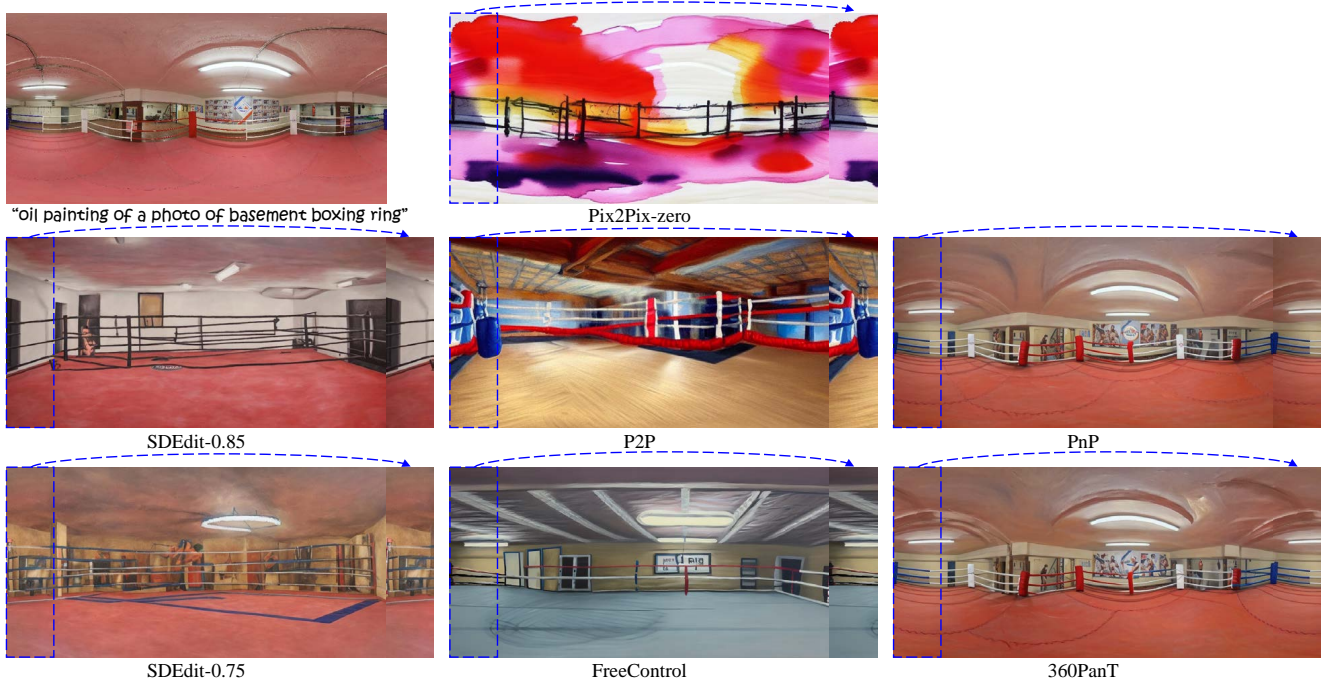


Figure 16. **Visual results on real-world 360-degree panorama.**

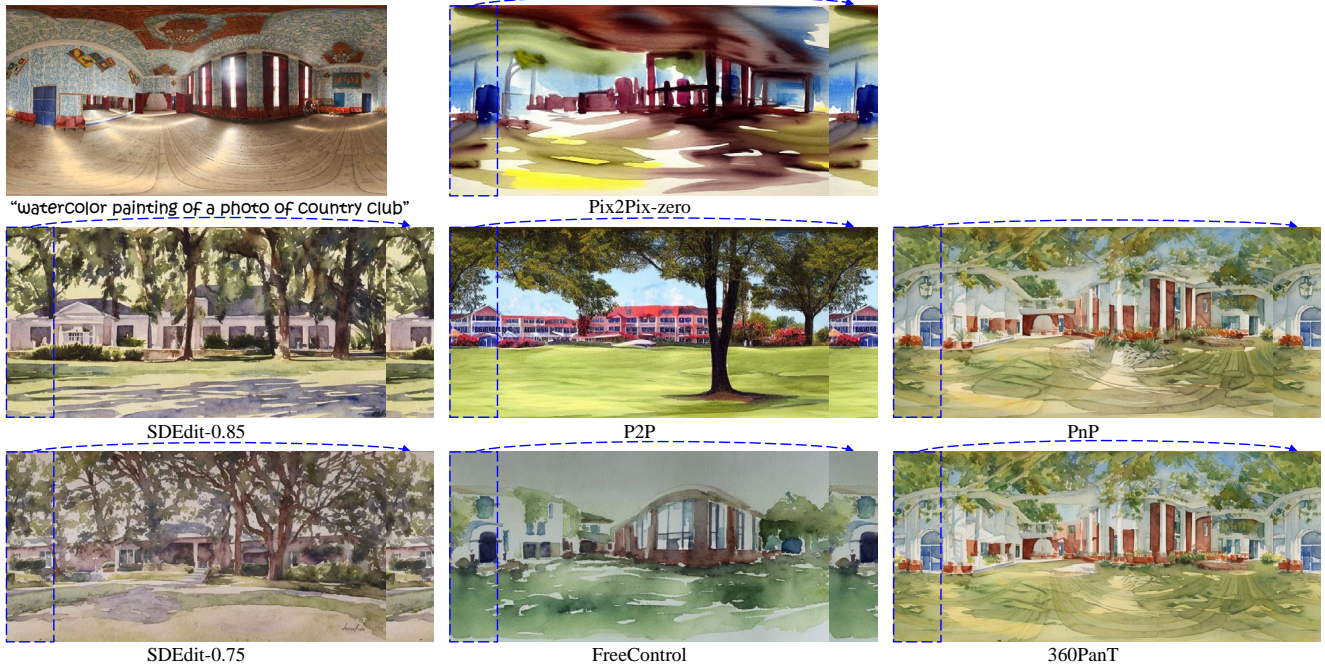


Figure 17. Visual results on real-world 360-degree panorama.

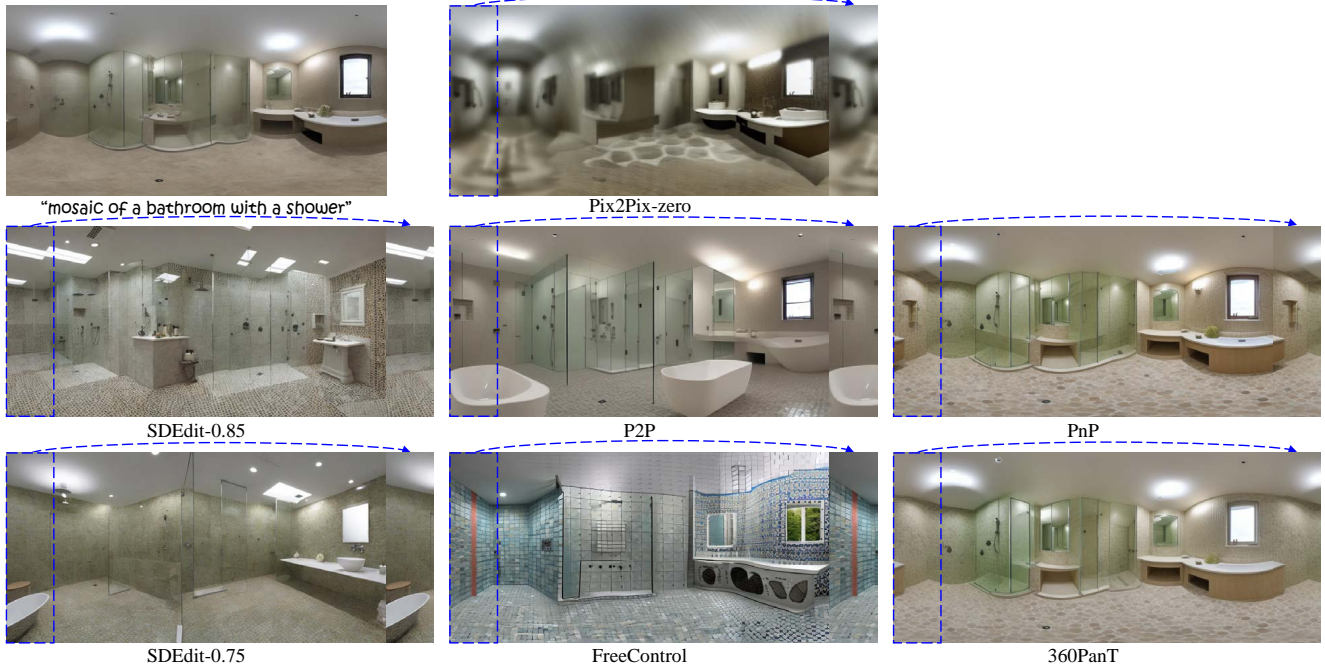


Figure 18. Visual results on synthesized 360-degree panorama. To easily identify visual continuity or discontinuity at the boundaries, we copy the left area of the panorama indicated by the blue dashed box and paste it onto the rightmost side of the image.

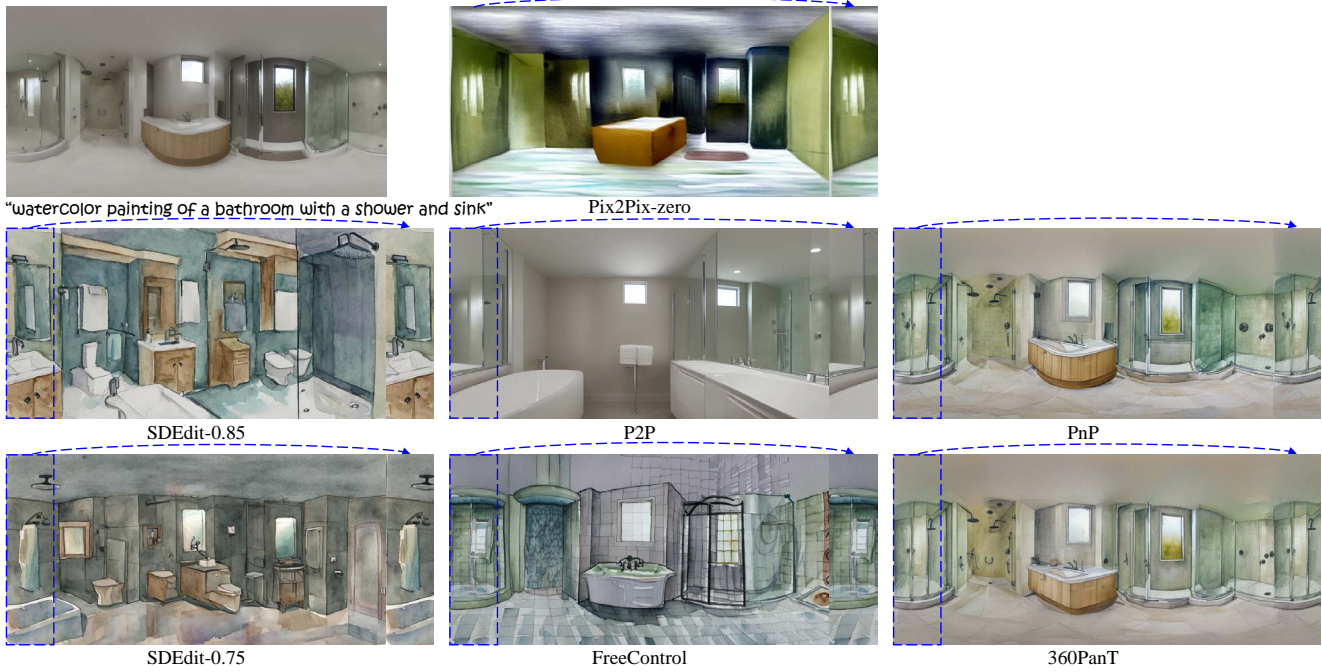


Figure 19. Visual results on synthesized 360-degree panorama.



Figure 20. Visual results on synthesized 360-degree panorama.