

Taming Stable Diffusion for Text to 360° Panorama Image Generation

Cheng Zhang^{1,3} Qianyi Wu¹ Camilo Cruz Gambardella^{1,3} Xiaoshui Huang^{2*}
Dinh Phung¹ Wanli Ouyang² Jianfei Cai^{1*}

¹Monash University ²Shanghai AI Laboratory ³Building 4.0 CRC, Caulfield East, Victoria, Australia



“A living room with a ceiling fan.”

“A historic lighthouse perched on a rugged coastline, overlooking the vast expanse of the open sea.”

Figure 1. Our PanFusion can generate realistic and consistent 360° horizontal by 180° vertical FOV panoramas from a single text prompt, compared to the limited FOV of current state-of-the-art method MVDiffusion [47]. Left: PanFusion addresses the problem of repetitive elements (duplicated “ceiling fans”) and inconsistency (the ceiling and wall in the center) of MVDiffusion. Right: While trained mostly on indoor scenes, PanFusion can generalize well to out-of-domain outdoor prompts with more reasonable layout.

Abstract

Generative models, e.g., *Stable Diffusion*, have enabled the creation of photorealistic images from text prompts. Yet, the generation of 360-degree panorama images from text remains a challenge, particularly due to the dearth of paired text-panorama data and the domain gap between panorama and perspective images. In this paper, we introduce a novel dual-branch diffusion model named *PanFusion* to generate a 360-degree image from a text prompt. We leverage the stable diffusion model as one branch to provide prior knowledge in natural image generation and register it to another panorama branch for holistic image generation. We propose a unique cross-attention mechanism with projection awareness to minimize distortion during the collaborative denoising process. Our experiments validate that *PanFusion* surpasses existing methods and, thanks to its dual-branch structure, can integrate additional constraints like room layout for customized panorama outputs. Code

is available at <https://chengzhag.github.io/publication/panfusion>.

1. Introduction

Creating a 360° panorama image from textual prompts is a nascent yet pivotal frontier in computer vision, with profound implications for applications that require extensive environmental representation, such as environmental lighting [1, 4, 51], VR/AR [60, 61], autonomous driving [59], and visual navigation [17]. Despite recent significant strides in text-to-image synthesis, the leap to generating full 360° horizontal by 180° vertical field-of-view (FOV) panorama remains challenging.

There are two major hurdles for achieving this goal. The first hurdle is data scarcity. The availability of text-to-panorama image pairs [20, 47] is significantly less compared with the abundance of text-to-common image pairs [38, 39]. The dearth of data complicates the training and finetuning of generative models. The second hurdle

*Corresponding author.

lies in the geometric and domain variations. Panorama images are distinct not only in their aspect ratio (2 : 1) but also in the underlying equirectangular projection (ERP) geometry [58]. This differs significantly from typical square images of the perspective projection that are used in most generative model training [38, 39].

To mitigate the scarcity of panorama-specific training data, the previous solutions follow a common principle that *leverages the prior knowledge of the pre-trained generative model* [17, 20, 47]. However, taming powerful models like stable diffusion [31, 34] to generate high-fidelity panorama images remains a non-trivial task. Early attempts turn to formulate the 360-degree generation as an iterative image inpainting or warping process [17, 20]. Such solutions suffer from error accumulation and fail to handle loop closure [47]. To address this, MVDiffusion [47] proposes to produce multiple perspective images simultaneously by introducing a correspondence-aware attention module to facilitate multiview consistency, and then stitch together the perspective images to form a complete panorama. Despite the improved performance, the pixel-level consistency between neighboring perspectives in MVDiffusion cannot ensure global consistency, often resulting in repetitive elements or semantic inconsistency, as illustrated in Fig. 1.

Therefore, in this paper, we propose a novel dual-branch diffusion model called *PanFusion* that is tailored to address the limitations of prior models for high-quality text to 360° panorama image generation. Specifically, PanFusion is designed to operate in both panorama and perspective domains, employing a global branch for creating a coherent panoramic “canvas” and a local branch that focuses on rendering detail-rich multiview perspectives. The local-global synergy of PanFusion significantly improves the resulting panoramas against the prevalent issues of error propagation and visual inconsistency that prior models have struggled with. To enhance the synergy between the two branches, we further propose an Equirectangular-Perspective Projection Attention (EPPA) mechanism, which respects the equirectangular projection for maintaining geometric integrity throughout the generation process. In addition, our adoption of parametric mapping for positional encoding is another leap forward, enhancing the model’s spatial awareness and further ensuring the consistency of the generated panorama. Moreover, the panorama branch of PanFusion can be easily leveraged to accommodate supplementary control inputs at the panorama level, such as room layout, allowing for the creation of images that adhere to precise spatial conditions. We summarize our primary contributions as follows.

- We pioneer a dual-branch diffusion model PanFusion, harnessing both the global panorama and local perspective latent domains, to generate high-quality, consistent 360° panoramas from text prompts.

- To enhance the interaction between the two branches, we introduce an “equirectangular-perspective projection attention” mechanism that establishes a novel correspondence between global panorama and local perspective branches, addressing the unique projection challenges of panorama synthesis.
- Our PanFusion not only surpasses prior models in quality and consistency but also supports extended control over the generation process with the inclusion of room layout. Extensive experimental results demonstrate the superiority of our proposed framework.

2. Related Work

Diffusion models. In recent years, diffusion models [6, 12, 41, 42, 44, 62] have taken the world of image generation by storm, as they have become faster [15, 19, 42] and more capable in terms of image quality and resolution [28, 31, 36]. This success has prompted the development of various applications for diffusion models, such as text-to-image [25, 28, 31, 36], image-conditioned generation [24, 64], in-painting [21, 35] and subject-driven generation [10, 33]. Most of these applications try to exploit the prior knowledge of a pre-trained diffusion model to mitigate the scarcity of task-specific data, by either finetuning with techniques like LoRA [14], or introducing auxiliary modules to distill the knowledge. We also adopt the same principle to harness the power of a pre-trained latent diffusion model [31] for panorama image generation.

Panorama generation. Panorama image generation encompasses different settings, including panorama outpainting and text-to-panorama generation. Panorama outpainting [1, 5, 26, 51, 54, 56] focus on generating a 360-degree panorama from a partial input image. Various methods, such as StyleLight [51] and BIPS [26], have addressed specific use cases, focusing on HDR environment lighting and robotic guidance scenarios. Recent works [54, 56] have improved realism with diffusion models, but often lack the exploitation of rich prior information from pre-trained models, limiting generalization. On the other hand, recent developments in generative models have opened new frontiers in synthesizing immersive visual content from textual input [8, 13, 43, 52, 53, 55, 56, 61, 63]. As an image-based representation, generating panorama from text has gained much attention. Text2Light [4] adopt the VQGAN [7] structure to synthesize an HDR panorama image from text. To generate in arbitrary resolution with pre-trained diffusion models, DiffCollage [65], MultiDiffusion [2] and SyncDiffusion [16] propose to fuse the diffusion paths, while PanoGen [17] solves by iteratively inpainting. However, they failed to model the equirectangular projection of 360-degree panoramas. Lu et al. [20] adopts an autoregressive framework, but suffers from inefficient issues. MVDiffusion [47] designs a correspondence-aware attention module

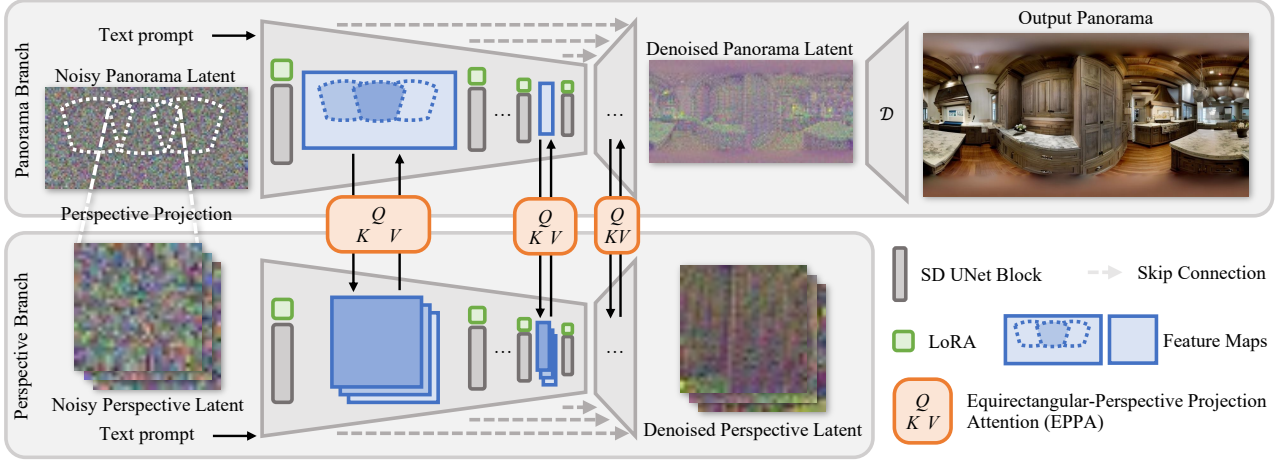


Figure 2. Our proposed dual-branch PanFusion pipeline. The panorama branch (upper) provides global layout guidance and registers the perspective information to get seamless panorama output. The perspective branch (lower) harnesses the rich prior knowledge of Stable Diffusion (SD) and provides guidance to alleviate distortion under perspective projection. Both branches employ the same UNet backbone with shared weights, while finetuned with separate LoRA layers. Equirectangular-Perspective Projection Attention (EPPA) modules are plugged into different layers of the UNet to pass information between the two branches.

to produce multi-view images simultaneously that can be stitched together but results in repetitive elements and inconsistency. In contrast, our proposed PanFusion, a dual-branch framework, addresses the limitations of existing methods by considering both global panorama views and local perspective views, providing a comprehensive solution for text-driven 360-degree panorama image generation.

3. Method

3.1. Preliminary

First introduced in [41], Diffusion models [12, 42, 44] aim to generate images from a noise distribution by iterative denoising with a learned prior distribution. However, the early diffusion models operate in the image space, which is of high dimension and complex. We have recently witnessed the huge success of latent diffusion models [31] that first transform an image x to a latent representation z with a learned encoder \mathcal{E} and then train a UNet [32] model ϵ_θ parameterized by θ for the reverse process in the latent space, formulating the training objective as:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x), t, \epsilon, y} [\|\epsilon - \epsilon_\theta(z_t, t, \tau(y))\|_2], \quad (1)$$

where $\tau(y)$ is an encoding of input condition y (e.g., text prompt or image), z_t is the latent map at time step t , and ϵ is sampled from a Gaussian noise. To sample from the model, z_t is first initialized from standard Gaussian distribution, then the reverse process is applied iteratively to generate z_0 , which is finally decoded into image space with a decoder \mathcal{D} .

3.2. Dual-Branch Diffusion Model

Directly employing pre-trained latent diffusion models [31], e.g., Stable Diffusion (SD) [34], to generate panorama from

multiple perspective images, either in an iterative manner [9, 13, 17] or in a synchronized way [2, 47], would fail to handle loop closure [47] or produce repetitive elements (Fig. 1) due to lack of global understanding. To address this issue, we propose a dual-branch diffusion model, which consists of a panorama branch and a perspective branch both based on the UNet of SD as shown in Fig. 2. The panorama branch is designed to provide global layout guidance and to register the perspective information to get the final panorama without stitching, while the perspective branch is designed to exploit the rich perspective image-generation capabilities of SD and to provide guidance to alleviate distortion under perspective projection. The two branches work together during the diffusion process to generate a denoised panorama latent map. Finally, this latent map runs through the pre-trained decoder \mathcal{D} of SD to produce the final panorama image.

Panorama Branch. Given a text prompt y , our goal is to generate a panorama $x \in \mathbb{R}^{3 \times H \times W}$, where $W = 2H$, $H = 512$. To account for the difference between the target resolution and the one that SD is trained on, we introduce LoRA [14] layers to adapt the model to the new resolution. SD with LoRA can already serve as a strong baseline for panorama generation, but the results are not loop-consistent. Previous works [8, 56] have attempted to overcome this problem using 90 degree rotations of the latent map at each diffusion step. However, the seams are still obvious [56]. On close inspection of the SD model we have observed that the loop inconsistency is mainly caused by the convolutional layers in the UNet backbone, due to the lack of a mechanism to pass information between the two ends of the panorama image. Therefore, we introduce an adaptation to the UNet by adding additional circular padding [40, 54, 66] before each convolutional layer, and then cropping the out-

put feature maps to the original size. In addition, we also add circular padding to the latent map before decoding to mitigate the less apparent loop inconsistency caused by the decoder. The combination of the techniques outlined above – latent rotation and circular padding – enable the generation of loop-consistent results with negligible computational cost, and can thus serve as another strong baseline. However, these alone do not make full use of the perspective generation capabilities that SD possess.

Perspective Branch. The perspective branch aims to exploit the outstanding capabilities that SD has shown in the generation perspective images. Since it does not generate panoramas directly, it can operate at a lower resolution. Specifically, we set its resolution to $H/2 \times H/2$ and add LoRA layers to adapt to the new resolution. To evenly distribute perspective cameras and fully cover the panorama image, we sample $N = 20$ cameras with poses $R^i \in SO(3), i \in 1, \dots, N$ on an icosahedron, similar to [27, 30], and set the FOV = 90° . We input the same prompt y to both branches, and leave it to the model to decide how to exploit the prompt.

Joint Latent Map Initialization. Previous work [22] has shown that the latent map initialization can affect the layout of the generated image, *i.e.*, the initial noise can be modified to manipulate the layout of the generated image. We find this is particularly important for multi-image generation with correspondence, as the overlapping regions in different views tend to generate different elements if noise is sampled independently. Under this observation, we propose to jointly sample noise for the panorama and perspective latent maps by projecting noise from the panorama to the perspective. To be more specific, we first initialize panorama latent map $z_T^* \in \mathbb{R}^{4 \times H/f \times W/f}$ as Gaussian noise, then initialize perspective latent maps with $z_T^i = P(z_T^*, R^i, \text{FOV}, (H/2f, H/2f))$, $z_T^i \in \mathbb{R}^{4 \times H/2f \times H/2f}$, where $P(\cdot)$ is the function that projects z_T^* to a perspective view, and f is the down-sampling factor the encoder \mathcal{E} used in SD. We use nearest interpolation for projection as it introduces fewer artifacts than bilinear interpolation.

3.3. EPP Attention

To pass the guidance between the perspective and panorama branches, a naive solution is to project the feature maps in different layers from one branch to the other, similar to [47]. However, this will introduce information loss during interpolation and will restrict the receptive field to a small region around corresponding pixels. Instead, we propose an Equirectangular-Perspective Projection Attention (EPPA) module to implicitly pass the guidance in a cross-attention way as shown in Fig. 2. The EPPA operates on feature maps in different layers of two UNet branches, denoted as $F^* \in \mathbb{R}^{c \times h \times w}$ and $F^i \in \mathbb{R}^{c \times h/2 \times h/2}$, where c is the channel dimension, h and w are the height and width

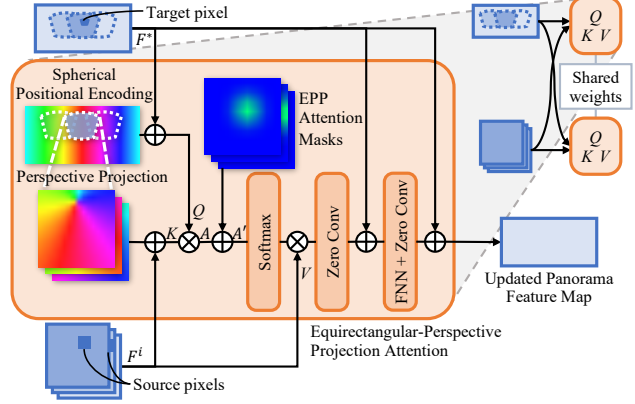


Figure 3. Equirectangular-Perspective Projection (EPP) Attention. As EPP attention module is designed to be bijective to pass information in both directions, we only illustrated the direction of registering perspective information to panorama.

of the feature maps, respectively. In addition, while cross-attention bypasses interpolation and provides benefits from the global receptive field of the attention mechanism, it is unaware of the projection between different formats. To address this issue, as shown in Fig. 3, we introduce two key components: EPP spherical positional encoding and EPP attention mask.

EPP Spherical Positional Encoding. Since the EPPA module tries to associate two different formats, it is important to add positional information in the same space so that the attention mechanism can learn the correspondence. We achieve this by introducing Spherical Positional Encoding (SPE) [4, 23] to the EPPA module. The $\text{SPE}(\theta, \phi) = (\gamma(\theta), \gamma(\phi))$ function maps polar coordinates (θ, ϕ) to a higher dimensional space \mathbb{R}^{4L} with Fourier positional encoding:

$$\gamma(\theta) = \left[\sin(2^0 \pi \theta), \cos(2^0 \pi \theta), \dots, \sin(2^{L-1} \pi \theta), \cos(2^{L-1} \pi \theta) \right]. \quad (2)$$

Here we set $L = c/4$ so that $\text{SPE}(\theta, \phi) \in \mathbb{R}^c$. In the EPPA module, we apply SPE by first computing an SPE map for the panorama feature map, then project it to each perspective feature map with projection function $P(\cdot)$, so that corresponding pixels in different formats share the same SPE vector. Finally, the SPE maps are added to the feature maps accordingly and go through a linear layer to get the query Q and key K , following a matrix product to get the affinity matrix A . Take the perspective-to-panorama direction as an example as shown in Fig. 3, where $Q \in \mathbb{R}^{c \times h \times w}$ and $K \in \mathbb{R}^{N \times c \times h/2 \times h/2}$ are reshaped and multiplied to get the affinity matrix $A \in \mathbb{R}^{hw \times Nh^2/4}$.

EPP Attention Mask. In addition to the SPE, we also propose an EPP attention mask to encourage the attention mechanism to focus around the corresponding pixels, inspired by [48]. For example, for a target pixel in the panorama feature map as in Fig. 3, we would like to

focus on registering the information from corresponding source pixels in the perspective feature maps. We achieve this by enhancing the affinity matrix A with a soft mask $M \in \mathbb{R}^{hw \times Nh^2/4}$ highlighting the corresponding pixels. To generate M , we first use $P(\cdot)$ to project a binary mask $M_{j,k}$ for each pixel (j, k) in panorama feature map to each perspective view i . Then we apply a Gaussian kernel to smooth the masks and normalized them to $[-1, 1]$. The masks are stacked and reshaped into M , which is then added to A to get the enhanced affinity matrix A' . The rest goes the same as the vanilla attention mechanism, where a softmax function is applied to A' to get the attention weights, which are finally multiplied with the value V to get the output.

Inspired by [47, 64], we add zero-initialized 1×1 convolutional layers to the output of cross-attention and add it as a residual to the target feature map. This ensures the UNet stays unmodified at the beginning of training and can be gradually adapted to the EPPA modules. We add independent EPPA modules after each down-sampling layer and before each up-sampling layer of UNet to connect two branches, detailed in Sec. A of the supplementary material. Considering that the guidance is passed in both directions following the same equirectangular-perspective projection rule, which is bijective in nature, we share the weights of the EPPA modules in the two directions.

3.4. Layout-Conditioned Generation

One of the most important applications for panorama generation is to generate panorama according to a given room layout. This is particularly useful for panorama novel view synthesis [57] and can potentially benefit indoor 3D scene generation [8, 13]. However, this problem is not well researched for diffusion-based panorama generation, largely due to the difficulty of introducing layout constraints while exploiting the rich prior knowledge of SD in perspective format at the same time. For panorama generation from multi-view [17, 47], one naive solution is to project layout condition into different views to locally condition the generation of perspective images. Instead, for our dual-branch diffusion model, we can naturally exploit the global nature of the panorama branch to enforce a much stronger layout constraint. Specifically, we render the layout condition as a distance map, then use it as the input of a ControlNet [64] to condition the panorama branch.

3.5. Training

During training, we employ the same technique for latent map initialization to jointly sample noise for the panorama and perspective view, which we denote as ϵ^* and ϵ^i , respectively. Given a GT panorama x^* , we apply supervision on the panorama branch with the loss in Eq. (1):

$$\mathcal{L}^* = \mathbb{E}_{\mathcal{E}(x^*), t, \epsilon^*, y} [\|\epsilon^* - \epsilon_\theta^*(z_t^*, t, \tau(y))\|_2], \quad (3)$$

where ϵ_θ^* is the predicted noise from the panorama branch. To encourage the synchronization between the two branches, we also apply supervision on the perspective branch with the following loss:

$$\mathcal{L}^i = \mathbb{E}_{\mathcal{E}(x^i), t, \epsilon^i, y} [\|\epsilon^i - \epsilon_\theta^i(z_t^i, t, \tau(y))\|_2], \quad (4)$$

where x^i is a perspective image projected from x^* and ϵ_θ^i is the predicted noise from the perspective branch. We joint train the EPPA modules and the LoRA layers in the two branches by combining the above two losses as $\mathcal{L} = \mathcal{L}^* + \frac{1}{N} \sum_{i=1}^N \mathcal{L}^i$. Note that the SD UNet blocks remain frozen.

4. Experiment

4.1. Experimental Setup

Dataset. We follow the MVDiffusion [47] to use the Matterport3D dataset [3], which has 10,800 panoramic images with 2,295 room layout annotations. We employ BLIP-2 [18] to generate a short description for each image.

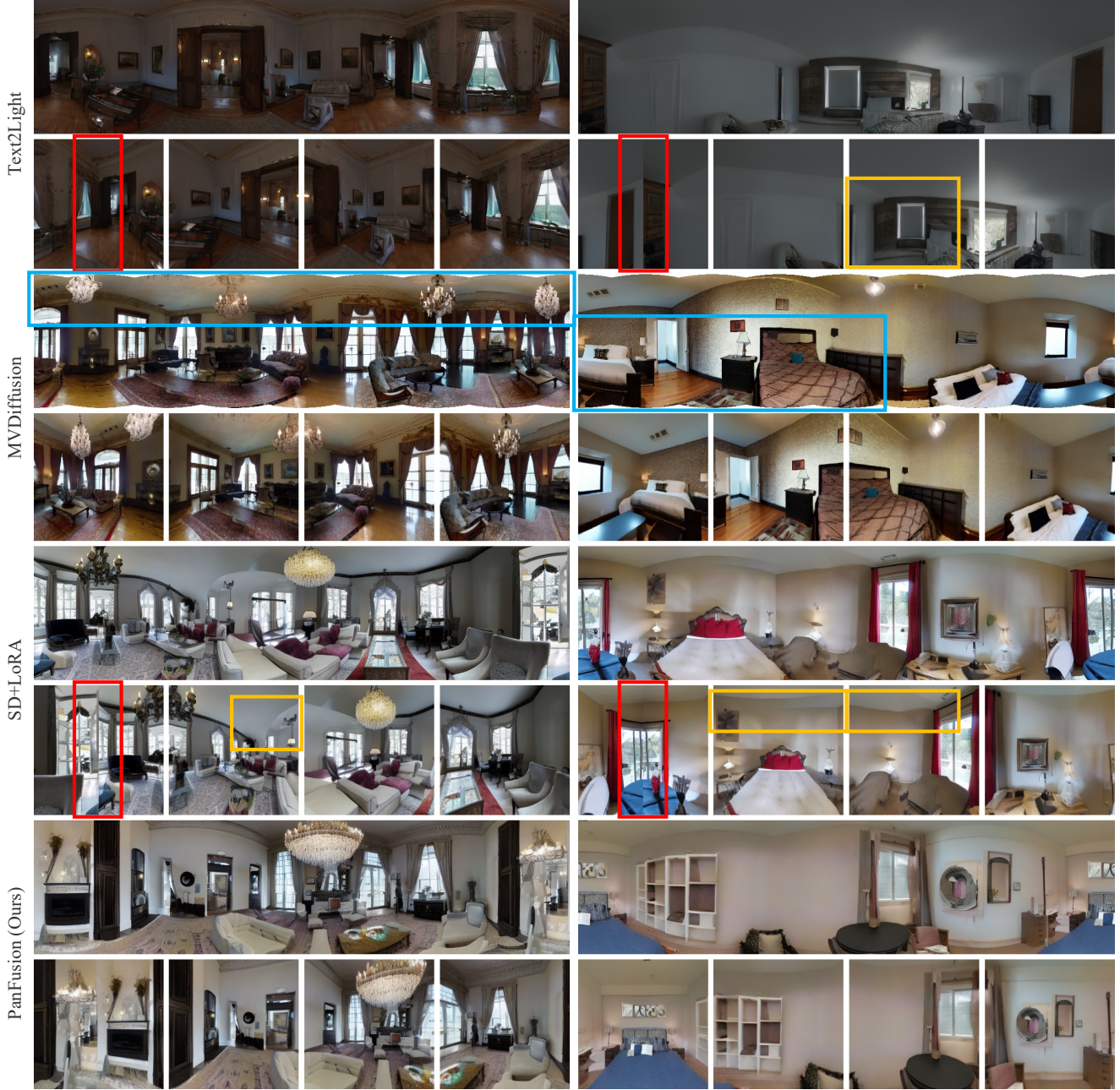
Implementation Details. For text-conditioned generation, the training and inference schedules are kept the same as MVDiffusion [47] to make a fair comparison. For text-layout conditioned generation, we train the additional ControlNet with other parameters fixed.

Evaluation Metrics. Following previous works, we evaluate image quality in panorama [4] and perspective [47] domain. For layout-conditioned generation, we propose a new metric to evaluate how well the generated panorama follows input layout. Specifically, we use the following metrics:

- *Panorama.* We follow Text2Light [4] to report Fréchet Inception Distance (FID) [11] and Inception Score (IS) [37] on panoramas to measure realism and diversity. Additionally, CLIP Score (CS) [29] is used to evaluate the text-image consistency. While FID is widely used for image generation, it relies on an Inception network [46] trained on perspective images, thus less applicable for panorama images. Therefore, a variant of FID customized for panorama, Fréchet Auto-Encoder Distance (FAED) [26], is used to better compare the realism.

- *Perspective.* To simulate the real-world scenario where the user can freely navigate a panorama by viewing from different perspective views, we also report FID and IS for 20 randomly sampled views to compare with methods that generate 180° vertical FOV. We also follow MVDiffusion [47] to report FID, IS and CS scores on 8 horizontally sampled views. It is worth noting that this group of metrics favors MVDiffusion by measuring its direct outputs, while ours involve interpolation for perspective views.

- *Layout Consistency.* We propose a layout consistency metric, which employs a layout estimation network HorizonNet [45] to estimate the room layout from the generated panorama and then compute its 2D IoU and 3D IoU [45] with the input layout condition.



“A living room with a chandelier.”

“A bedroom with a bed and a table.”

Figure 4. Qualitative comparisons of text-conditioned panorama generation. We show panoramas cropped to the vertical FoV of MVDiffusion [47]. Below each panorama, we show 4 evenly spaced perspective projections, with the first view crossing the left and right boundaries. We highlight the **loop inconsistency**, **distorted lines** and **repetitive objects and unreasonable furniture layout** of baseline methods with corresponding colors of boxes, which are addressed by our method. More results can be found in Sec. E of the supplementary.

See Supplementary Sec. B for more details of the above.

4.2. Comparisons with Previous Methods

Baselines. We compare our PanFusion with the following baselines (see Supplementary Sec. B for details).

- *MVDiffusion* [47] utilizes a multi-view diffusion model to generate 8 horizontal views that can be stitched into a

panorama with vertical FOV of 90° . It requires separate prompts for training while providing an option to generate from a single prompt.

- *Text2Light* [4] generates a 180° vertical FOV panorama from a text prompt in a two-stage auto-regressive manner.
- *SD+LoRA* is our baseline model that uses LoRA [14] to finetune a Stable Diffusion [31] on panorama images.

Method	Panorama				20 Views		Horizontal 8 Views [47]		
	FAED ↓	FID ↓	IS ↑	CS ↑	FID ↓	IS ↑	FID ↓	IS ↑	CS ↑
Text2Light [4]	97.24	76.5	3.60	27.48	36.25	5.67	43.66	4.92	<u>25.88</u>
MVDiffusion [47]	-	-	-	-	-	-	25.27	6.90	26.34
SD+LoRA [14, 31]	<u>7.19</u>	51.69	<u>4.40</u>	28.83	<u>19.32</u>	<u>6.90</u>	20.68	6.48	24.77
Pano Branch	7.90	<u>50.40</u>	4.54	<u>28.67</u>	20.10	7.06	<u>20.56</u>	6.37	24.85
PanFusion (Ours)	6.04	46.47	4.36	28.58	17.04	6.85	19.88	<u>6.50</u>	24.98

Table 1. Comparison with SoTA methods. We evaluate the panorama image quality with Fréchet Auto-Encoder Distance (FAED), Fréchet Inception Distance (FID), Inception Score (IS), and CLIP Score (CS). In addition, we evaluate the perspective image quality in two settings. We first randomly sample 20 views, which is the closest to the real-world scenario where the user can freely navigate the panorama to view the scene from different perspectives. We then follow MVDiffusion [47] to horizontally sample 8 evenly spaced views.

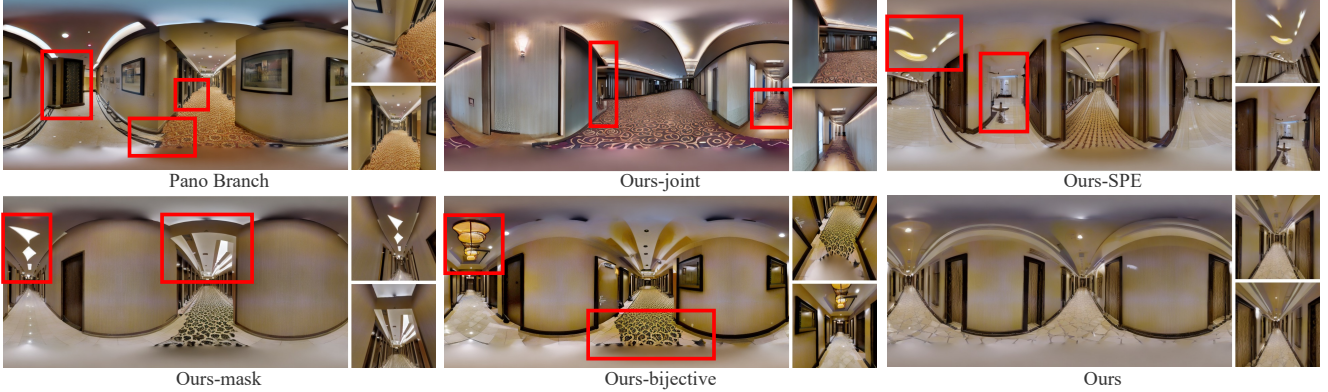


Figure 5. Ablation study. Artifacts are highlighted with red boxes and projected to perspective views. Prompt: “A hallway in a hotel.”

Method	Panorama				20 Views	
	FAED ↓	FID ↓	IS ↑	CS ↑	FID ↓	IS ↑
Pano Branch	7.90	50.40	4.54	28.66	20.10	<u>7.06</u>
Ours-joint	7.75	59.43	4.34	27.94	29.68	6.82
Ours-SPE	<u>6.63</u>	49.55	4.65	28.66	20.82	7.00
Ours-mask	6.77	45.49	4.35	28.66	15.81	6.81
Ours-bijjective	7.36	48.35	<u>4.58</u>	28.66	18.78	7.12
Ours	6.04	<u>46.47</u>	4.36	<u>28.58</u>	<u>17.04</u>	6.85

Table 2. Ablation study. We compare the ablated versions of our method on both panorama and perspective domains. Here, “-” indicates removing the subsequent component.

- *Pano Branch* is SD+LoRA with additional modifications as described in Sec. 3.2 to ensure loop consistency.

Quantitative Results. Tab. 1 shows the quantitative comparison results. Here, we assign the highest value to realism in image generation, measured in FAED and FID. On these two metrics, our method outperforms baseline methods in both panorama and perspective. For IS, our method’s performance is slightly lower than baselines. This is likely due to the fact that IS evaluates diversity of objects in generated images, using a classifier, and our model, unlike the baselines, tends to not generate unexpected objects. Similarly, it is possible to say that baseline models present slightly higher CS is due to the repetition of objects reinforcing alignment with prompts. Considering SD+LoRA is superior to Pano Branch on FAED and is on par in other metrics, we only qualitatively compare with SD+LoRA below.

Qualitative Results. Fig. 4 shows the qualitative comparison results. **Loop inconsistency** can be observed for Text2Light and SD+LoRA due to lack of message passing between left-right boundaries. They also suffer from **distorted lines** in perspective views, meaning the generated panorama does not follow the correct equirectangular projection. MVDiffusion on the other hand suffers from **repetitive objects and unreasonable furniture layout**, which is likely due to the lack of global context. Our method generates the most realistic scenes and aligns to text condition the best, also with less distortion in perspective views.

4.3. Ablation Study

In Sec. 4.2 and Tab. 1, we show that our full model outperforms Pano Branch, the baseline model of our method without the perspective branch. Here, as shown in Tab. 2 and Fig. 5, we further conduct an ablation study to validate the effectiveness of each component in our method. For a consistent comparison, we keep the layout similar between different ablated versions by sampling the same noise for latent map initialization, exploiting the observation of [22].

Joint latent map initialization. We ablate the joint latent map initialization by initializing the latent maps of the panorama and perspective branches separately (**Ours-joint**). Significant performance drop can be observed in all metrics and qualitative results, demonstrating the importance of joint latent map initialization. Interestingly, Ours-



“A hallway in a building.”

“A living room with a view of the ocean.”

Figure 6. Layout-conditioned generation comparisons. We showcase how layout-conditioned panorama generation can benefit from our dual-branch structure. The input layout is drawn on the generated panorama as yellow lines, while the panorama is cropped to the FoV of MVDiffusion and projected to one view to highlight the layout consistency.

Method	Layout Consistency		Horizontal 8 Views [47]		
	3D IoU \uparrow	2D IoU \uparrow	FID \downarrow	IS \uparrow	CS \uparrow
MVDiffusion [47]	61.06	64.43	28.83	5.60	26.79
PanFusion (Ours)	68.46	71.82	22.58	5.10	26.04
GT image	74.31	77.15	-	-	-

Table 3. Layout-conditioned comparisons. We evaluate layout consistency between the condition and the layout of each generated image extracted by a layout estimator HorizonNet [45]. “GT image” indicates the upper bound of the layout estimator.

joint is even worse than Pano Branch in FID. It is likely due to that joint latent map initialization helps the corresponding pixels to share a similar noise distribution from the beginning of the diffusion process, which is essential for EPPA to align the content of overlapped regions.

EPP SPE and attention mask. We ablate spherical positional encoding (**Ours-SPE**) and attention mask (**Ours-mask**) from EPPA module. From Tab. 2, we can see that missing SPE hurts FAED and FID, which is likely due to that SPE helps the model to learn the relative position of the pixels between the two branches. Missing attention mask gets FID better, but it hurts FAED, which is more accurate in evaluating the panorama quality as it is customized for the target dataset. Both result in clear artifacts around point light sources, inconsistent textures of the floor, and distortions in the highlighted projections, as shown in Fig. 5.

Bijjective EPPA. We ablate the bijjective EPPA (**Ours-bijjective**) by using separate parameters for the two directions in the EPPA module. Both FAED and FID get worse for Ours-bijjective. In addition, the ablated version struggles to generate consistent textures for the floor and ceiling for the two directions of the hallway in Fig. 5. On the contrary, our full model generates floor and ceiling with consistent style, showing a better global understanding of the scene.

4.4. Application: Layout-Conditioned Generation

To showcase the advantage of our method in generating panorama images with additional layout conditions, we build a baseline model by adding a ControlNet to MVDiffusion as described in Sec. 3.4. We render the layout condition into a distance map and then project it to perspective views to condition its generation of multi-view images.

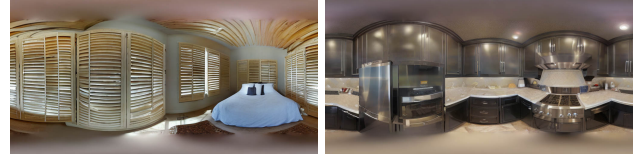


Figure 7. Failure case. Our method sometimes generates indoor scenes with no entrance.

The training settings are kept the same as our PanFusion. As shown in Tab. 3, our method outperforms the baseline model in layout consistency while keeping the realism advantage of perspective projections. We visualize the layout condition as wireframes overlaid on generated panorama images in Fig. 6, where we can see our generated panorama images follow their layout conditions better, especially as highlighted in perspective views. More details can be found in Sec. F of the supplementary material.

5. Conclusion

In this paper, we have proposed PanFusion, a novel text-to-360°-panorama image generation method that can generate high-quality panorama images from a single text prompt. Particularly, a dual-branch diffusion architecture has been introduced to harness prior knowledge of Stable Diffusion in the perspective domain while addressing the repetitive elements and inconsistency issues observed in previous works. An EPPA module has been further introduced to enhance the information passing between the two branches. We have also extended our PanFusion for the application of layout-conditioned panorama generation. Comprehensive experiments have demonstrated that PanFusion can generate high-quality panorama images with better realism and layout consistency than previous methods.

Limitations. Although the dual-branch architecture of PanFusion combines the advantages in both panorama and perspective domains, it comes with a cost of higher computational complexity. Additionally, our method sometimes fails to generate entrance for indoor scenes, as shown in Fig. 7, which is essential for use cases like virtual tour.

Acknowledgement: This research is supported by Building 4.0 CRC and the National Key R&D Program of China

(NO.2022ZD0160101), and was partially done at Shanghai AI Laboratory.

References

- [1] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In *CVPR*, pages 11441–11450, 2022. 1, 2
- [2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, pages 1737–1752. PMLR, 2023. 2, 3
- [3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *3DV*, 2017. 5, 1, 2
- [4] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM TOG*, 41(6):1–16, 2022. 1, 2, 4, 5, 6, 7
- [5] Mohammad Reza Karimi Dastjerdi, Yannick Hold-Geoffroy, Jonathan Eisenmann, Siavash Khodadadeh, and Jean-François Lalonde. Guided co-modulated gan for 360° field of view extrapolation. In *3DV*, pages 475–485. IEEE, 2022. 2
- [6] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 2
- [7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2
- [8] Chuan Fang, Xiaotao Hu, Kunming Luo, and Ping Tan. Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. *arXiv preprint arXiv:2310.03602*, 2023. 2, 3, 5
- [9] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *arXiv preprint arXiv:2302.01133*, 2023. 3
- [10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2
- [11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 5
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 3
- [13] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, pages 7909–7920, 2023. 2, 3, 5
- [14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 3, 6, 7, 4
- [15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, pages 26565–26577, 2022. 2
- [16] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. In *NeurIPS*, 2023. 2
- [17] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. In *NeurIPS*, 2023. 1, 2, 3, 5
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 5, 1, 2
- [19] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, pages 5775–5787, 2022. 2
- [20] Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. *arXiv preprint arXiv:2309.03467*, 2023. 1, 2
- [21] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 2
- [22] Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. Guided image synthesis via initial image editing in diffusion model. In *ACM MM*, pages 5321–5329. ACM, 2023. 4, 7
- [23] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421. Springer, 2020. 4
- [24] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhonggang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2
- [25] 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 *ICML*, pages 16784–16804. PMLR, 2022. 2
- [26] 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 *ECCV*, pages 352–371. Springer, 2022. 2, 5, 1
- [27] Chi-Han Peng and Jiayao Zhang. High-resolution depth estimation for 360deg panoramas through perspective and panoramic depth images registration. In *WACV*, pages 3116–3125, 2023. 4
- [28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Asbell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 5, 4
- [30] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360deg monocular depth estimation. In *CVPR*, pages 3762–3772, 2022. 4
- [31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 6, 7, 1, 4
- [32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 3
- [33] 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 *CVPR*, pages 22500–22510, 2023. 2
- [34] RunwayML. Stable diffusion. <https://github.com/runwayml/stable-diffusion>, 2021. 2, 3
- [35] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2
- [36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 2
- [37] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, pages 2226–2234, 2016. 5
- [38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 2
- [39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, pages 25278–25294, 2022. 1, 2
- [40] Ka Chun Shum, Hong-Wing Pang, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Conditional 360-degree image synthesis for immersive indoor scene decoration. In *ICCV*, pages 4478–4488, 2023. 3
- [41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2, 3
- [42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3, 1
- [43] Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Zhao Yang. Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture. In *ACM MM*, pages 6898–6906. ACM, 2023. 2, 4
- [44] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, pages 11895–11907, 2019. 2, 3
- [45] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *CVPR*, pages 1047–1056, 2019. 5, 8
- [46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 5
- [47] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdifussion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023. 1, 2, 3, 4, 5, 6, 7, 8
- [48] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhil Alsian, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, pages 16773–16783, 2023. 4
- [49] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022. 1
- [50] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Layoutmp3d: Layout annotation of matterport3d. *arXiv preprint arXiv:2003.13516*, 2020. 1
- [51] Guangcong Wang, Yinyo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *ECCV*, pages 477–492. Springer, 2022. 1, 2
- [52] Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. Perf: Panoramic neural radiance field from a single panorama. *arXiv preprint arXiv:2310.16831*, 2023. 2
- [53] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *WACV*, 2024. 2
- [54] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. In *ACM MM*, pages 6811–6821. ACM, 2023. 2, 3
- [55] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 2
- [56] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Ipo-ldm: Depth-aided 360-degree indoor rgb panorama outpainting via latent diffusion model. *arXiv preprint arXiv:2307.03177*, 2023. 2, 3
- [57] Jiale Xu, Jia Zheng, Yanyu Xu, Rui Tang, and Shenghua Gao. Layout-guided novel view synthesis from a single indoor panorama. In *CVPR*, pages 16438–16447, 2021. 5

- [58] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *JSTSP*, 14(1):5–26, 2020. [2](#)
- [59] Jian-Ru Xue, Jian-Wu Fang, and Pu Zhang. A survey of scene understanding by event reasoning in autonomous driving. *MIR*, 15(3):249–266, 2018. [1](#)
- [60] Bangbang Yang, Yinda Zhang, Yijin Li, Zhaopeng Cui, Sean Fanello, Hujun Bao, and Guofeng Zhang. Neural rendering in a room: amodal 3d understanding and free-viewpoint rendering for the closed scene composed of pre-captured objects. *ACM TOG*, 41(4):1–10, 2022. [1](#)
- [61] Bangbang Yang, Wenqi Dong, Lin Ma, Wenbo Hu, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. Dreamspace: Dreaming your room space with text-driven panoramic texture propagation. *arXiv preprint arXiv:2310.13119*, 2023. [1](#), [2](#)
- [62] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM CSUR*, 2022. [2](#)
- [63] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023. [2](#)
- [64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. [2](#), [5](#)
- [65] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *CVPR*, 2023. [2](#)
- [66] Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *AAAI*, pages 3653–3661, 2022. [3](#)
- [67] Chuhan Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods. *IJCV*, 129:1410–1431, 2021. [1](#)

Taming Stable Diffusion for Text to 360° Panorama Image Generation

Supplementary Material

The supplementary material is organized as follows. In Sec. A, we provide more details about the network architecture. In Sec. B, we provide more details about the experiment setup and baseline methods. In Sec. E, we provide more qualitative comparisons. In Sec. F, we provide more details about the layout conditioned generation. In Sec. G, we provide more generalization results to out-domain prompts.

A. Network Architecture

In Sec. 3 of the main paper, we introduced our dual-branch architecture for panorama generation. Here we provide more details about the position of inserting EPPA into the UNet of SD and the feature dimension of each layer in Tab. A.1. We found that inserting EPPA earlier than the first DownSampler (Table A.1(8)) and later than the last UpSampler (Table A.1(47)) is memory-consuming due to large feature maps with no better performance. Therefore, we insert EPPA right after the DownSampler and before the UpSampler of each block.

B. Experiment Details

As mentioned in Sec. 4 of the main paper, here we provide more details about the experiment setup and baseline methods.

Dataset.

Matterport3D dataset [3] is a large-scale scene understanding dataset with 10,800 panoramic images from 90 building-scale scenes. For text conditioned generation, we utilize BLIP-2 [18] to generate a short description of the full image with a prompt of “a 360 - degree view of”. We use the same data split as [47], which contains 9,820 for training and 1,092 for evaluation. We note that the original Matterport3D dataset contains blurry regions near the upper and lower edges, as shown in Fig. B.1a. Therefore, our model is trained to generate images with similar blurry regions. For text and layout conditioned generation, we use the MatterportLayout [50, 67] dataset, which annotates room layout for 2,295 indoor panoramic images in the Matterport3D dataset, with 1,648 for training, 191 for validation, and 459 for testing.

Implementation Details. We implement our model in PyTorch based on the implementation of Stable Diffusion [31] from Diffusers [49]. When training our dual-branch model for text-conditioned generation, we jointly train the EPPA module and finetune the two branches with rank-4 LoRA to the new resolution. We randomly sample 20 views as the input of the perspective branch to encourage EPPA to un-

derstand the correspondence provided by SPE and attention mask instead of remembering the fixed camera poses. Following MVDiffusion [47], we train for 10 epochs with the AdamW optimizer, using a batch size of 4 and learning rate of $2e-4$ for training, and a DDIM sampler [42] is used with a step size of 50 for inference. When training an additional ControlNet for text-layout conditioned generation, we extend training to 100 epochs due to less room layout annotations. Training is conducted on 4 NVIDIA A100 GPUs and takes about 8 hours for text conditioned generation and 15 hours for text-layout conditioned generation.

Perspective transformation details. In the main paper, we denote the transformation from equirectangular panorama $I^* \in \mathbb{R}^{C \times H \times W}$ to perspective image $I \in \mathbb{R}^{C \times h \times w}$ as $I = P(I^*, R, \text{FoV}, (h, w))$, where the rotation matrix $R \in SO(3)$ describes the camera extrinsic matrix and FoV and image size (h, w) define the camera intrinsic matrix K . Specifically, given a pixel $p \in \mathbb{R}^2$ on the image plane of I , we shoot a ray $K^{-1}[p, 1]^T$ from the camera center, and then transform it to the 3D coordinate of panorama as $v = R^{-1}K^{-1}[p, 1]^T$. Subsequently, its corresponding pixel p^* on the image plane of I^* can be computed as:

$$p^* = \left[\frac{W(\tan 2(v_y, v_x) + \pi)}{2\pi}, \frac{H(\tan 2(v_z, \sqrt{v_x^2 + v_y^2}) + \pi/2)}{\pi} \right],$$

which is used to bilinearly interpolate I from I^* . Note that we use different symbols here for easy explanation.

Evaluation Metrics. Previous works MVDiffusion [47] and Text2Light [4] both address the problem of text conditioned image generation, but in different domains. MVDiffusion generates 8 horizontal views with 90° FoV, thus limiting the evaluation to perspective images. Text2Light generates a full 180° vertical FoV, therefore focusing on evaluating the panorama quality. Ours is closer to the latter, but to showcase the effectiveness of our proposed method, we conduct a comparison in both. We also detail the implementation of layout consistency evaluation in the following.

- *In the panorama domain*, we value Fréchet Auto-Encoder Distance (FAED) [26] more, since it is customized for panorama and uses an auto-encoder trained on the target dataset as the feature extractor. Specifically, we train the auto-encoder similar to [26] but with RGB images instead of RGBD by removing the depth branch. The auto-encoder is trained on the training set of the Matterport3D dataset for 60 epochs with Adam optimizer and batch size of 4. An exponential learning rate scheduler is used with an initial learning rate of $1e-4$ and decay rate of 0.99 for every epoch.
- *In the perspective domain*, the CS is measured between the perspective image and the text prompt captioned from



Figure B.1. An example of a panoramic image from the Matterport3D dataset [3] (a) and its layout rendered as distance map (b). Regions near the upper and lower edges of the panoramic image are blurry in the original dataset.

GT view using BLIP-2 [18].

- *When evaluating layout consistency*, to make the comparison fair for MVDiffusion, we mask out pixels outside its vertical FoV before feeding the generated panorama to HorizonNet for our method, so that we do not benefit from the larger FoV when estimating the layout. We finetune HorizonNet on the masked training set of MatterportLayout dataset for 100 epochs with Adam optimizer and batch size of 4. The initial learning rate is set to $1e-4$ and halved if the validation loss does not decrease for 10 epochs.

Additional comparison with previous methods. We follow [47] to use the released weights of Text2Light [4] and MVDiffusion [47] as two of our baselines. For Text2Light, we use its first stage without super-resolution inverse tone mapping stage to get panoramic images at a resolution of 512×1024 , which takes 80.6 seconds per image on a single NVIDIA A100 GPU. For MVDiffusion, we use its direct outputs for quantitative comparison in main paper Tab. 1. This favors MVDiffusion by avoiding inconsistency in stitching and interpolating in projection. Therefore, to make a comprehensive comparison, we additionally evaluate MVDiffusion in different settings in Tab. B.1. We detail these settings in the following.

- *MVDiffusion* is in its original setting that does not involve stitching and projection, and is used for comparison in main paper Tab. 1. It outputs 8 horizontal views with 90° FoV at a resolution of 512×512 , which takes 102.2 seconds. One only difference from the original MVDiffusion paper is that we downsample the output images to 256×256 before evaluation to match the resolution of GT images.
- *MVDiffusion (projection)* uses the same weight as MVDiffusion, but we stitch its outputs into a panorama and then project the panorama back to perspective views for evaluation. This strictly follows our setting of panorama generation by considering the inconsistency between the perspective images. The performance drops significantly, which shows that inconsistency is a major issue for MVDiffusion.
- *MVDiffusion+LoRA* is MVDiffusion finetuned with

LoRA on a lower resolution at 256×256 . With lower resolution, the inference time is reduced to 27.4 seconds for a panorama with 90° vertical FoV at the resolution of 256×1024 , while our method takes 15.1 seconds to generate a panorama with 180° vertical FoV at the resolution of 512×1024 . This setting skips the stitching and projection thus does not reflect the actual panorama generation ability of MVDiffusion.

- *MVDiffusion+LoRA (projection)* follows the evaluation setting of MVDiffusion (projection) but uses the same weight as MVDiffusion+LoRA. The FID is better than MVDiffusion (projection), but still significantly worse than ours. This version is used for layout conditioned generation in Tab. 3 of the main paper, detailed in Sec. F.

While our method achieves better realism than baseline methods, it comes with a cost of higher computational complexity as discussed in Sec. 5 of the main paper. Specifically, the average inference time is 2.8 and 2.9 seconds per panorama for SD+LoRA and Pano Branch, respectively. However, we note that our model can be further optimized for higher speed as a significant amount of numpy operations are used for the EPPA module.

C. Loop Consistency Analysis

In Sec. 3.2, we describe two techniques to eliminate loop inconsistency, *i.e.*, latent rotation and circular padding. Qualitative results in Fig. C.1 show the stitched ends of generated panoramas with each column corresponding to one input text. We can see that latent rotation (b) can only mitigate loop inconsistency of SD+LoRA (a), while the results with circular padding combined (c) or alone (d) are more seamless.

D. Repetition analysis.

In Sec. 4.2, we qualitatively highlight the repetition issue of MVDiffusion. Here, we try to evaluate the repetition by projecting panorama to cubemap and computing a score

#	Layer	Output		Additional Inputs
		Pers Branch (20×)	Pano Branch	
(1)	Latent Map	4 × 32 × 32	4 × 64 × 128	
(2)	Conv.	320 × 32 × 32	320 × 64 × 128	
CrossAttnDownBlock1				
(3)	ResBlock	320 × 32 × 32	320 × 64 × 128	Time emb.
(4)	AttnBlock	320 × 32 × 32	320 × 64 × 128	Prompt emb.
(5)	ResBlock	320 × 32 × 32	320 × 64 × 128	Time emb.
(6)	AttnBlock	320 × 32 × 32	320 × 64 × 128	Prompt emb.
(7)	DownSampler	320 × 16 × 16	320 × 32 × 64	
(8)	EPPA	320 × 16 × 16	320 × 32 × 64	
CrossAttnDownBlock2				
(9)	ResBlock	640 × 16 × 16	640 × 32 × 64	Time emb.
(10)	AttnBlock	640 × 16 × 16	640 × 32 × 64	Prompt emb.
(11)	ResBlock	640 × 16 × 16	640 × 32 × 64	Time emb.
(12)	AttnBlock	640 × 16 × 16	640 × 32 × 64	Prompt emb.
(13)	DownSampler	640 × 8 × 8	640 × 16 × 32	
(14)	EPPA	640 × 8 × 8	640 × 16 × 32	
CrossAttnDownBlock3				
(15)	ResBlock	1280 × 8 × 8	1280 × 16 × 32	Time emb.
(16)	AttnBlock	1280 × 8 × 8	1280 × 16 × 32	Prompt emb.
(17)	ResBlock	1280 × 8 × 8	1280 × 16 × 32	Time emb.
(18)	AttnBlock	1280 × 8 × 8	1280 × 16 × 32	Prompt emb.
(19)	DownSampler	1280 × 4 × 4	1280 × 8 × 16	
(20)	EPPA	1280 × 4 × 4	1280 × 8 × 16	
DownBlock				
(21)	ResBlock	1280 × 4 × 4	1280 × 8 × 16	Time emb.
(22)	ResBlock	1280 × 4 × 4	1280 × 8 × 16	Time emb.
MidBlock				
(23)	ResBlock	1280 × 4 × 4	1280 × 8 × 16	Time emb.
(24)	AttnBlock	1280 × 4 × 4	1280 × 8 × 16	Prompt emb.
(25)	ResBlock	1280 × 4 × 4	1280 × 8 × 16	Time emb.
(26)	EPPA	1280 × 4 × 4	1280 × 8 × 16	
UpBlock				
(27)	ResBlock	1280 × 4 × 4	1280 × 8 × 16	(22), Time emb.
(28)	ResBlock	1280 × 4 × 4	1280 × 8 × 16	(21), Time emb.
(29)	ResBlock	1280 × 4 × 4	1280 × 8 × 16	(19), Time emb.
(30)	EPPA	1280 × 4 × 4	1280 × 8 × 16	
(31)	UpSampler	1280 × 8 × 8	1280 × 16 × 32	
CrossAttnUpBlock1				
(32)	ResBlock	1280 × 8 × 8	1280 × 16 × 32	(18), Time emb.
(33)	AttnBlock	1280 × 8 × 8	1280 × 16 × 32	Prompt emb.
(34)	ResBlock	1280 × 8 × 8	1280 × 16 × 32	(16), Time emb.
(35)	AttnBlock	1280 × 8 × 8	1280 × 16 × 32	Prompt emb.
(36)	ResBlock	1280 × 8 × 8	1280 × 16 × 32	(13), Time emb.
(37)	AttnBlock	1280 × 8 × 8	1280 × 16 × 32	Prompt emb.
(38)	EPPA	1280 × 8 × 8	1280 × 16 × 32	
(39)	UpSampler	1280 × 16 × 16	1280 × 32 × 64	
CrossAttnUpBlock2				
(40)	ResBlock	640 × 16 × 16	640 × 32 × 64	(12), Time emb.
(41)	AttnBlock	640 × 16 × 16	640 × 32 × 64	Prompt emb.
(42)	ResBlock	640 × 16 × 16	640 × 32 × 64	(10), Time emb.
(43)	AttnBlock	640 × 16 × 16	640 × 32 × 64	Prompt emb.
(44)	ResBlock	640 × 16 × 16	640 × 32 × 64	(7), Time emb.
(45)	AttnBlock	640 × 16 × 16	640 × 32 × 64	Prompt emb.
(46)	EPPA	640 × 16 × 16	640 × 32 × 64	
(47)	UpSampler	640 × 32 × 32	640 × 64 × 128	
CrossAttnUpBlock3				
(48)	ResBlock	320 × 32 × 32	320 × 64 × 128	(6), Time emb.
(49)	AttnBlock	320 × 32 × 32	320 × 64 × 128	Prompt emb.
(50)	ResBlock	320 × 32 × 32	320 × 64 × 128	(4), Time emb.
(51)	AttnBlock	320 × 32 × 32	320 × 64 × 128	Prompt emb.
(52)	ResBlock	320 × 32 × 32	320 × 64 × 128	(2), Time emb.
(53)	AttnBlock	320 × 32 × 32	320 × 64 × 128	Prompt emb.
(54)	GroupNorm	320 × 32 × 32	320 × 64 × 128	
(55)	SiLU	320 × 32 × 32	320 × 64 × 128	
(56)	Conv.	4 × 32 × 32	4 × 64 × 128	

Table A.1. Detailed PanFusion pipeline. We highlight the inserted EPPA modules in orange.

Method	Horizontal 8 Views [47]		
	FID ↓	IS ↑	CS ↑
MVDiffusion [47]	25.27	6.90	26.34
MVDiffusion (projection)	32.56	6.40	25.70
MVDiffusion+LoRA	21.76	6.55	25.22
MVDiffusion+LoRA (projection)	30.04	5.69	24.90
PanFusion (Ours)	19.88	6.50	24.98

Table B.1. More quantitative comparison. We compare our method with MVDiffusion [47] in different settings. MVDiffusion with projection considers stitching and projection, which is closer to our setting. We also finetune MVDiffusion with LoRA [14] on low resolution to have a fair comparison for time efficiency and layout-conditioned generation.



Figure C.1. Loop consistency analysis. We stitch both ends of each generated panorama. Here, each column corresponding to one same input text. It is shown that latent rotation (b) can only mitigate loop inconsistency of SD+LoRA (a), while the results with circular padding combined (c) or alone (d) are more seamless.

$RS(I_i, I_j) = \max(100 * \cos(E_i, E_j), 0)$ between each pair of 4 horizontal views, where E_* is the CLIP embedding of image I_* . RS is averaged over all image pairs of 1,092 test samples, with higher values indicating more repetition. It is shown in Tab. D.1 that our method has the lowest RS while MVDiffusion has the most repetition.

	Text2Light	MVDiffusion	PanFusion (Ours)	GT image
RS ↓	88.81	90.79	88.13	86.49

Table D.1. Repetition analysis. Inspired by CLIP Score [29], we report the repetition score (RS) that measures the similarity between different parts of the generated panorama images. Lower RS indicates less repetition.

Method	Layout Consistency		Horizontal 8 Views [47]		
	3D IoU ↑	2D IoU ↑	FID ↓	IS ↑	CS ↑
SD+LoRA [14, 31]	68.02	71.41	21.39	5.03	25.84
PanFusion (Ours)	68.46	71.82	22.58	5.10	26.04

Table F.1. Layout-conditioned comparison with SD+LoRA. Our method achieves comparable or better results.

E. More Qualitative Comparisons

In Sec. 4.2 Fig. 4 of the main paper, we compared our method with previous methods qualitatively. Due to space limitations, we cropped the generated images to the vertical FoV of MVDiffusion for all methods. Here we provide more qualitative comparisons without cropping in Figs. E.1 to E.10, where Fig. E.1 has the same prompts as Fig. 4 in the main paper. Similarly, we evenly sample 4 horizontal views from the generated panorama for each panorama, in which the first view crosses the left and right borders to show how loop consistency is handled.

F. Layout Conditioned Generation Details

In Sec. 3.4 of the main paper, we showcased the benefits of our dual-branch method with the application of layout conditioned generation. Specifically, the room layout is rendered as a distance map, as shown in Fig. B.1b, and normalized to the range of $[-1, 1]$ as an additional spatial condition. To add layout condition to MVDiffusion, we follow [43] to project the layout condition to perspective views as a distance map instead of a depth map to ensure consistency among overlapped regions. However, when training the ControlNet for MVDiffusion at the original resolution of 512×512 , it suffers from gradient explosion. Instead, we found finetuning MVDiffusion with LoRA on a lower resolution of 256×256 can make the training of the ControlNet converge, and also improve the realism of MVDiffusion. Therefore, we use MVDiffusion+LoRA as the base model for layout conditioned generation in main paper Tab. 3 to serve as a stronger baseline. In Figs. F.1 to F.2, we provide more quantitative comparison with MVDiffusion. We also compare with SD+LoRA in Tab. F.1 to show that our method can get comparable or better results.

G. Generalization to Out-domain Prompts

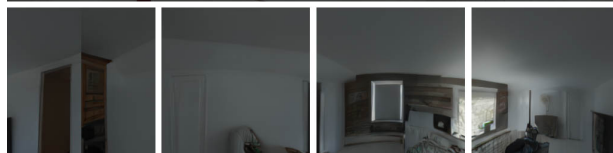
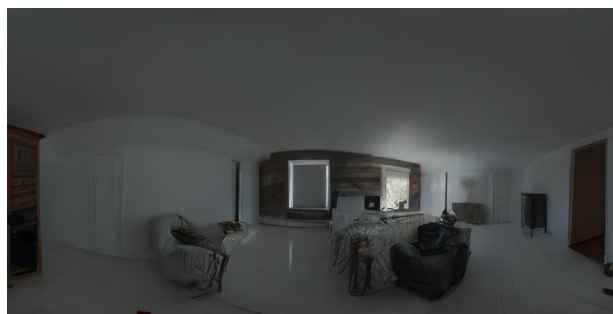
While our method is trained on the Matterport3D dataset, which contains mostly indoor scenes, we show that it can

generalize to out-domain prompts and transfer its knowledge of layout understanding to outdoor scenes, as shown in Figs. G.1 to G.4.

H. Future Works

Future works might include introducing more controls over the style and content of the generated panorama images to support applications like virtual house tour, or extending the method to enable outpainting by exploiting the perspective branch to extract guidance from the input image. The dual-branch architecture can also potentially benefit texture generation for 3D models, where the global branch can operate on UV maps and the perspective branch can operate on rendered images.

Text2Light



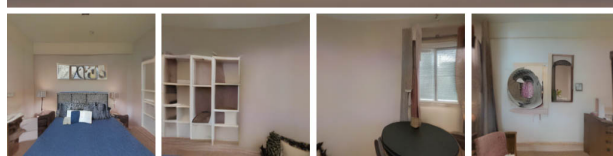
MVDiffusion



SD+LoRA



PanFusion (Ours)

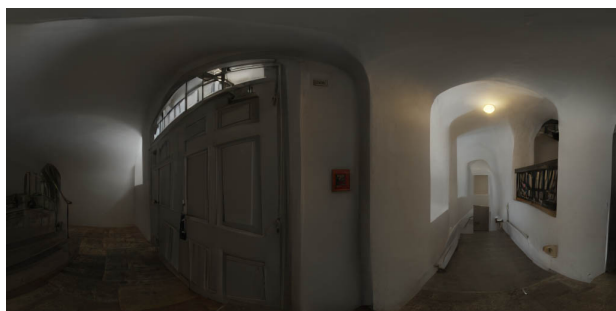


"A living room with a chandelier."

"A bedroom with a bed and a table."

Figure E.1. More qualitative comparisons.

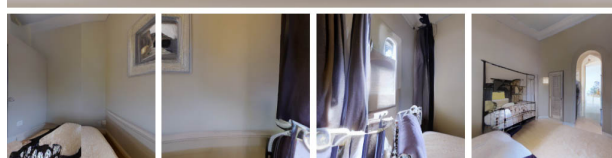
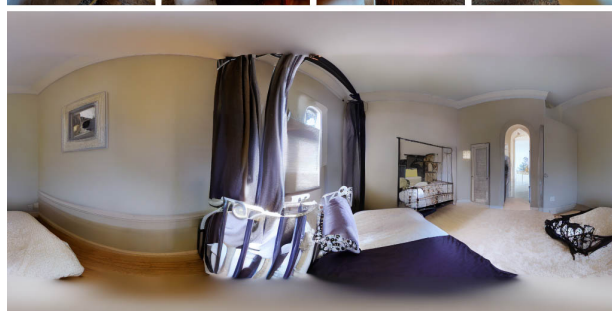
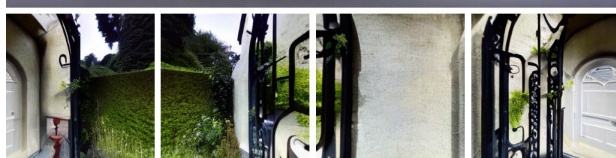
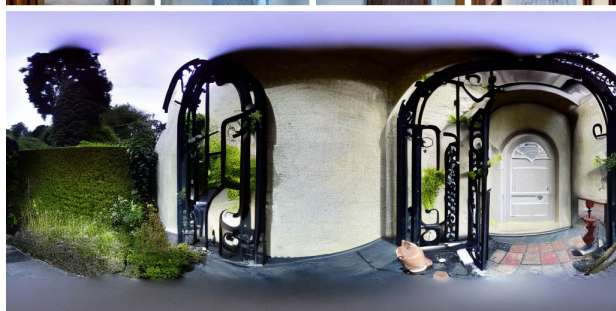
Text2Light



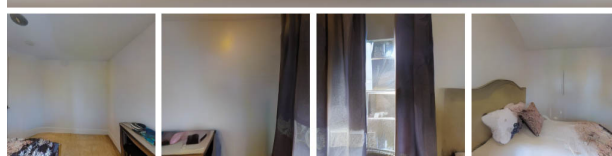
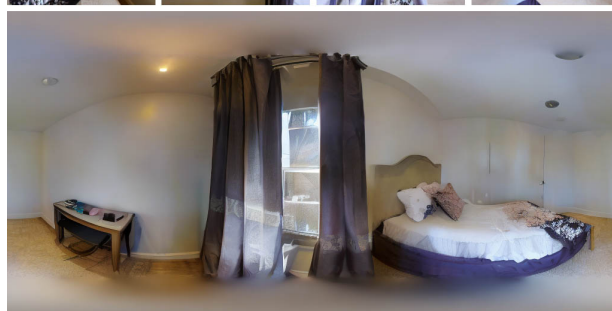
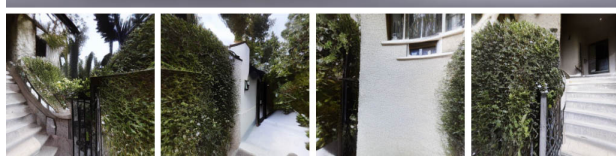
MVDiffusion



SD+LoRA



PanFusion (Ours)



"An entrance to a house."

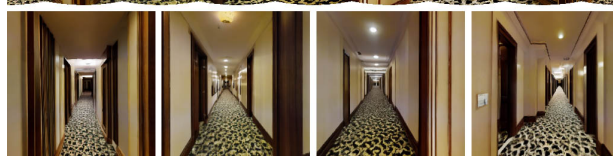
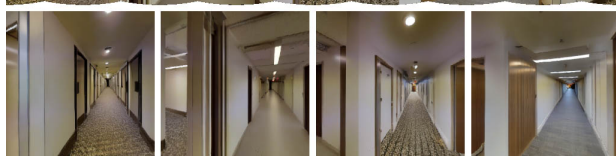
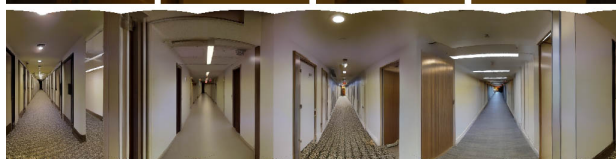
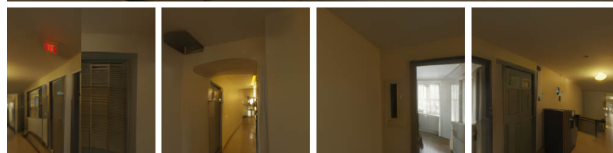
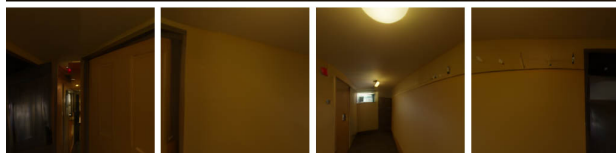
"A bedroom with a bed."

Figure E.2. More qualitative comparisons.

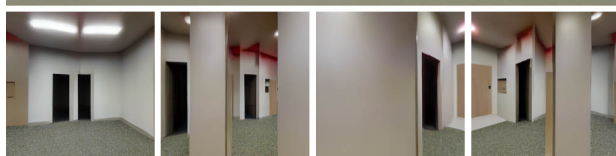
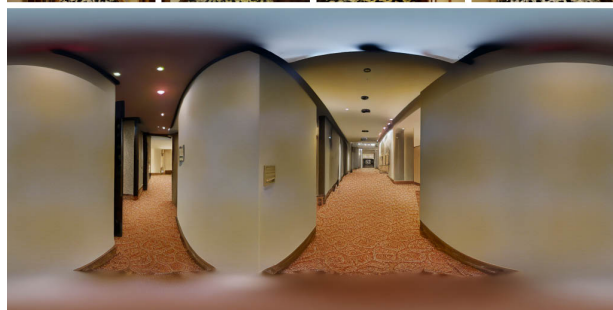
Text2Light



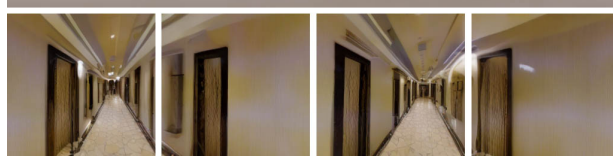
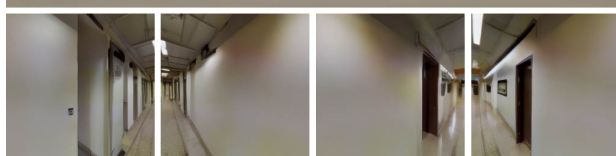
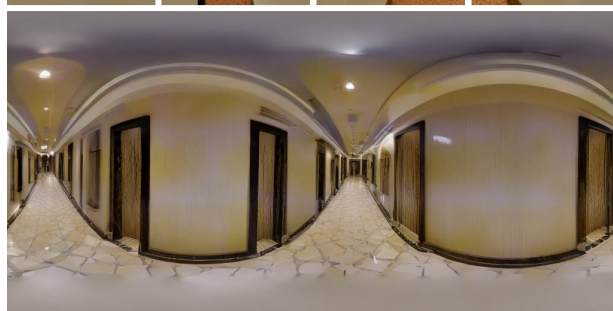
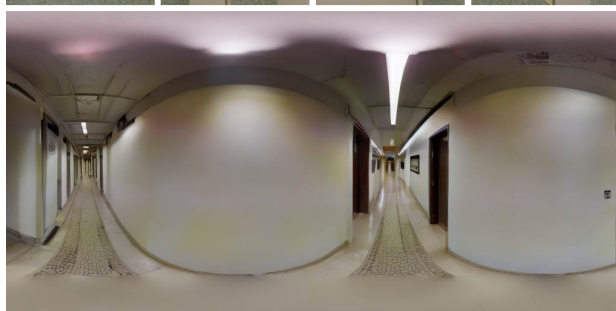
MVDiffusion



SD+LoRA



PanFusion (Ours)

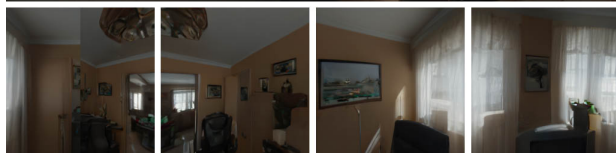
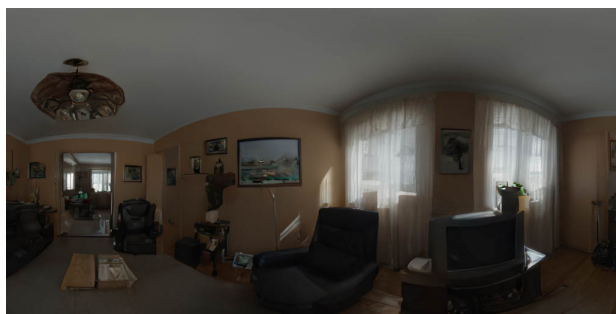


"A hallway in a building."

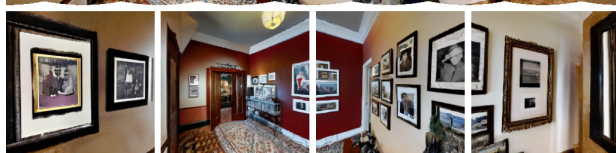
"A hallway in a hotel."

Figure E.3. More qualitative comparisons.

Text2Light



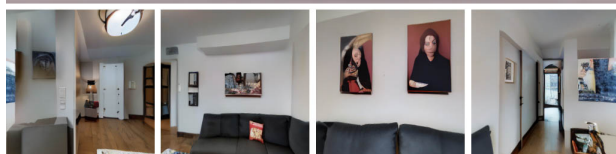
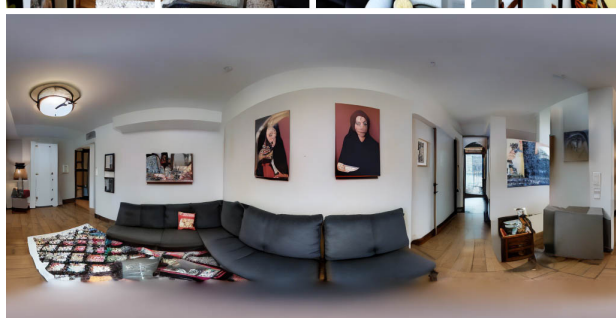
MVDiffusion



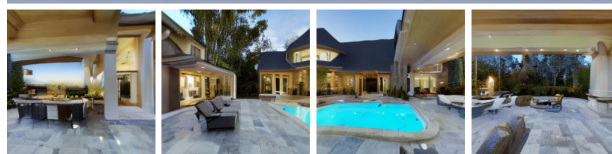
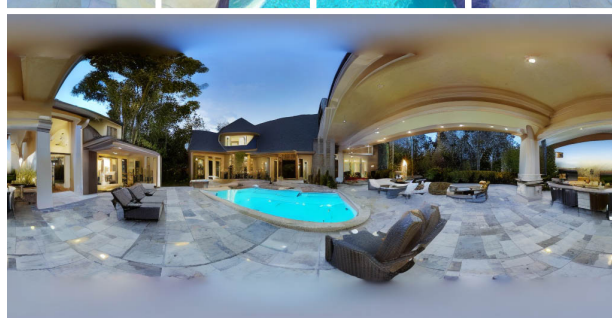
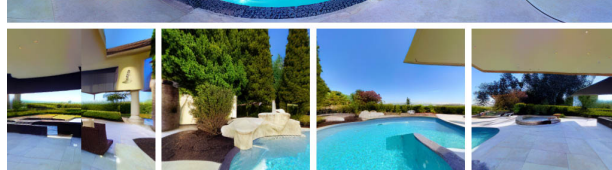
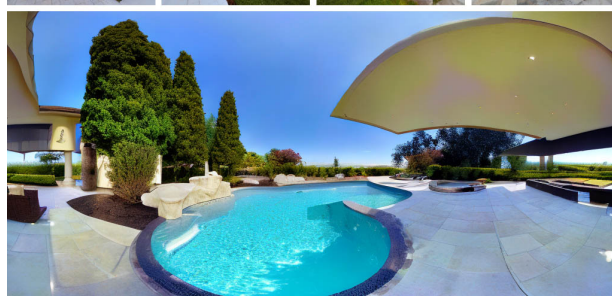
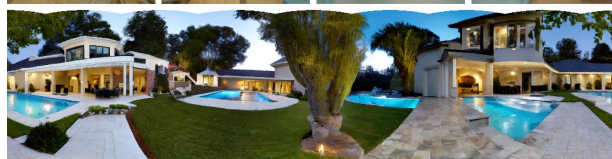
SD+LoRA



PanFusion (Ours)



"A living room with pictures on the wall."



"A home with a pool and patio."

Figure E.4. More qualitative comparisons.

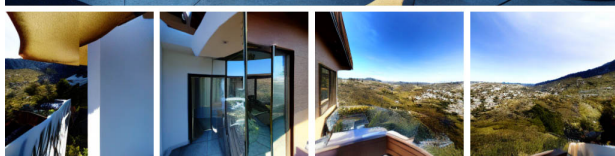
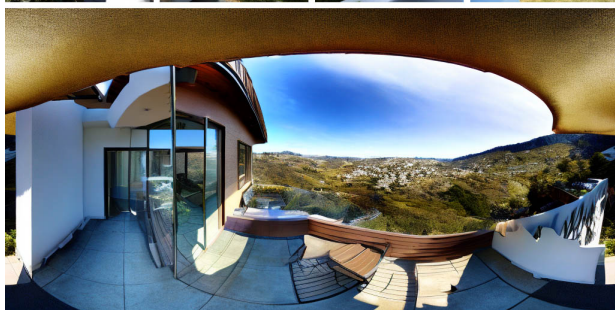
Text2Light



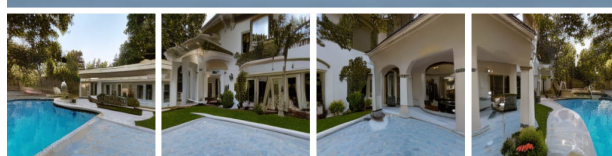
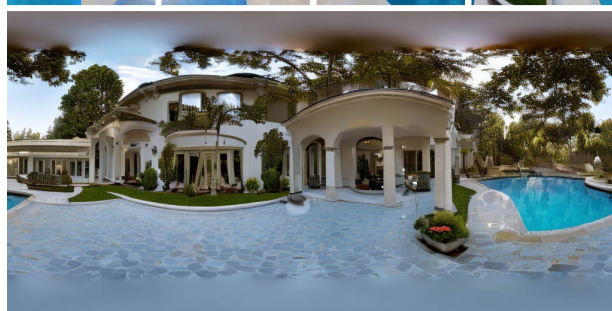
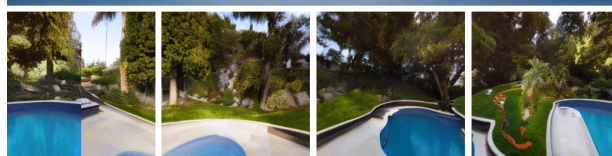
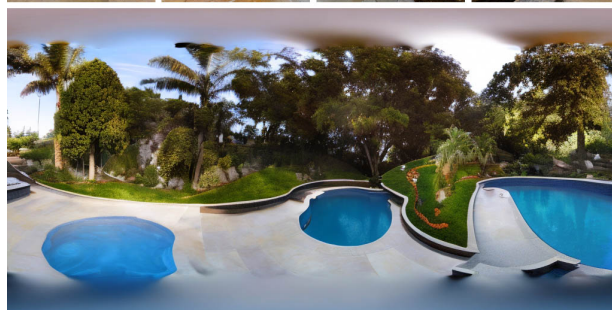
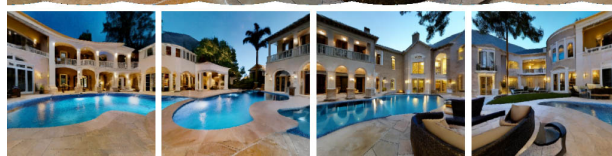
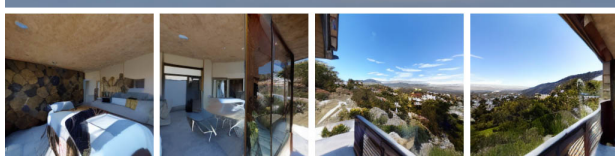
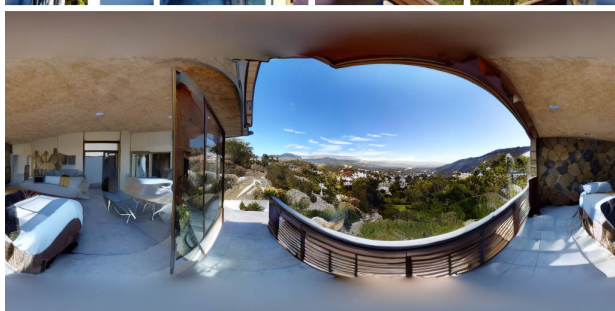
MVDiffusion



SD+LoRA



PanFusion (Ours)



“A house with a view of the mountains.”

“A large home with a pool.”

Figure E.5. More qualitative comparisons.

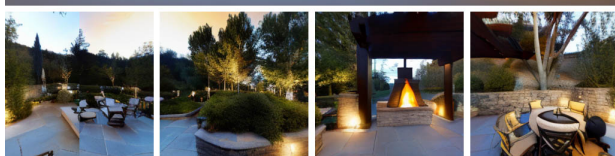
Text2Light



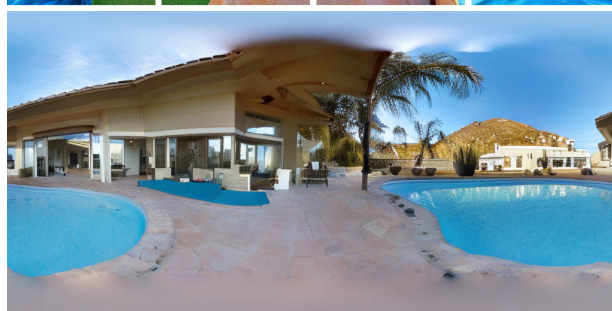
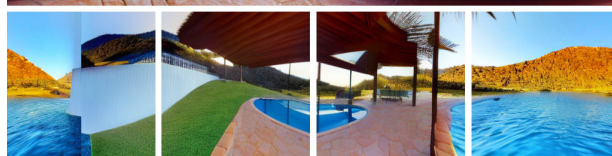
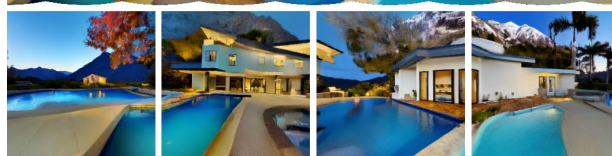
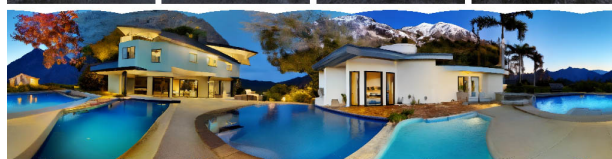
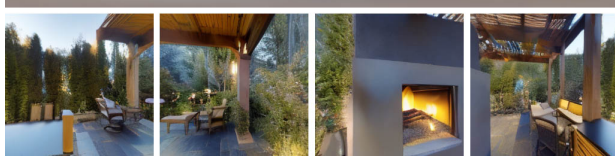
MVDiffusion



SD+LoRA



PanFusion (Ours)

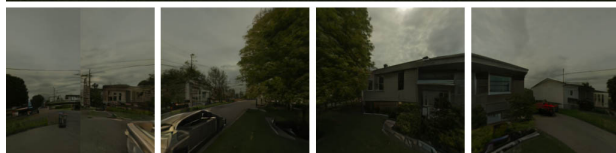


"An outdoor patio with a fireplace."

"A house with a pool and mountains in the background."

Figure E.6. More qualitative comparisons.

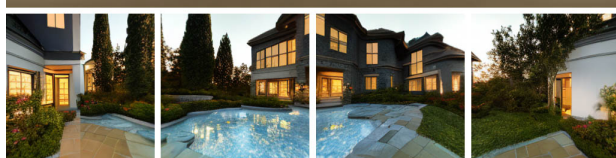
Text2Light



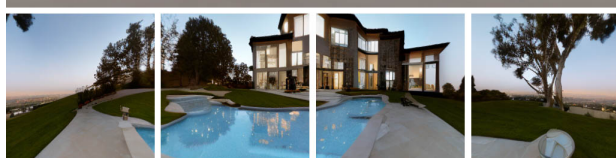
MVDiffusion



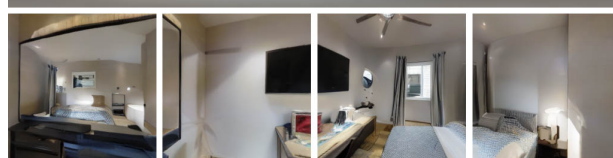
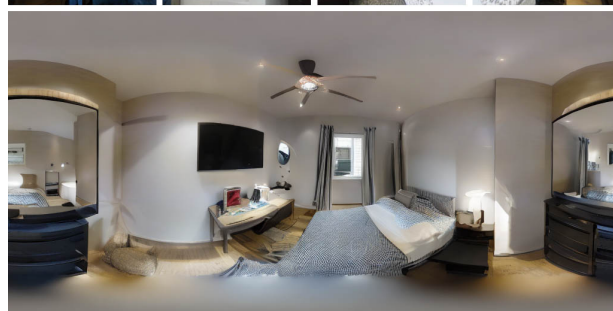
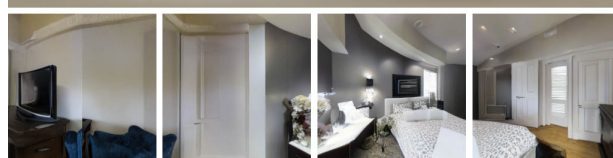
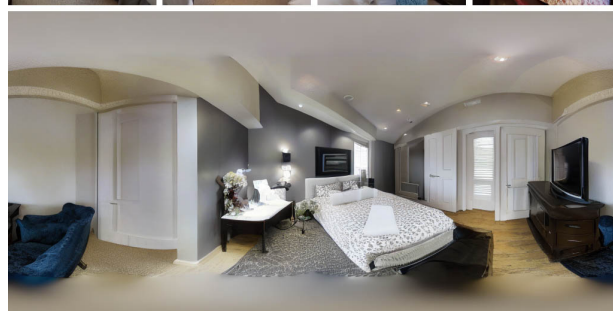
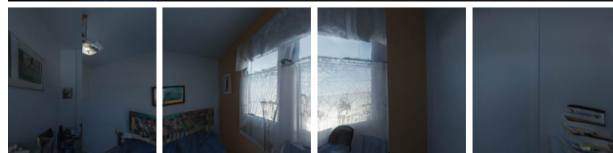
SD+LoRA



PanFusion (Ours)



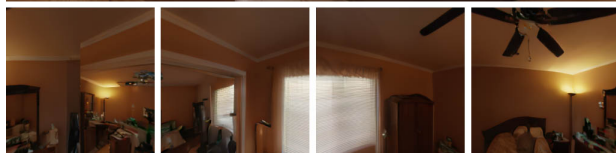
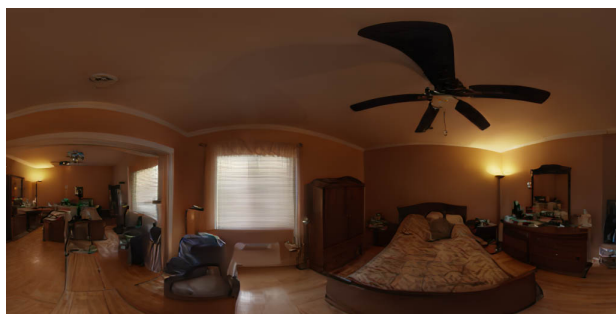
“A luxury home at dusk.”



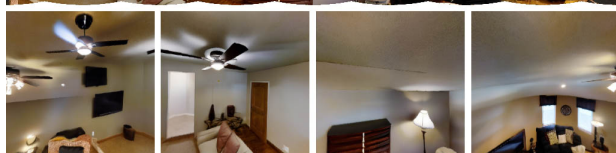
“A bedroom with a bed and TV.”

Figure E.7. More qualitative comparisons.

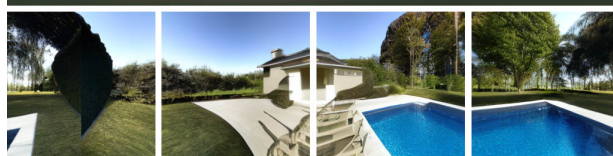
Text2Light



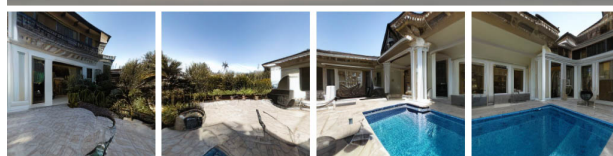
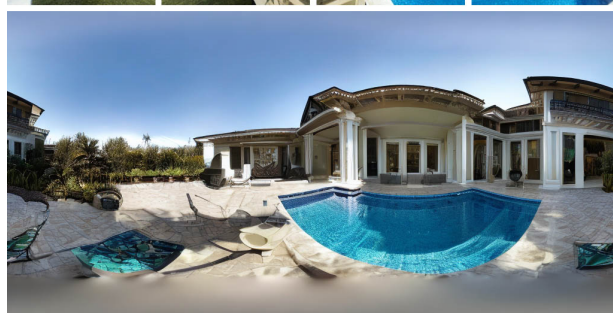
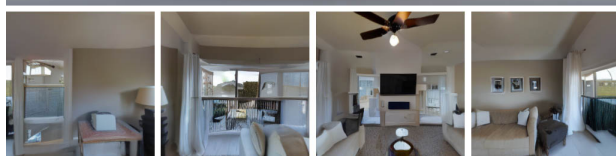
MVDiffusion



SD+LoRA



PanFusion (Ours)

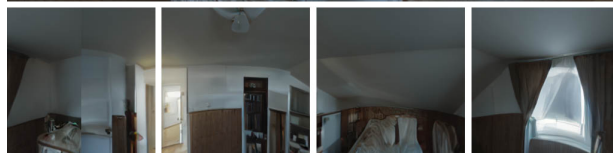
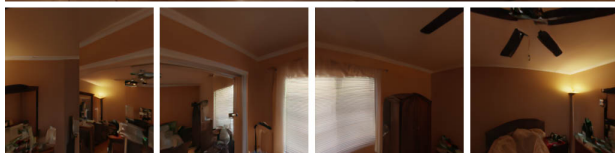
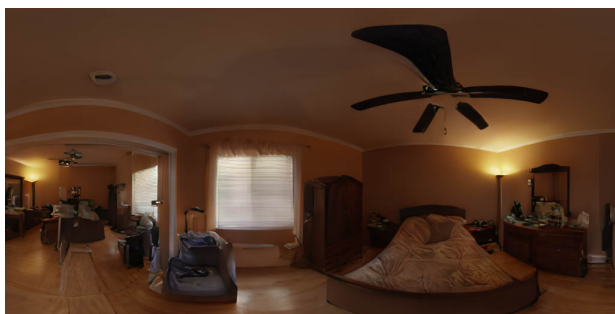


"A living room with a ceiling fan."

"A house with a pool."

Figure E.8. More qualitative comparisons.

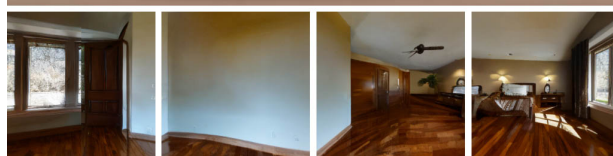
Text2Light



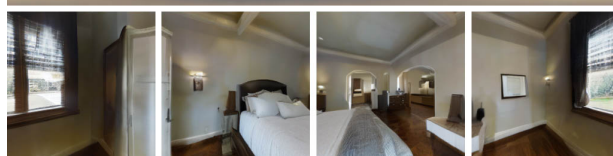
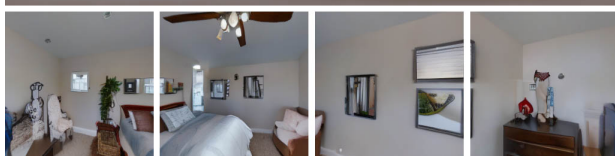
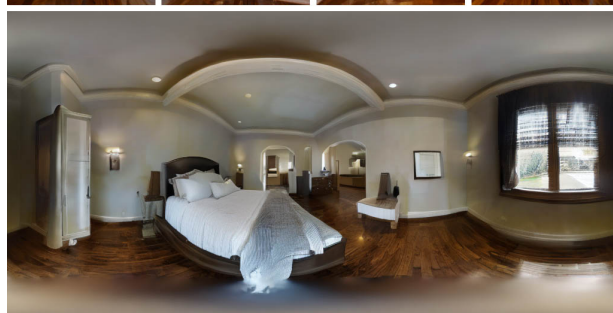
MVDiffusion



SD+LoRA



PanFusion (Ours)

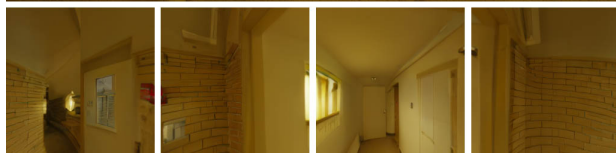
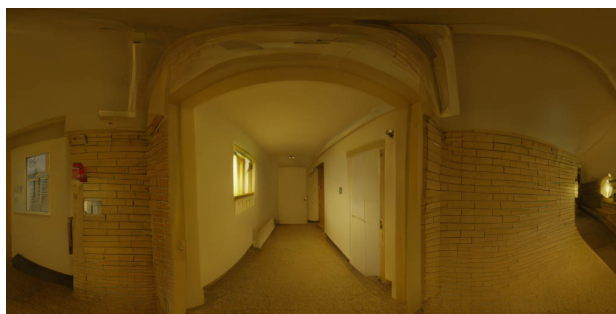


"A bedroom with a ceiling fan."

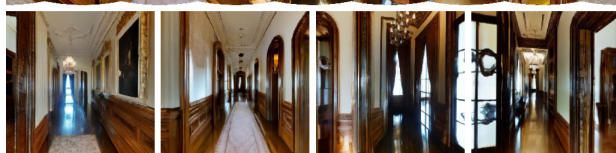
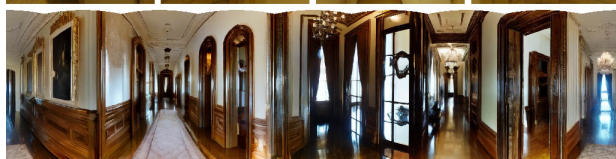
"A bedroom with hardwood floors."

Figure E.9. More qualitative comparisons.

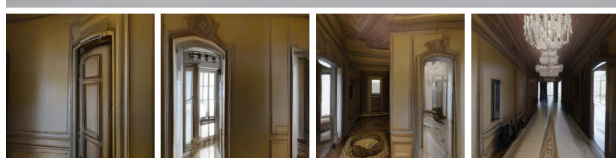
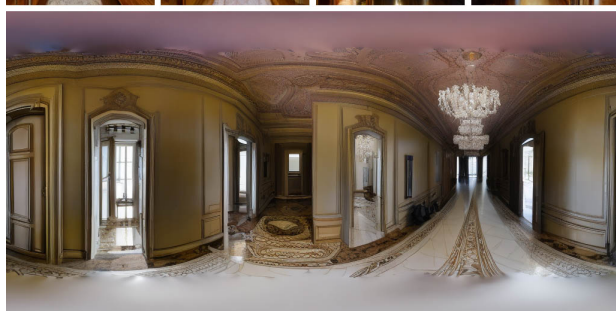
Text2Light



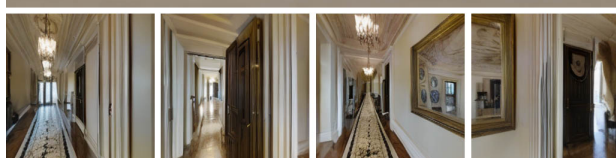
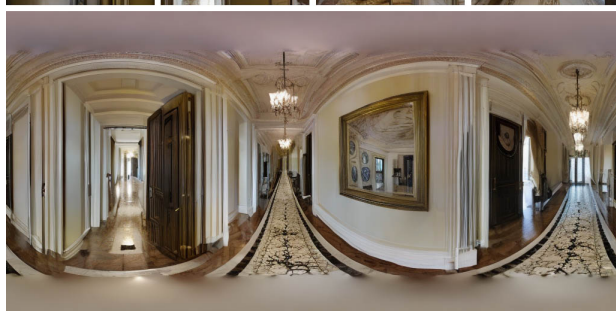
MVDiffusion



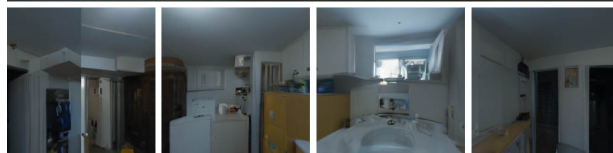
SD+LoRA



PanFusion (Ours)

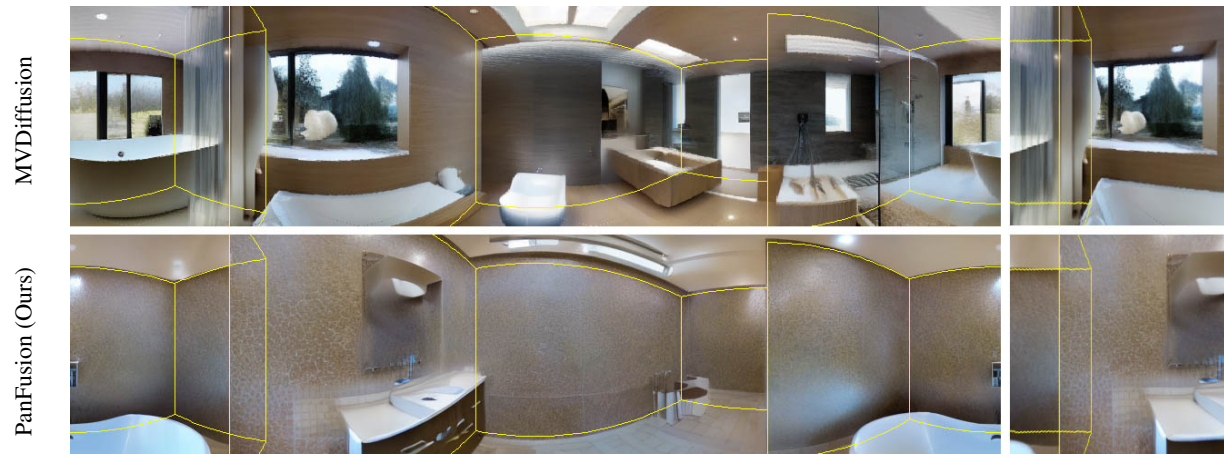


“A hallway in a mansion.”

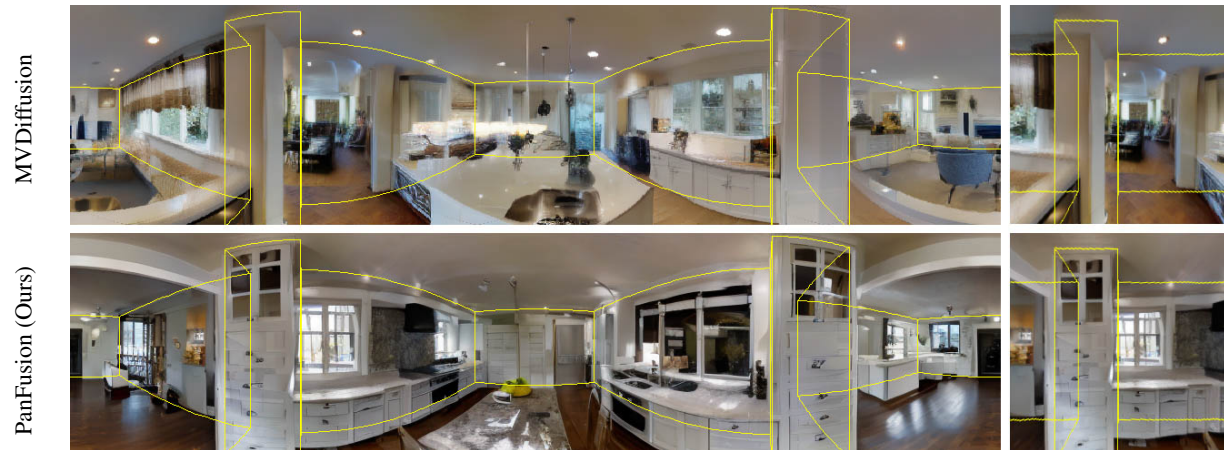


“The inside of a kitchen.”

Figure E.10. More qualitative comparisons.



“A bathroom with a tub and sink.”



“A kitchen and dining room.”



“A hallway in an office.”

Figure F.1. More layout-conditioned generation comparisons.



“An office with glass walls.”



“A kitchen and dining room.”

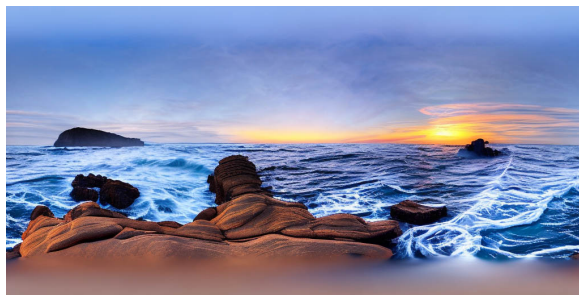


“A living room and dining room.”

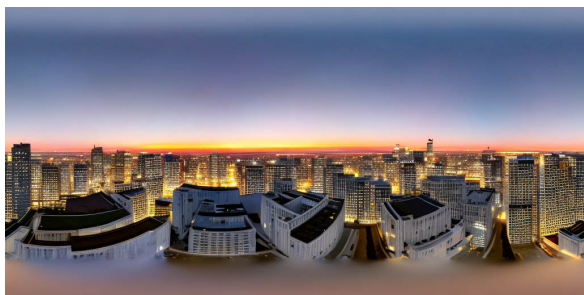
Figure F.2. More layout-conditioned generation comparisons.



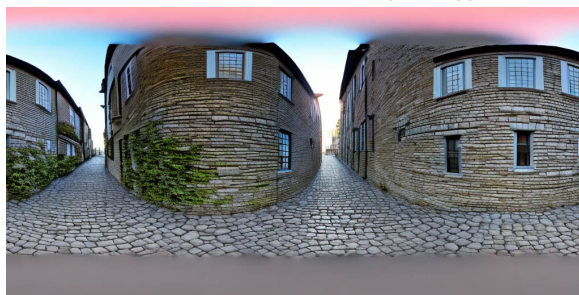
“A futuristic kitchen.”



“Coastal cliff at sunset, waves crashing on rugged rocks.”



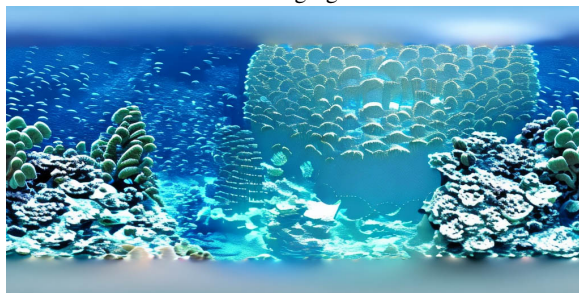
“Urban skyline at twilight, city lights twinkling in the distance”



“Cobblestone alley, historic architecture bathed in soft morning light.”



“Snow-covered cottage, smoke rising from a charming stone chimney.”



“An underwater scene, where coral reefs teem with colorful fish beneath the clear blue ocean.”



“A peaceful coastal village at sunrise, with fishing boats docked along the quiet harbor.”



“The interior of a historic library, filled with rows of antique books, leather-bound and dust-covered.”



“A tranquil botanical garden, with exotic plants, blooming flowers, and meandering stone pathways.”



“The calm waters of a secluded lake, reflecting the colors of the surrounding autumn foliage.”

Figure G.1. Generalization to out-domain prompts.



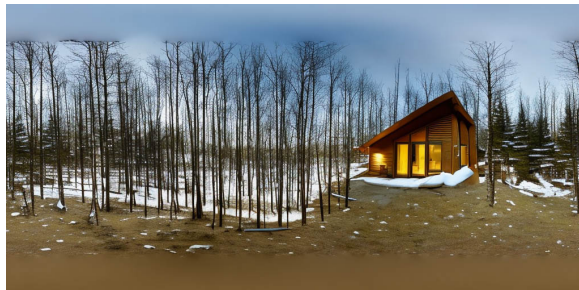
“Lighthouse in stormy seas.”



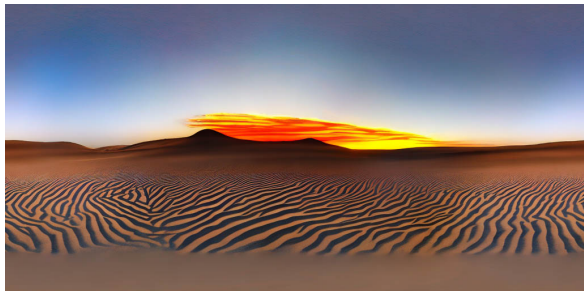
“Desert canyon, sculpted sandstone.”



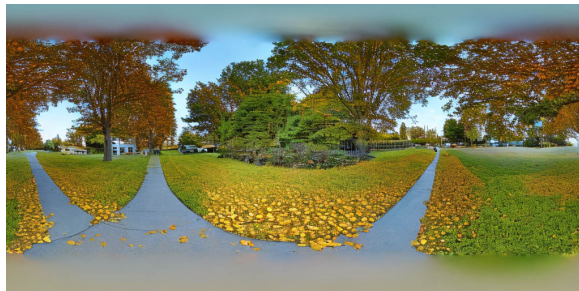
“Balcony garden, blooming serenity.”



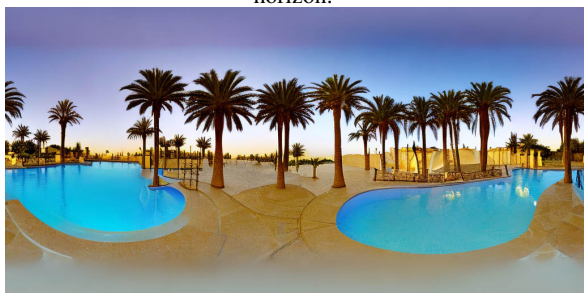
“Firelit cabin, crackling warmth amid snowy woods.”



“Desert sunrise, silhouettes painted against the golden horizon.”



“Suburban street, autumn leaves carpeting the sidewalk in hues.”



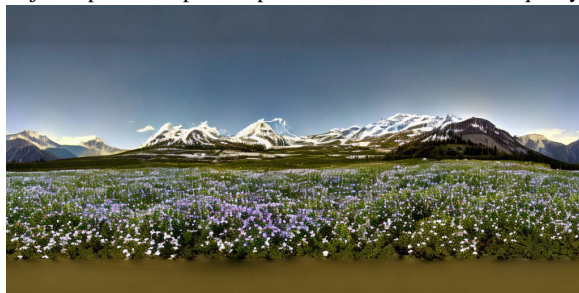
“Desert oasis, palm trees surrounding a pristine pool, an emerald jewel amid golden sands—an Arabian mirage.”



“Alpine village, snow-covered rooftops, nestled between majestic peaks—a picture-perfect scene of winter tranquility.”



“Rustic farmhouse, weathered by time, surrounded by fields of golden wheat—a pastoral scene capturing the essence of simplicity.”



“Alpine meadow, wildflowers swaying in a mountain breeze, snow-capped peaks embracing a serene panorama—a high-altitude sanctuary.”

Figure G.2. Generalization to out-domain prompts.



“A futuristic cityscape with floating skyscrapers and neon lights reflected in a calm river.”



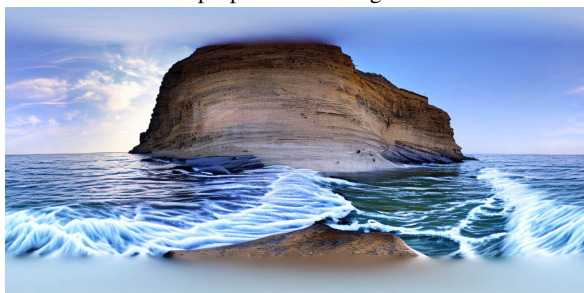
“In the heart of a bustling market, the aroma of exotic spices mingles with the vibrant colors of fresh produce.”



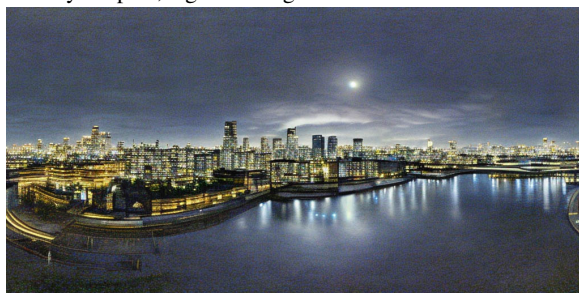
“Steampunk airship, navigating cloudy skies, gears turning, propellers whirring.”



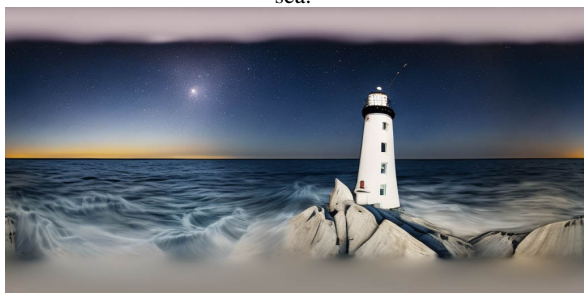
“Urban rooftop garden, vibrant blooms against a backdrop of skyscrapers, a green refuge amid concrete and steel.”



“Coastal cliffside, waves crashing on rugged rocks, seagulls soaring in the salty breeze—a dramatic meeting of land and sea.”



“Moonlit cityscape, reflections shimmering on rain-kissed streets, a quiet metropolis under the night sky—an urban nocturne.”



“Coastal lighthouse, guiding ships through the moonlit night.”



“Rooftop garden, city lights below, a quiet urban oasis.”

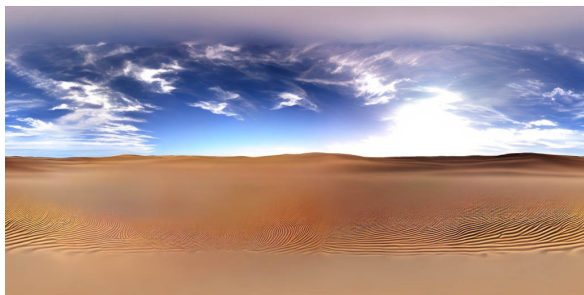


“Zen garden, raked pebbles, and bonsai trees—a serene oasis.”



“Tropical paradise, palm trees swaying, turquoise waters lapping sandy shores.”

Figure G.3. Generalization to out-domain prompts.



“Desert dunes, endless golden waves.”



“Antique bookstore, leather-bound treasures.”



“Alpine cabin, snow-capped serenity.”



“Rain-soaked city streets, glistening reflections.”



“Inside a bustling space station, people from different galaxies interact amid futuristic architecture and advanced robotics.”



“On the surface of a distant planet, a landscape of alien rock formations and swirling, multicolored gases.”



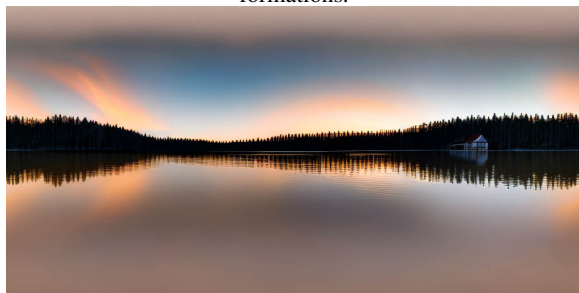
“A cozy coffee shop on a rainy day, with the comforting scent of freshly brewed coffee and the sound of rain on the windows.”



“Standing on the edge of the Grand Canyon, marveling at the vastness of the canyon and the layers of colorful rock formations.”



“A spaceship interior adorned with holographic displays, sleek metallic surfaces, and advanced technology.”



“A serene lakeside cabin at dawn, with mist rising from the water and the first light of the day illuminating the landscape.”

Figure G.4. Generalization to out-domain prompts.