HORIZON: High-Resolution Semantically Controlled Panorama Synthesis

Kun Yan[1], Lei Ji[2], Chenfei Wu[2], Jian Liang[3], Ming Zhou[4], Nan Duan[2], Shuai Ma[1]

[1]SKLSDE Lab, Beihang University,
[2]Microsoft Reseach Asia,
[3]Peking University,
[4]Langboat Technology
{kunyan,mashuai}@buaa.edu.cn, {leiji,chewu,nanduan}@microsoft.com, j.liang@stu.pku.edu.cn, zhouming@langboat.com

Abstract

Panorama synthesis endeavors to craft captivating 360-degree visual landscapes, immersing users in the heart of virtual worlds. Nevertheless, contemporary panoramic synthesis techniques grapple with the challenge of semantically guiding the content generation process. Although recent breakthroughs in visual synthesis have unlocked the potential for semantic control in 2D flat images, a direct application of these methods to panorama synthesis yields distorted content. In this study, we unveil an innovative framework for generating high-resolution panoramas, adeptly addressing the issues of spherical distortion and edge discontinuity through sophisticated spherical modeling. Our pioneering approach empowers users with semantic control, harnessing both image and text inputs, while concurrently streamlining the generation of high-resolution panoramas using parallel decoding. We rigorously evaluate our methodology on a diverse array of indoor and outdoor datasets, establishing its superiority over recent related work, in terms of both quantitative and qualitative performance metrics. Our research elevates the controllability, efficiency, and fidelity of panorama synthesis to new levels.

Introduction

Panoramic images and videos are becoming increasingly popular, due to the ability to provide an unlimited field of view (FOV) compared with traditional, planar images. With panoramic images, viewers can navigate 360° views and shift the viewing perspective in all directions, capturing a wealth of environmental detail. Additionally, these images provide an immersive experience that opens up a range of possibilities for interactive applications in a variety of domains, such as advertising, entertainment, and the design industry. However, the process of panorama acquisition typically requires significant human efforts or specialized panoramic equipment. Thus, the development of automated panoramic synthesis techniques is becoming increasingly important as virtual and augmented reality technology and devices, such as head-mounted displays and glasses, continue to evolve. This technique not only helps designers save time and effort when creating and editing blueprints, but it also reduces the cost associated with specialized panoramic equipment.

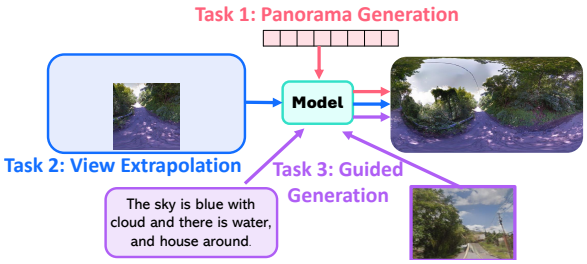


Figure 1: HORIZON supports multitype panorama synthesis

Panoramic images possess two unique characteristics: *spherical distortion* in distinct spatial locations and *continuity* between the left and right boundaries as compared to planar images. Current research on panoramic image synthesis primarily focuses on generating a spherical image with a large FOV from a single or a sequence of FOV images.(Sumantri and Park 2020) introduced the use of equirectangular projection for the generation of realistic spherical images from multiple images, addressing distorted projection issues commonly encountered when working with flat images. (Hara, Mukuta, and Harada 2021) proposed a method for generating spherical images without discontinuity. However, these methods lack the ability for *user control*, which is crucial in the synthesis of virtual worlds.

In particular, the ability to control generated content with style or semantic guidance is crucial for achieving desired images. As designers often invest a significant amount of time into creating and editing images with similar backgrounds but different semantics or styles. Research has been conducted in order to mimic human's capability to easily imagine 360-degree panoramic sceneries. This includes methods for image guidance for view extrapolation based on similar scene categories (Zhang et al. 2013), scene label guidance for controlling style using a co-modulated GAN (Karimi Dastjerdi et al. 2022), and text guidance for image synthesis (Chen, Wang, and Liu 2022). Both view extrapolation and panoramic image synthesis tasks can benefit from a variety of inputs to guide the generation process and make this technique more flexible and useful in real-world applications. As shown in Figure 1, the inputs can be text descrip-

tions and/or visual inputs for semantic and style guidance.

It is worth noting that recent text-to-image generation methods, such as VAE (Ramesh et al. 2021), GAN (Esser, Rombach, and Ommer 2021), and Diffusion (Dhariwal and Nichol 2021), have achieved great success in terms of semantic relevance and controllability for planar images. These works leverage the capabilities of attention mechanism (Vaswani et al. 2017) to take various types of guidance as generating conditions. However, these methods are difficult to apply directly to panoramic image synthesis due to a lack of consideration for spherical characteristics. To the best of our knowledge, there are currently no existing frameworks that are general enough to handle all of these spherical properties and controlling guidance for panorama synthesis in a unified framework. This is the motivation behind our work, which aims to explore advanced guided image generation techniques and design specific modules, mechanisms, and training strategies for spherical structures.

Additionally, it is imperative to generate *high-resolution* panoramic images in order to enhance the immersive experience for Virtual Reality and Augmented Reality. However, simply adopting current state-of-the-art methods such as DALL·E-like (Esser, Rombach, and Ommer 2021) for high-resolution panorama generation can result in spherical distortion and low efficiency. Recent efforts have attempted to address this issue by using separate models to first generate low-resolution panoramas and then upscaling to high-resolution, such as (Akimoto, Matsuo, and Aoki 2022a; Chen, Wang, and Liu 2022). These methods, however, can still result in artifacts caused by error accumulation. It is worth noting that we find properly managing high-resolution context within a defined module can alleviate this problem and produce better results.

In this paper, we introduce a novel, versatile framework for generating high-resolution panoramic images that have well-preserved spherical structure and easy-to-use semantic controllability. Specifically, we employ a two-stage procedure that includes learning an image encoder and decoder in the first step and a reconstruction model in the second step. In the second stage, we propose a new method called Spherical Parallel Modeling(SPM) that not only improves efficiency through parallel decoding, but also addresses the issue of local distortion through the use of spherical relative embedding and spherical conditioning improvement. Additionally, we have found that the panoramic pictures generated by SPM no longer have the problem of screen tearing when the left and right edges are spliced, which means that the generated panoramic pictures can be directly viewed in VR devices without the need for further editing.

The contributions can be summarized as:

- Alleviating the spherical distortion and edge incontinuity problem through spherical modeling.
- Supporting semantic control through both image and text guidance.
- Effectively generating high-resolution panoramas through parallel decoding.

Related Work

Panorama Synthesis Panorama synthesis, a well-established task in computer vision, involves various input types such as overlapped image sequences (Szeliski 2006; Brown and Lowe 2007), sparse images (Sumantri and Park 2020), and single images (Akimoto et al. 2019; Hara, Mukuta, and Harada 2021). Traditional methods employed image matching and stitching (Szeliski 2006), while recent generative models utilize GAN-based methods (Akimoto et al. 2019; Koh et al. 2022) and autoregressive models (Rockwell, Fouhey, and Johnson 2021) for panorama generation.

In computer graphics, view synthesis techniques, including geometry and layout prediction, optical flow, depth, and illumination estimation (Song and Funkhouser 2019; Xu et al. 2021; Zhang, Wang, and Liu 2022; Wang et al. 2022; Somanath and Kurz 2021), are often studied. Spherical structure and texture are modeled using cube maps (Han and Suh 2020), cylinder convolution (Liao et al. 2022), predicted panoramic three-dimensional structures (Song et al. 2018), and scene symmetry (Hara, Mukuta, and Harada 2021).Most previous generative models are limited in handling fixed scenes and low-resolution images. Recent research (Sumantri and Park 2020; Akimoto, Matsuo, and Aoki 2022a) addresses these limitations using hierarchical synthesis networks (Sumantri and Park 2020), U-Net structures (Akimoto, Matsuo, and Aoki 2022a), and separate models for generating and upscaling low-resolution images (Chen, Wang, and Liu 2022). Our proposed method tackles high-resolution panorama context and preserves spherical structure within a single module.

User-controlled semantic content generation is essential for interactive panorama generation. Recent works adopt scene symmetry with CVAE-based methods (Hara, Mukuta, and Harada 2021) or scene category with GAN-based methods (Karimi Dastjerdi et al. 2022) for view extrapolation. Our versatile framework addresses spherical, user-controlling, and high-resolution panoramas in a unified manner, incorporating mechanisms to handle spherical distortion, continuity, and semantic guidance without additional tuning.

We put a straightforward comparison of these most recent relevant efforts in Table 1, and detailed discussion can be found in the appendix.

Image Generation Previous image generation works primarily employ generative adversarial networks (GAN) (Goodfellow et al. 2014; Reed et al. 2016; Xu et al. 2018; Qiao et al. 2019; Zhang et al. 2021a), VAE-based methods (Kingma and Welling 2013; van den Oord, Vinyals, and kavukcuoglu 2017), and denoising diffusion models (Nichol et al. 2021; Gu et al. 2021; Kim and Ye 2021). With the advent of transformer models, two-stage methods have emerged as a new paradigm for pretraining with web-scale image and text pairs, demonstrating effectiveness in generalizing high semantically related open-domain images (Ramesh et al. 2021; Ding et al. 2021; Zhang et al. 2021b). These methods tokenize images into discrete tokens using VQVAE (van den Oord, Vinyals, and kavukcuoglu 2017) or VAGAN (Esser, Rombach, and

Method	High Resolution?	Spherical Coherence?	Global Semantic Condition?	Multiscale Semantic Editing?	Inference Efficiency?
COCO-GAN (2019)	✗	◗	◗	✗	🚀
InfinityGAN (2022)	✓	✗	✓	✓	🚀
LDM (2021)	✓	✗	✓	✗	🚹
Text2light (2022)	✓	◗	◗	✗	🚹
OminiDreamer (2022b)	✗	◗	✗	✗	🚗
Horizon	✓	✓	✓	✓	🚀

Table 1: The comparison between our method and the most recent relevant methods. ◗ represents partially satisfying the property. Find detailed evaluations and explanations in the appendix.

Ommer 2021), then generate visual tokens for decoding into real images.

Efforts have been made to improve high-resolution planar image generation (Esser, Rombach, and Ommer 2021; Chang et al. 2022; Wu et al. 2022). However, these models often produce blurred or teared artifacts, making them unsuitable for panoramic scenarios. Guided image generation methods achieve superior performance (Ramesh et al. 2021; Ding et al. 2021; Zhang et al. 2021b), but applying existing models to panoramic generation without considering the unique spherical structure remains challenging.

Method

The overall training is a two-stage procedure, similar to (Ramesh et al. 2021; Ding et al. 2021). The first stage is to train an encoder for image/view representation (discrete visual tokens in this paper) and a decoder for image generation, both of which are frozen in the second stage. The second stage is to learn a reconstruction model based on the discrete visual representation.

- **Stage 1.** Every equirectangular projected panoramic image with a resolution of 768x1,536 is first divided into 3x6=18 RGB view patches, each with a resolution of 256x256. Then we train a VQGAN(Esser, Rombach, and Ommer 2021) on every view patch separately. The encoder of VQGAN compresses each RGB view patch into a 16×16 grid of view tokens. The overall view token dictionary has a size of 16,384 possible values. As a result, each panoramic image has 18×16×16 view tokens.
- **Stage 2.** All 18 groups of view tokens from a single panoramic image are modeled as a whole context to incrementally learn the reconstruction of all view tokens. We progressively develop auto-regressive modeling, local parallel modeling, and the newly proposed spherical parallel modeling detailed described in the following subsections.

We apply the off-the-shelf model in the first stage and mainly devote our effort to effectively modeling the prior in the second stage. Due to the large number of view tokens for a single panorama (18x16x16=4608), it is still a non-trivial problem to learn the reconstruction. We should balance the quality, efficiency, and controllability with considerable refinement. In the following subsections, we will describe our progressive attempts and corresponding design considerations.

Auto-Regressive Modeling

One intuitive way is to directly employ a auto-regressive transformer decoder to generate 4608 view tokens one by one. However, due to the quadraticity of the attention mechanism of the transformer itself, directly inputting 4608 tokens into the model will bring huge memory consumption and great difficulties to the training of the model. At the same time, the sequence is too long for the model to converge efficiently.

We noticed that, as the relative distance increases, the impact of adjacent tokens becomes weak or even negative for the quality. Therefore, we shrink the attention scope for both efficiency and effectiveness consideration. Specifically, the range of attention for each view patch is limited to 2 surrounding view patches on the left and above, and autoregressively performs the prediction within the current patch. The interval of attention is shown at the top of Figure 2. After making this improvement, we have been able to generate decent high-resolution panoramas, which are also used as our first baseline ARM.

Local Parallel Modeling

Although ARM makes high-resolution panorama generation basically feasible, flattening the view patch into a one-dimensional sequence of tokens in raster scan order is still not an optimal and efficient modeling solution. Since the length of the autoregressive sequence still grows quadratically, it not only presents a challenge for modeling long-term correlations but also makes decoding intractable. Inspired by MaskGIT(Chang et al. 2022), we adopt the Masked Visual Token Modeling(MVTM) into the view modeling process, which can be formulated as:

$$\mathcal{L}_{\text{LPM}} = -\mathbb{E}\left[\sum_{\forall i \in [1,N], mask_i=1} \log p\left(y_i \mid Y_{\overline{\mathbf{M}}}; Y_W\right)\right] \quad (1)$$

For every training pass, sample a subset of tokens and replace them with a special [MASK] token. The number of masked tokens is parameterized by a scheduling function $\lceil \cos(r * \pi/2) \cdot N \rceil$, r is a real number uniformly sampled from 0 to 1, N is the total number of view tokens in current view patch. Masked token sequence $Y_{\overline{\mathbf{M}}}$ and ground truth view

tokens from surrounding view patch Y_W are fed into a multi-layer bidirectional transformer to predict the probabilities $p\left(y_i \mid Y_{\overline{\mathbf{M}}}; Y_W\right)$ for each masked token, where the negative log-likelihood is computed as the cross-entropy between the ground-truth and prediction.

During inference, we use a similar iterative decoding technique with a constant step T, initially all tokens in the current patch are masked, at each step t only $\lceil(1-(\cos\left(\frac{\pi t}{2T}\right)))N\rceil$ tokens with higher confidence are kept, others will be refined during further steps. We named this adapted version of MaskGIT as Local Parallel Modeling (LPM), that is, the modeling is applied in parallel for each local view patch. By applying this strategy, we shorten the inference speed by 64 times. However, we also observe a significant performance drop compared with ARM.

Spherical Parallel Modeling

Both ARM and LPM modules regard each patch view equally in the image, which do not take the spherical characteristics of panoramic images into consideration in model design. They assume that the visual features after spherical projection are translation-invariant on the two-dimensional plane. This is obviously not in line with the actual situation. We observed that under the ARM method, the model can still maintain a strong relative positional relationship, and according to the current sequence order, it can be deduced what degree of deformation should be used to generate the current view token. However, under the LPM method, the generation of the current token is no longer strictly constrained by the previous token, and sequence order is no longer an important factor for model learning goals. Naturally, the panoramic images generated by LPM have distorted local details, which are relatively weak in performance indicators such as FID.

In this section, we describe a new Spherical Parallel Modeling(**SPM**) method that not only maintains the efficiency of parallel decoding but also alleviates the local distortion problem through the spherical relative embedding and spherical conditioning improvement. Besides, we found that the panoramic images generated by SPM no longer have the problem of screen tearing when the left and right edges are spliced.

Spherical Relative Embedding Relative positional embedding effectively captures positional information during attention, particularly for spherical properties. Formally, a positional encoding function $f(\mathbf{x}, l)$ is defined for item x at position l. For items $\mathbf{q}$ and $\mathbf{k}$ at positions m and n, the inner product between $f(\mathbf{q}, m)$ and $f(\mathbf{k}, n)$ depends on $\mathbf{q}$, $\mathbf{k}$, and their relative position $m-n$. The dot product between two vectors is a function of their magnitudes and the angle between them.

Rotary Position Embedding (RoPE) (Su et al. 2021) encodes text token embedding with an absolute position using a rotation matrix, incorporating explicit relative position dependency in self-attention. Embeddings are treated as complex numbers and positions as pure rotations. During attention, if both query and key are shifted by the same amount, changing the absolute but not relative position, both representations are rotated similarly, maintaining the angle and dot product between them.

The function solution that satisfies the above requirement can be formulated as below:

$$
\begin{aligned}
f(\mathbf{q}, m) &= \begin{pmatrix} M_1 & & & \\ & M_2 & & \\ & & \ddots & \\ & & & M_{d/2} \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_d \end{pmatrix} = \mathbf{R_m Q_m} \\
&= \mathbf{R_m W_q X_m}
\end{aligned} \tag{2}
$$

where $M_j = \begin{pmatrix} \cos m\theta_j & -\sin m\theta_j \\ sinm\theta_j & \cos m\theta_j \end{pmatrix}$, $\mathbf{R_m}$ is the block diagonal rotation matrix, $\mathbf{W_q}$ is the learned query weights, and $\mathbf{X_m}$ is the embedding of the m-th token. For query $\mathbf{k}$, a similar corresponding equation is applied. When extending to the 2-dimensional case, the rotation matrix should correlate with both coordinates x and y:

$$
\boldsymbol{M}_{x,y} = \left(\begin{array}{cc:cc} \cos x\theta & -\sin x\theta & 0 & 0 \\ \sin x\theta & \cos x\theta & 0 & 0 \\ \hdashline 0 & 0 & \cos y\theta & -\sin y\theta \\ 0 & 0 & \sin y\theta & \cos y\theta \end{array}\right) \tag{3}
$$

However, only two-dimensional relative position embedding cannot represent the relative positional relationship of the spherical surface. This is manifested in two aspects. One is that the distance between the plane and the spherical surface is measured in different ways. Second, the coordinates of the same latitude have a ring-shaped positional relationship, that is, for a token sequence $0 \ldots m$ at the same latitude, the positional embedding of token 0 and the positional embedding of token m should be as close as possible. To satisfy this property, we re-derived the rotation matrix, instead of the Θ in the original RoPE:

$$
\Theta = \left\{\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, \ldots, d/2]\right\} \tag{4}
$$

We define Θ_{sphere} as:

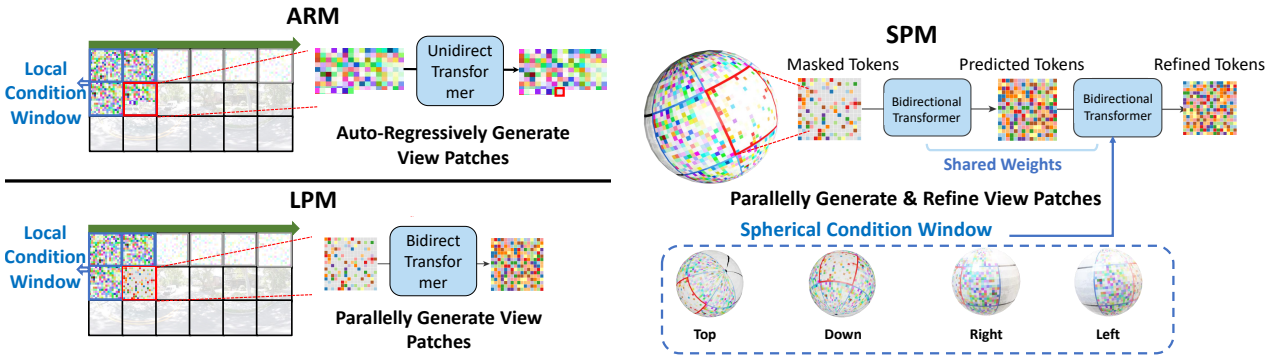
$$
\Theta_{shpere,x} = \left\{\theta_i = \frac{-2(i-1)*2\pi}{d*w}, i \in [1, 2, \ldots, d/2]\right\}, \tag{5}
$$

$$
\Theta_{shpere,y} = \left\{\theta_i = \frac{-2(i-1)*\pi}{d*h}, i \in [1, 2, \ldots, d/2]\right\}, \tag{6}
$$

where x, y are the different axis of the spherical surface, w is the length of token sequences along the x axis, and h is the length along the y axis. Note that as $\Theta_{shpere,x}$ represents latitude, the numerator has a factor 2π, which makes the rotation of the sequence head and tail as close as possible. While $\Theta_{shpere,y}$ does not keep this property, as the poles of a sphere are far from each other naturally.

Spherical Relative Embedding(SRE) applies to self-attention as follows:

$$
\begin{aligned}
SRE(\boldsymbol{q}_{(x1,y1)}^{\top}) SRE(\boldsymbol{k}_{(x2,y2)}) = &\left(\boldsymbol{R}_{\Theta,(x1,y1)}^{d} \boldsymbol{W}_q \boldsymbol{x}_{(x1,y1)}\right)^{\top} \\
&\left(\boldsymbol{R}_{\Theta,(x2,y2)}^{d} \boldsymbol{W}_k \boldsymbol{x}_{(x2,y2)}\right)
\end{aligned} \tag{7}
$$



(a) ARM and LPM (b) SPM

Figure 2: Modeling Strategy: we progressively improve modeling strategy from ARM to LPM and eventually SPM, achieving both high efficiency and high fidelity. In this Figure, the red boxes show the current view patch to be generated, while the blue boxes present the condition window.

Spherical Conditioning The autoregressive one-directional transformer decoder generates the tokens from left to right, which leads to discontinuity between the left-most and right-most boundaries. When viewing these images with the panorama viewer, we see obvious tearing artifacts at the stitching seams. To make the left-most and right-most pixels consistent, these pixels should be generated with consideration of each other. This local detail also implies a deeper defect of the aforementioned method. If the context information of the complete spherical structure is not considered when predicting the token of the current position, the consistency and integrity of the final overall result cannot be guaranteed.

In order to fix this defect, we redesign the conditions for generating each view patch. Through the two-pass mechanism, the model no longer only autoregressively focuses on the small window on the upper left but also focuses on the entire hemispherical area around the current patch view. Specifically, in the training phase, the model learns both $\mathcal{L}_{\text{LPM}}$ and $\mathcal{L}_{\text{SPM}}$ in each iteration step. The form of $\mathcal{L}_{\text{SPM}}$ is as follows:

$$\mathcal{L}_{\text{SPM}} = -\mathbb{E}\left[\sum_{\forall i \in [1,N], mask_i=1} \log p\left(y_i \mid Y_{\overline{\mathbf{M}}}; Y_S\right)\right] \quad (8)$$

It is worth noting that Y_S is different from Y_W in $\mathcal{L}_{\text{LPM}}$. For the view patch at row i and column j, the corresponding Y_W contains view patches of upper and left, while Y_S contains these of upper, down, left, and right. Specifically, Y_S comprises of $[(i-1,j),(i,j-1),(i-1,j-1)]$, and Y_S comprises of $[(i-1,j-1),(i-1,j),(i-1,j+1),(i,j-1),(i,j+1),(i+1,j-1),(i+1,j),(i+1,j+1)]$. When the above coordinates are out of bounds, Y_W will not consider the part beyond the boundary, and Y_S will extend the part beyond the x-axis to the other side. In addition, each Y_S also applies the spherical relative embedding transformation described above and a special learnable phase embedding $\mathbf{I}$ to let the model know which pass it is currently in.

During inference, the model first performs a complete LPM decoding and retains all tokens, and then superimposes the spherical relative embedding and phase embedding $\mathbf{I}$ for each token in the second pass to optimize the generation result of the first pass. In this way, although the inference time is doubled, it is still faster than ARM and can further greatly improve the generation quality, surpassing both ARM and LPM.

Guided Semantic-Condition To further enhance the capabilities of the HORIZON model, we employ semantics as an additional input condition to guide the panoramic image generation process. Specifically, we use FOV=90 degrees to cut out the front, back, left, and right perspective pictures of the panoramic image and encode them with the pretrained CLIP(Radford et al. 2021) visual module to get four semantic vectors for each panorama. We then take those vectors as semantic conditions sequentially appended to sphere conditions, and train in an end-to-end way. Similarly, we also employ the pretrained CLIP text module to encode text as a semantic condition to control the generation. During inference, we can either input visual conditions or text conditions respectively.

Experiment Setup

Dataset

We evaluate our model on the high-resolution StreetLearn dataset (Mirowski et al. 2019), which consists of Google Street View panoramas. We use the Pittsburgh dataset containing 58k images, split into 52.2k for training and 5.8k for testing. The equirectangular panorama RGB images are stored as high-quality JPEGs with dimensions of 1664 x 832. In our experiments, we resize all panoramas to 1536 x 768. The experiments are conducted on 64 V100 GPUs, each with 32GiB memory. Our model is evaluated on three typical panorama tasks: panorama generation, view extrapolation, and guided generation.

To further show the flexibility of our method on arbitrary resolutions, we also conducted experiments on a higher resolution setting of 3072x1536. However, as most previous works are unable to handle such large images, we present

only qualitative demonstrations in the appendix.To show our method are applicable to diverse scenes, we conduct experiments on Matterport 3D with different available baselines(Chen, Wang, and Liu 2022; Lin et al. 2019, 2022). Please find it in the appendix.

Task 1: Panorama Generation

The widely used Frechet Inception Distance (FID) score (Heusel et al. 2017) evaluates image quality by measuring feature distribution distances between real and fake images. However, FID treats all image positions equally, while information density in spherical images varies spatially. The top and bottom of an image often contain sparse information, while the middle holds denser information.

To account for this, we propose a novel *spherical FID* for evaluating panoramic image quality, which dynamically considers information density variation. FID is calculated on different view patch sets within an image. For a panorama with a 3-row and 6-column grid of view patches, we label each row as top, middle, and bottom, and calculate FID scores for these subsets. Table 2 reports spherical FID scores for various locations. As a baseline, we train the text2light model (Chen, Wang, and Liu 2022) on the Streetlearn dataset, with detailed settings in the appendix.

Our SPM(+SRE+SC) model generates state-of-the-art results, significantly outperforming baseline models. Both spherical relative embedding and spherical conditioning are effective mechanisms.The second pass balances efficiency and effectiveness by revising the first pass. LPM is the most efficient model, albeit with lower performance.

View Discontinuity Problem The discontinuity of synthesis occurs when the left-most and the right-most boundary merged into one spherical image. Among the four snapshots from the viewer tool, the last (4th) image rendered is the merged image in which the middle is exactly the boundary between the left most and right most. From the showcases in Figure 3, we can see that the results of the baseline algorithm have an obvious separator in the middle while our algorithm considering the spherical attention generate a smooth connection. Moreover, we evaluate this continuity quantitatively by gradient based metrics as shown in Table 2. Inspired by the metric Left-Right Consistency Error (LRCE) (Shen et al. 2022) for depth estimation, we evaluate the consistency of the left-right boundaries by calculating the horizontal gradient between the both sides of the panorama. In details, the horizontal gradient G_I^H of the image I can be written as $G_I = \max_{dim=-1} |I^{col}_{first} - I^{col}_{last}|$, where $I^{col}_{first}/I^{col}_{last}$ represents the RGB values in the first/last columns of the image I. Note that, different from LRCE, the generated panorama can not minus ground truth gradient to alleviate natural discontinuity. We choose to calculate the distribution distance instead of the absolute distance between the predicted horizontal gradient and real panorama gradient to measure the boundary continuity. The final calculation of LRCS(left-right continuity score) is as follows:

$$LRCS = KL(\mathcal{N}_{\mathcal{P}}, \mathcal{N}_{\mathcal{GT}}) \tag{9}$$

$\mathcal{N}_{\mathcal{P}}$ and $\mathcal{N}_{\mathcal{GT}}$ are two normal distributions estimated from the horizontal gradient of the predicted panorama($\{G_P\}$) and ground truth($\{G_{GT}\}$), respectively. KL means KL-distance. The lower LRCS means the panorama is more seamless. We show the result in Table 2, though Parallel Decoding increase the discontinuity, after using SRE and SC our final results has significantly resolve the problem and achieves 10 times lower LRCS than Text2light. All the above results demonstrates the effectiveness of our proposed spherical attention module.

Figure 3: Discontinuity v.s. Continuity. The three randomly selected cases present results from baseline and our models. The top images are the generated images of the baseline(LPM) method and the bottom examples are from our model(SPM). There is an obvious split line in the middle of each image on the top examples while the boundary is smooth on the bottom examples.

Task 2: View Extrapolation

We conduct quantitative experiments and adopt structural similarity (SSIM) and peak-to-signal-noise-ratio (PSNR) and FID as evaluation metrics specific for the view extrapolation tasks. Our generative model demonstrates superior performance compared to baseline methods as demonstrated in Table 3. To further illustrate the effectiveness of our approach, we have included a comparison of our method with Omnidream (Akimoto, Matsuo, and Aoki 2022a) in the appendix, where we have constrained the resolution to 1024x512 in accordance with their capabilities.

Our generative model demonstrated exceptional performance in view extrapolation, as validated by the examples shown in Figure 4. The generated panoramas not only seamlessly filled in unseen content, but also possessed reasonable structure and rich semantics. These results showcase the superior capabilities of our model in generating high-resolution, coherent panoramas.

Task 3: Guided Generation

We illustrate the guided generation showcases in Figure 5 and Figure 6. The visual guidance and text guidance are encoded by CLIP model. From these examples, we can observe that our framework can edit and modify semantic elements in the panorama by providing reference view images or text hints at specific locations. More cases can be found in the supplemental material.

Visual Guidance As shown in Figure 5, the case presents the results given the visual guidance. The left case shows the

	FID↓	Spherical FID↓				Continuity
		mean	top	middle	bottom	LRCS↓
Text2light(Chen, Wang, and Liu 2022)	36.33	56.31	48.05	60.28	60.62	0.0224
ARM	25.36	41.21	26.16	32.17	65.32	0.0283
LPM	45.71	57.13	64.34	46.71	60.35	0.2032
SPM(+SRE)	10.74	39.18	28.72	26.44	62.39	0.0726
SPM(+SRE+SC)	**7.79**	**20.97**	**15.82**	**21.80**	**25.29**	**0.0020**

Table 2: Generation results. SPM is the spherical model in our paper, SRE is spherical relative embedding and SC is spherical conditioning.

	$SSIM\uparrow$	$PSNR\uparrow$	FID↓
ARM	0.521	15.39	11.78
LPM	0.508	14.94	17.62
SPM	0.542	15.49	5.53

Table 3: View Extrapolation Results.

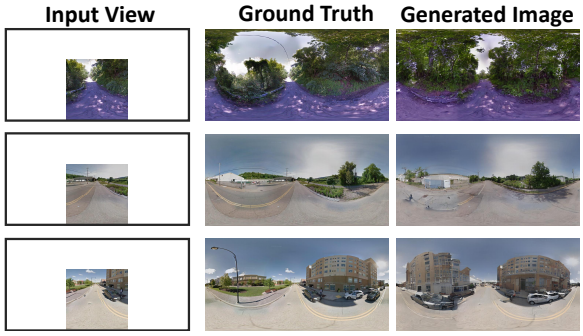


Figure 4: View Extrapolation. The first column gives the input samples, the second column presents the ground truth examples, and the third column demonstrates the generated panorama.

original panoramic image (bottom) as well as 4 FOV images rendered (top 4 images). The middle and right cases demonstrate the generated results given the 4th guided image highlighted in the red box. The middle case edits the final view with "a single tree", and the right case edits the final view with "lush trees".This demonstrates the model is capable of generating panoramic images with both semantic and style controls. Please note, in order to generate a consistent image, the context view may be changed accordingly.

Text Guidance As shown in Figure 6,the cases presents the results given the text guidance. In the Figure, the first row contains the original panorama images, the second row presents the text guidance and the third row illustrates the generated panoramic images. As shown in these cases, the highlighted red box shows the corresponding region modified. The semantic text guidance is "lush trees", and the trees in these images are modified accordingly.

Conclusion

In this study, we present an innovative framework for crafting high-resolution panoramic visuals, skillfully integrating spherical structure and semantic control. By employing spherical modeling, we adeptly tackle spherical distortion and edge continuity challenges while facilitating generation through image and text cues. Future endeavors will focus on embedding interactive features and enhancing inference speed, ultimately positioning the model as a viable alternative to current human-built interfaces.

Figure 5: Visual Guided generation. When we replace the guidance with different visual semantics as shown in the middle and right columns, we can manipulate the generated panoramas as we need.

Figure 6: Textual Guided generation. We can also use natural language to edit or embellish target panoramas. In each case, the top role are the original panoramas, and the bottom role are the embellished panoramas according to the text hint shows in the middle.

Acknowledgements

This work was conducted during Kun Yan's internship at Microsoft Research Asia and supported in part by NSFC 61925203 and U22B2021.

References

Akimoto, N.; Kasai, S.; Hayashi, M.; and Aoki, Y. 2019. 360-degree image completion by two-stage conditional gans. In *2019 IEEE International Conference on Image Processing (ICIP)*, 4704–4708. IEEE.

Akimoto, N.; Matsuo, Y.; and Aoki, Y. 2022a. Diverse Plausible 360-Degree Image Outpainting for Efficient 3DCG Background Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11441–11450.

Akimoto, N.; Matsuo, Y.; and Aoki, Y. 2022b. Diverse Plausible 360-Degree Image Outpainting for Efficient 3DCG Background Creation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11431–11440.

Brown, M.; and Lowe, D. G. 2007. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74(1): 59–73.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. MaskGIT: Masked Generative Image Transformer. *ArXiv*, abs/2202.04200.

Chen, Z.; Wang, G.; and Liu, Z. 2022. Text2Light: Zero-Shot Text-Driven HDR Panorama Generation. *ACM Transactions on Graphics (TOG)*, 41(6): 1–16.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2021. Vector quantized diffusion model for text-to-image synthesis. *arXiv preprint arXiv:2111.14822*.

Han, S. W.; and Suh, D. Y. 2020. PIINET: A 360-degree Panoramic Image Inpainting Network Using a Cube Map. *arXiv preprint arXiv:2010.16003*.

Hara, T.; Mukuta, Y.; and Harada, T. 2021. Spherical Image Generation from a Single Image by Considering Scene Symmetry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1513–1521.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NIPS*.

Karimi Dastjerdi, M. R.; Hold-Geoffroy, Y.; Eisenmann, J.; Khodadadeh, S.; and Lalonde, J.-F. 2022. Guided Co-Modulated GAN for 360° Field of View Extrapolation. *arXiv e-prints*, arXiv–2204.

Kim, G.; and Ye, J. C. 2021. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv preprint arXiv:2110.02711*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koh, J. Y.; Agrawal, H.; Batra, D.; Tucker, R.; Waters, A.; Lee, H.; Yang, Y.; Baldridge, J.; and Anderson, P. 2022. Simple and Effective Synthesis of Indoor 3D Scenes. *arXiv preprint arXiv:2204.02960*.

Liao, K.; Xu, X.; Lin, C.; Ren, W.; Wei, Y.; and Zhao, Y. 2022. Cylin-Painting: Seamless 360 {\deg} Panoramic Image Outpainting and Beyond with Cylinder-Style Convolutions. *arXiv preprint arXiv:2204.08563*.

Lin, C. H.; Chang, C.; Chen, Y.; Juan, D.; Wei, W.; and Chen, H. 2019. COCO-GAN: Generation by Parts via Conditional Coordinating. In *IEEE International Conference on Computer Vision (ICCV)*.

Lin, C. H.; Cheng, Y.-C.; Lee, H.-Y.; Tulyakov, S.; and Yang, M.-H. 2022. InfinityGAN: Towards Infinite-Pixel Image Synthesis. In *International Conference on Learning Representations*.

Mirowski, P.; Banki-Horvath, A.; Anderson, K.; Teplyashin, D.; Hermann, K. M.; Malinowski, M.; Grimes, M. K.; Simonyan, K.; Kavukcuoglu, K.; Zisserman, A.; et al. 2019. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1505–1514.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, 1060–1069. PMLR.

Rockwell, C.; Fouhey, D. F.; and Johnson, J. 2021. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14104–14113.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Shen, Z.; Lin, C.; Liao, K.; Nie, L.; Zheng, Z.; and Zhao, Y. 2022. PanoFormer: Panorama transformer for indoor 360 depth estimation. *arXiv e-prints*, arXiv–2203.

Somanath, G.; and Kurz, D. 2021. HDR Environment Map Estimation for Real-Time Augmented Reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11298–11306.

Song, S.; and Funkhouser, T. 2019. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6918–6926.

Song, S.; Zeng, A.; Chang, A. X.; Savva, M.; Savarese, S.; and Funkhouser, T. 2018. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3847–3856.

Su, J.; Lu, Y.; Pan, S.; Wen, B.; and Liu, Y. 2021. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*.

Sumantri, J. S.; and Park, I. K. 2020. 360 panorama synthesis from a sparse set of images with unknown field of view. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2386–2395.

Szeliski, R. 2006. Image alignment and stitching: a tutorial, foundations and trends in computer graphics and computer vision. *Now Publishers*, 2(1): 120.

van den Oord, A.; Vinyals, O.; and kavukcuoglu, k. 2017. Neural Discrete Representation Learning. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, G.; Yang, Y.; Loy, C. C.; and Liu, Z. 2022. StyleLight: HDR Panorama Generation for Lighting Estimation and Editing. *arXiv preprint arXiv:2207.14811*.

Wu, C.; Liang, J.; Hu, X.; Gan, Z.; Wang, J.; Wang, L.; Liu, Z.; Fang, Y.; and Duan, N. 2022. NUWA-Infinity: Autoregressive over Autoregressive Generation for Infinite Visual Synthesis.

Xu, J.; Zheng, J.; Xu, Y.; Tang, R.; and Gao, S. 2021. Layout-guided novel view synthesis from a single indoor panorama. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16438–16447.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1316–1324.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021a. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *CVPR*.

Zhang, W.; Wang, Y.; and Liu, Y. 2022. Generating High-Quality Panorama by View Synthesis Based on Optical Flow Estimation. *Sensors*, 22(2): 470.

Zhang, Y.; Xiao, J.; Hays, J.; and Tan, P. 2013. Framebreak: Dramatic image extrapolation by guided shift-maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1171–1178.

Zhang, Z.; Ma, J.; Zhou, C.; Men, R.; Li, Z.; Ding, M.; Tang, J.; Zhou, J.; and Yang, H. 2021b. M6-UFC: Unifying Multi-Modal Controls for Conditional Image Synthesis. *arXiv preprint arXiv:2105.14211*.