

Spherical Manifold Guided Diffusion Model for Panoramic Image Generation

Xiancheng Sun¹ Mai Xu¹ Shengxi Li^{1*} Senmao Ma¹ Xin Deng¹ Lai Jiang¹ Gang Shen²
¹Beihang University ²China Tower



Figure 1. A subjective comparison of zero-shot text-conditioned panoramic image generation between our SMGD model and the state-of-the-art method PanFusion [53]. Our model demonstrates superior performance in generating images with superior overall natural content, enhanced subjective quality, and better alignment with the text description, while benefiting reduced model parameters.

Abstract

Panoramic image essentially acts as a pivotal role in emerging virtual reality and augmented reality scenarios; however, the generation of panoramic images are essentially challenging due to the intrinsic spherical geometry and spherical distortions caused by equirectangular projection (ERP). To address this, we start from the very basics of S^2 manifold inherent to panoramic images, and propose a novel spherical manifold convolution (SMConv) on S^2 manifold. Based on the SMConv operation, we propose a spherical manifold guided diffusion (SMGD) model for text-conditioned panoramic image generation, which can well accommodate the spherical geometry during generation. We further develop a novel evaluation method by calculating grouped Fréchet inception distance (FID) on cube-map projections, which can well reflect the quality of generated panoramic images, compared to existing methods that randomly crop ERP-distorted content. Experiment results demonstrate that our SMGD model achieves the state-of-the-art generation quality and accuracy, whilst retaining

the shortest sampling time in the text-conditioned panoramic image generation task. Codes are publicly available at <https://github.com/chronos123/SMGD>.

1. Introduction

Being able to be controlled by text descriptions, panoramic image generation with a full 360° horizontal and 180° vertical field-of-view [44] opens a new avenue for various applications in virtual reality [48], augmented reality [47], autonomous driving [46], etc. Although recent developments in deep probabilistic models including generative adversarial networks (GANs) [14, 20, 34, 57] and diffusion models [10, 32, 33] have been acting as the workhorse in generating photo-realistic images, panoramic image generation still remains challenging due to their intrinsic spherical geometry property and distorted scene structures. In particular, the equirectangular projection (ERP) [50] employed to convert the panoramic content into a planar rectangular format, introduces spherical distortions and breaks the continuity at the left and right edges. Within the ERP format, rectangular panoramic images with an aspect ratio of 1 : 2 exhibit a

*Corresponding author.

non-uniform pixel distribution. In other words, the pixel density is notably sparse near the polar regions and significantly denser around the equatorial region. Consequently, existing state-of-the-art deep probabilistic models designed for generating planar images [31, 34, 49, 51] still yet to generalize well for the panoramic image generation.

To address the panoramic image generation task, several methods proposed to generate rectangular images, including autoregressive generation based on the latent codes [11], stitching latent features [3], and directly autoregressive generation on image patches [24]. To retain the continuity between the left and right boundaries, strategies such as circular convolution [5], circular token mechanisms [2], and exploiting the symmetry of panoramic images [15] have been further proposed. However, the above methods fail to effectively mitigate the spherical distortions, thus motivating the usage of deformable convolution [43] and attention mechanisms [53]. Unfortunately, existing methods still suffer from sub-optimal performance, which lacks well-defined global structure and optimal preservation of spherical geometry, as illustrated in Fig. 1, whilst compensating this with heavy computational complexity.

In this paper, we start from the intrinsic property of panoramic format, and propose the spherical manifold guided diffusion (SMGD) model operating on S^2 manifold, to generate panoramic images from text descriptions. Since panoramic images can be recognized as pixels distributed on S^2 manifold, we propose the spherical manifold convolution (SMConv), to effectively address spherical distortions and preserve spherical geometry in panoramic images. To the best of our knowledge, there exists no spherical convolution designed for the panorama generation task, in which our SMConv fills this void by performing the correct convolution on spherical surfaces. Based on SMConv, we establish spherical manifold encoder (SME) and spherical manifold decoder (SMD) blocks, constituting a spherical manifold U-Net (SMUNet). Our SMGD model is subsequently constructed by employing the SMUNet as the denoising network, with the VQGAN [11] serving as its image encoder. Unlike previous spherical convolutions [8, 36, 37, 45, 55] applied in classification and detection tasks, our SMConv is supported by the intrinsic manifold theory. Guided by exponential mapping, our SMConv incorporates the convolution within spherical neighborhoods under the measure of the geodesic distance [40]. Moreover, to further address the spherical geometry of panoramic images, we develop a spherical loss to optimize the proposed SMGD model. We also adopt an enhanced strategy to evaluate the quality of generated panoramic images by calculating the grouped Fréchet inception distance (FID) [17] in cubemap projection (CMP) [52] format. Experimental results demonstrate that our method achieves superior generation performances by effectively addressing the unique challenges associated with panoramic

image generation, such as complex scene structure, spherical distortions caused by ERP and the sparse pixel distribution near the poles. In summary, our contributions are three-fold:

- We set out the first attempt to propose the spherical convolution, namely, SMConv, operating on S^2 manifold, which is able to optimally capture the intrinsic spherical geometry in panoramic images.
- We propose to incorporate SMConv into spherical manifold guided blocks, including the SME and SMD blocks, so as to mitigate spherical distortions and preserve spatial coherence across the spherical domain.
- Building upon these foundational blocks, we propose the SMGD model as the first generative approach to incorporate spherical convolution and achieve the state-of-the-art performance in generating panoramic images from text descriptions.

2. Related Works

Panoramic Image Generation. Early research on panoramic image generation can be achieved by the image outpainting task, which generates panoramic images from partial view [1, 2, 9, 15, 28, 38, 42]. However, these outpainting methods are limited by the freedom of control over the generated content. Most recently, with the rapid advancements in generative models, text-conditioned panoramic image generation [3, 5, 23, 24, 39, 41, 43, 53, 54] has received increasing research efforts. More specifically, Text2Light [5] first proposed to generate panoramic images from text descriptions, using CLIP [30] to encode text together with VQGAN [11] to encode panoramic images. To generate images at arbitrary sizes, DiffCollage [54], Multi-Diffusion [3] and SyncDiffusion [23] aggregated the intermediate outputs of pre-trained diffusion models. PanoGen [24] proposed to recursively outpaint the images generated by pretrained models to create panoramic images. However, those methods either fail to handle spherical distortions nor to ensure continuity between the left and right boundaries. To address this problem, MVDiffusion [39] generated perspective images and stitches them into panoramic images. However, the stitched panoramic image may witness its limitation in the field of view. PanFusion [53] introduced an attention mechanism for perspective images to enhance the generation of local details, while the main branch focused on generating the overall panoramic image. However, PanFusion utilizes two pretrained diffusion models [18] whereby multiple perspective images are processed in a separate branch; this can lead to unnatural global structures and result in heavy computational complexity.

Convolution Operation for Panoramic Images. Several attempts have introduced modifications to the convolution operation [7, 8, 12, 21, 29, 36, 37, 45, 55] in order to effectively manage spherical distortions for panoramic image classification and segmentation tasks. The pioneering work [36]

adopted kernels with different shapes at different latitudes in object detection task. Inspired by these works, deformable kernels [12, 55] were adopted to handle the spherical geometry. On the other hand, Coors *et al.* [8] retained the kernel unchanged on the tangent plane of the sphere and projected it back to the sphere using inverse gnomonic projection in image classification task. For the same task, Xu *et al.* [45] proposed to enlarge the receptive field of convolution kernel by defining the kernel area in sphere. Furthermore, the rotation invariance property has been considered for designing spherical convolution in image classification [7] and image segmentation tasks [29]. However, spherical convolution operations have not yet been proposed for generative models that essentially suffer from improper spherical geometry. Therefore, we propose to incorporate the guidance from SMConv by exponential mapping, operating on S^2 manifold for high quality panoramic image generation.

3. Methodology

3.1. Spherical Manifold Convolution

The spherical distortion introduced by ERP leads to object deformation and interrupts the continuity of content on the sphere, which together significantly hinder the correct information aggregation of standard convolution. In contrast, performing convolution operation directly on the spherical domain is of significant potential to preserve the spherical geometry of panoramic images and to mitigate spherical distortions. By inspecting that panoramic images can be viewed as pixels distributed on a spherical surface, *i.e.*, the S^2 manifold, we propose the SMConv operation that performs convolution on the S^2 manifold. More specifically, we employ exponential mapping [13] to project uniform convolution kernels from the tangent space onto the S^2 manifold, enabling SMConv with kernels to uniformly aggregate the content across the S^2 manifold. More importantly, due to the length-preserving property of exponential mapping [13], SMConv retains the uniformity of geodesic distances, known as great-circle distances [40], on the S^2 manifold. This enables SMConv to better handle spherical distortions, providing an effective solution for panoramic image generation. Compared to previous spherical convolution based on inverse gnomonic projection [8], employing exponential mapping to construct SMConv kernels enables an expanded receptive field, refined kernel patterns, and improved preservation of spherical geometry, as demonstrated in Fig. 2.

More specifically, the S^2 manifold is defined as a set of points $x \in \mathbb{R}^3$ with norm 1, which can be parameterized by spherical coordinates $\theta \in [0, 2\pi]$ and $\varphi \in [-\pi/2, \pi/2]$. Through a transformation between spherical and Cartesian coordinates, we can obtain 3D Cartesian coordinates and then apply the exponential mapping. For instance, a point $\mathbf{p} = (\theta, \varphi)$ on S^2 manifold can be converted to a corre-

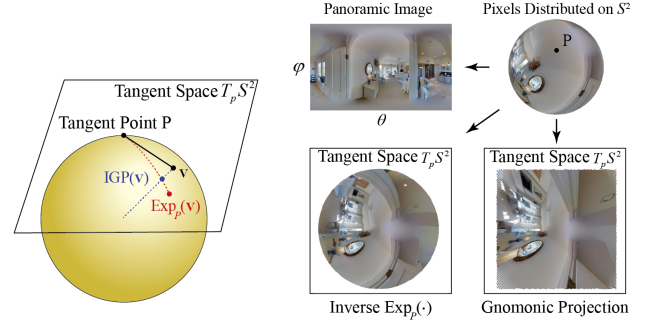


Figure 2. Comparison of exponential mapping (denoted as $\text{Exp}_p(\cdot)$) and inverse gnomonic projection (denoted as $\text{IGP}(\cdot)$) [8] through forward and backward mappings. Left: visualization of a point \mathbf{v} in the tangent space $T_p S^2$ at point \mathbf{p} , mapped onto the sphere. Right: projection of a panoramic image, the spherical data, onto the tangent space $T_p S^2$ centered at $\mathbf{p} = (\theta, \varphi)$ whereby $\theta = 0$ and $\varphi = \frac{\pi}{4}$.

sponding point $\mathbf{p} = (x, y, z)$ with Cartesian coordinates. The exponential mapping [13] can be formulated as follows:

$$\mathbf{q} = \text{Exp}_p(\mathbf{v}) = \cos(\|\mathbf{v}\|)\mathbf{p} + \sin(\|\mathbf{v}\|)\frac{\mathbf{v}}{\|\mathbf{v}\|}, \quad (1)$$

where \mathbf{q} denotes the mapped point, \mathbf{p} is the tangent point, and \mathbf{v} is a vector representing the coordinates of the kernel points $\mathbf{K}_{T_p}(m, n)$ in tangent space $T_p S^2$ centered at \mathbf{p} .

In practice, given a predefined position (u_m, v_n) of a kernel point $\mathbf{K}_{T_p}(m, n)$ in the tangent space $T_p S^2$, the corresponding vector \mathbf{v} in $T_p S^2$ can be obtained as follows:

$$\begin{aligned} \mathbf{v} &= u_m \hat{\mathbf{e}}_\theta + v_n \hat{\mathbf{e}}_\varphi, \\ \hat{\mathbf{e}}_\theta &= (-\sin \theta, \cos \theta, 0), \\ \hat{\mathbf{e}}_\varphi &= (-\sin \varphi \cos \theta, -\sin \varphi \sin \theta, \cos \varphi), \end{aligned} \quad (2)$$

where $\mathbf{p} = (\theta, \varphi)$ denotes the center point in $T_p S^2$ and the vectors $\hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_\varphi$ serve as the basis for the coordinate system in $T_p S^2$. Note that the shape and the position of the convolution kernel are provided by (u_m, v_n) in $T_p S^2$ and we can calculate the position $\mathbf{q} = \text{Exp}_p(\mathbf{v}) = (\hat{x}_{m,n}, \hat{y}_{m,n}, \hat{z}_{m,n})$ of the kernel point $\mathbf{K}(m, n)$ on S^2 manifold in Cartesian coordinates, via (1) and (2). Here, the kernel \mathbf{K} is centered at the point \mathbf{p} on S^2 manifold. Due to the properties of the exponential mapping [13], the norm of the mapped point $\mathbf{q} = (\hat{x}_{m,n}, \hat{y}_{m,n}, \hat{z}_{m,n})$ is guaranteed to be 1, ensuring that the point remains on the S^2 manifold. Since we need the position $\mathbf{q} = (\theta_m, \varphi_n)$ in spherical coordinates for performing SMConv, we can simply apply the inverse transformation from Cartesian to spherical coordinates.

Furthermore, suppose $\tilde{\mathbf{F}}_{S^2}, \mathbf{F}_{S^2}$ are features on S^2 manifold. The operation of SMConv can be formulated by

$$\begin{aligned} \tilde{\mathbf{F}}_{S^2}(\theta, \varphi) &= \sum_{m,n} \mathbf{F}_{S^2}(\theta + \theta_m, \varphi + \varphi_n) \cdot \mathbf{K}(m, n), \\ (\theta_m, \varphi_n) &= \mathbf{q} = \text{Exp}_p(\mathbf{K}_{T_p}(m, n)), \end{aligned} \quad (3)$$

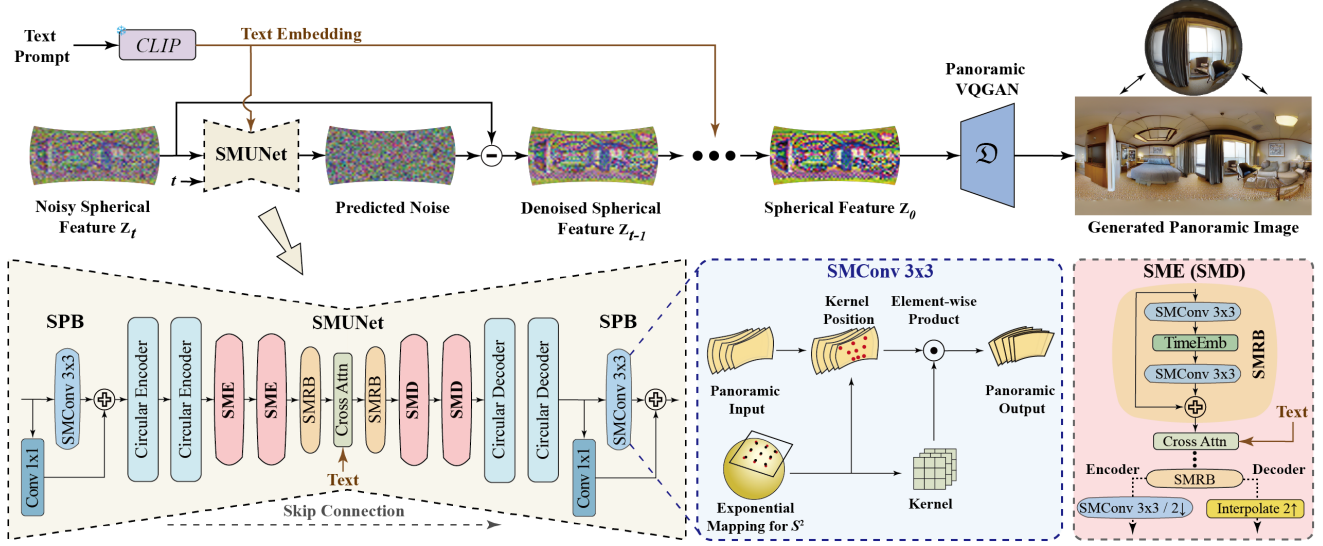


Figure 3. Overall architecture of the proposed SMGD model, which introduces SMUNet as the denoising network and a panoramic VQGAN as the image encoder. The SMUNet is primarily constructed by the SMConv, operating on the S^2 manifold, and integrates the SPB along with circular encoder and decoder blocks.

In (3), \mathbf{q} is the position of the kernel point $\mathbf{K}(m, n)$ for spherical features in ERP format, which is obtained by performing exponential mapping for the kernel point $\mathbf{K}_{T_p}(m, n)$ in the tangent space $T_p S^2$. Moreover, \mathbf{K} represents the kernel centered at point \mathbf{p} on S^2 manifold and the weight of \mathbf{K} is identical to that of \mathbf{K}_{T_p} . By uniformly choosing u_m and v_m and constructing a regular kernel in the tangent space $T_p S^2$, we can perform SMConv with kernels uniformly distributed across the S^2 manifold using exponential mapping, according to (3).

3.2. SMConv Inspired Diffusion Model

Building upon SMConv, we introduce the SMGD model to achieve high-quality panoramic image generation. The overall architecture of our SMGD model is shown in Fig. 3. The diffusion process can be formulated as

$$q(\mathbf{z}_t | \mathbf{z}_{t-1}) := \mathcal{N}(\mathbf{z}_{t-1}; \sqrt{1 - \beta_t} \mathbf{z}_{t-1}, \beta_t \mathbf{I}),$$

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\psi(\mathbf{z}_t, t, \mathbf{T}) \right) + \sqrt{\beta_t} \eta, \quad (4)$$

where $\mathbf{z}_{t-1}, \mathbf{z}_t$ denote the noisy spherical features with the same dimensionality as the data, at time step $t - 1 \in [0, T]$ and $t \in [0, T]$ respectively. \mathbf{T} is the text embedding obtained by CLIP model [31], i.e., from the pretrained CLIP ViT/L-14 model. In addition, $q(\mathbf{z}_t | \mathbf{z}_{t-1})$ represents the forward process gradually adding noise to the image, $\epsilon_\psi(\mathbf{z}_t, t)$ is the noise predicted by our SMUNet at time step t , ψ represents the parameters of SMUNet and η denotes a random Gaussian noise. Furthermore, the noise scheduler configuration β_t is a hyperparameter that controls the strength of the noise involved at each time step. The parameter α_t is defined as

$\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t$ is given by $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, the default setting within diffusion models.

As the core component of the SMGD model, our SMUNet predicts noise at each time step t from the noisy spherical feature \mathbf{z}_t and facilitates the recovery of the spherical feature \mathbf{z}_0 on S^2 manifold by subtracting the predicted noise. We employ spherical processor blocks (SPBs), along with circular encoder and decoder blocks based on circular convolution [2] at shallower layers of SMUNet. Then, we apply SME and SMD blocks at deeper layers with middle blocks consisted of spherical manifold residual blocks (SMRBs) and cross-attention layers to integrate text embeddings \mathbf{T} . The design of a hybrid architecture to integrate circular convolution at shallower layers with SMConv at deeper layers is motivated by the following considerations. As resolution increases, spherical convolution operation including SMConv, consumes substantially higher GPU memory cost compared to standard convolution [8] due to its bilinear interpolation. Furthermore, in deeper layers of SMUNet where spatial resolution is reduced, the sparse pixel distribution near the poles becomes increasingly pronounced, exacerbating the limitations of standard convolution. Consequently, the proposed hybrid architecture well balances between output quality, computational efficiency, and memory utilization.

Then, the SME block comprises 2 SMRBs, 2 cross-attention layers, and an SMConv with a stride of 2 for downsampling. For the l -th SME block, the SMRB can be formulated as

$$\tilde{\mathbf{z}}_t^{E_l} = \text{SMConv}[\text{SMConv}[\mathbf{z}_t^{E_l}] + g_\psi(t)] + \mathbf{z}_t^{E_l}, \quad (5)$$

where the calculation of SMConv follows (3) and $g_\psi(t)$ is the learnable time-step embedding function [32]. Moreover,

$\mathbf{z}_t^{E_l}$ denotes the input, while $\tilde{\mathbf{z}}_t^{E_l}$ represents the output for this layer. The structure of the SMD block closely mirrors that of the encoding SME block, with the upsampling layer using nearest-neighbor interpolation. Note that the final SME block and the first SMD block omit upsampling and downsampling operations. Moreover, the SPB employs the SMConv combined with a 1×1 residual convolutional layer to enhance the transformation between different feature channels. The circular encoder and decoder blocks follow the structure in LDM [32], with all planar convolutions replaced by circular convolution operations [2]. Notably, the circular encoder blocks use average pooling layers for downsampling. In addition to the SMUNet, we propose to employ a panoramic VQGAN [11] with a downsampling factor of 16 as the image encoder and decoder for our SMGD model. In the panoramic VQGAN, all planar convolutions in the layers of decoder are replaced by circular convolutions [2].

3.3. Spherical Guided Training Strategy

The training process involves two stages: training the panoramic VQGAN, and then training the SMGD model. During the training of the panoramic VQGAN, reconstruction loss \mathcal{L}_{rec} [11], commitment loss $\mathcal{L}_{\text{Commit}}$ [11] and perceptual loss \mathcal{L}_{Per} [19] are employed, which can be formulated as

$$\mathcal{L}_{\text{VQGAN}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{commit}} \mathcal{L}_{\text{commit}} + \lambda_{\text{per}} \mathcal{L}_{\text{Per}}, \quad (6)$$

where λ_{commit} and λ_{per} are hyperparameters to control the weights of different loss components during training.

During the training of the SMGD model, a spherical loss based on the weighted spherically uniform PSNR (WS-PSNR) [50] is employed. The total loss can be formulated as

$$\mathcal{L}_{\text{Diff}} = \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}} + \lambda_{\text{SMSE}} \mathcal{L}_{\text{SMSE}}, \quad (7)$$

where λ_{MSE} and λ_{SMSE} are also balancing hyperparameters. Furthermore, the \mathcal{L}_{MSE} and $\mathcal{L}_{\text{SMSE}}$ are given by

$$\begin{aligned} \mathcal{L}_{\text{MSE}} &= \mathbb{E}_{t, \mathbf{z}_0, \epsilon} [\|\epsilon - \epsilon_{\psi}(\mathbf{z}_t, t, \mathbf{T})\|_2^2], \\ \mathcal{L}_{\text{SMSE}} &= \mathbb{E}_{t, \mathbf{z}_0, \epsilon} [\|\mathbf{W} \odot (\epsilon - \epsilon_{\psi}(\mathbf{z}_t, t, \mathbf{T}))\|_2^2], \end{aligned} \quad (8)$$

where \mathbf{W} is given by the WS-PSNR [50] to reflect the panoramic characteristics as

$$\mathbf{W}_j = \cos\left(\left(\frac{j+0.5}{H} - 0.5\right)\pi\right), \quad (9)$$

where $j \in [0, H]$ is the index of height in \mathbf{z}_t and H denotes the height of the spherical feature \mathbf{z}_t . $\mathcal{L}_{\text{SMSE}}$ can improve the preservation of spherical geometry, *e.g.*, a sparse pixel distribution near the poles.

3.4. Quality Evaluation for Panoramic Images

FID score is the *de facto* standard for evaluating the quality of generated images, whereas performing a direct calculation

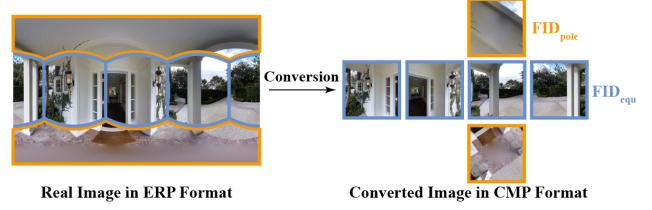


Figure 4. Illustration of the conversion between ERP and CMP formats. In the CMP format, four blue-bordered images represent equatorial regions, while two orange-bordered images represent polar areas. These images are grouped accordingly, with FID_{equ} and FID_{pole} calculated separately for each group.

of FID on panoramic images in rectangular ERP format introduces severe distortion, leading to inaccuracy. We propose to convert the generated panoramic images from ERP format to CMP format consisted of 6 square images for calculating FID score. The mapping from ERP to CMP format [50, 52] provides a robust solution to evaluate the quality of generated panoramic images. Note that although sharing similar goals with the concurrent work [6], our FID-based metrics are able to reflect detailed quality for different regions when evaluating generated panoramic images. As illustrated in Fig. 4, the conversion reveals the continuity across the left and right boundaries and represents the spherical geometry by obtaining 4 images representing areas near the equator and 2 images representing areas near the poles. Thus, the 6 square images within CMP format can be categorized into 2 groups, allowing for separate FID calculations to assess the quality of generated images. More specifically, FID_{equ} is calculated using the 4 images near the equator, while FID_{pole} is derived from the 2 images near the poles. FID_{equ} primarily assesses the quality and visual fidelity of the panoramic content, whereas FID_{pole} evaluates the preservation of spherical geometry. Analyzing FID_{equ} and FID_{pole} provides a comprehensive measure of the quality of the generated panoramic image.

4. Experiment

4.1. Experimental Settings

Dataset. We employed the widely-used Matterport3D [4] dataset to evaluate the performance for text-conditioned panoramic image generation. More specifically, we obtained 10,912 panoramic images with a resolution of 1024×512 according to [26]. We then utilized the BLIP-2 [25] model to generate text descriptions for each panoramic image, forming paired text-image data for training and testing. To enable FID-10K calculation, we allocated 10,000 images for training and testing.

Implementation Details. We adopted the panoramic VQGAN according to Sec. 3.2. SMUNet, the denoising network for our SMGD model, was configured with 320 and 640 channels for the two circular encoder and decoder blocks,

Table 1. Comparisons in terms of FID and CLIP score (denoted as CS) values for panoramic images generated by our method and existing baseline methods, along with evaluations of parameter number (denoted by Para.) and time required to inference (denoted as t_{sample}). We represent the best numbers by **red color** and the second best by **blue color**.

| Methods | FID ↓ | FAED ↓ | OmniFID ↓ | FID _{avg} ↓ | FID _{cent} ↓ | FID _{bord} ↓ | FID _{rand} ↓ | FID _{equ} ↓ | FID _{pole} ↓ | CS ↑ | Para. ↓ | $t_{\text{sample}}(\text{s})$ ↓ |
|---------------------|--------------|-------------|--------------|----------------------|-----------------------|-----------------------|-----------------------|----------------------|-----------------------|--------------|-------------|---------------------------------|
| Multi-Diffusion [3] | 69.96 | 1.12 | 68.70 | 56.74 | 58.19 | 56.53 | 55.50 | 27.26 | 87.73 | 0.182 | 1.3B | 39.45 ± 0.31 |
| SD2+LoRA [18, 32] | 59.32 | 1.47 | 69.07 | 50.04 | 51.78 | 49.16 | 49.17 | 24.49 | 88.68 | 0.171 | 1.4B | 6.48 ± 0.49 |
| Text2Light [5] | 48.42 | 4.10 | 63.16 | 43.68 | 46.93 | 46.64 | 37.47 | 30.87 | 72.06 | 0.188 | 1.0B | 50.12 ± 11.18 |
| MVDiffusion [39] | 56.21 | 0.93 | 94.53 | 38.15 | 38.88 | 37.60 | 37.96 | 27.58 | 125.48 | 0.181 | 1.4B | 97.14 ± 0.79 |
| PanFusion [53] | 16.15 | 0.35 | 27.64 | 14.34 | 14.91 | 13.96 | 14.14 | 14.81 | 31.99 | 0.182 | 2.3B | 38.51 ± 11.57 |
| SMGD (Ours) | 9.85 | 0.22 | 18.24 | 9.70 | 9.77 | 9.62 | 9.71 | 7.88 | 21.05 | 0.189 | 1.4B | 2.69 ± 0.39 |

and 960 and 1280 channels for the two SME and SMD blocks utilizing the 3×3 SMConv. At the first training stage, we trained the panoramic VQGAN with the Adam optimizer [22], using a batch size of 1 and a learning rate of 4.5×10^{-6} for 300 epochs, with $\lambda_{\text{commit}} = 1$ and $\lambda_{\text{per}} = 0.8$. At the second stage, we trained the SMGD model from the scratch with the AdamW optimizer [27], a batch size of 8, and learning rate of 1×10^{-5} for 800 epochs by setting $\lambda_{\text{MSE}} = 0.8$ and $\lambda_{\text{SMSE}} = 0.2$. Moreover, the noise scheduler configuration β_t followed that of LDM [32]. For inference, we used a DDIM sampler [35] with 50 steps and a classifier-free guidance scale [32] of 2.5.

Baselines. We compared our SMGD model with existing text-conditioned panoramic image generation methods, including Multi-Diffusion [3], Text2Light [5], MVDiffusion [39] and PanFusion [53]. We also compared our SMGD model with the widely-used stable diffusion model [32] by performing a rank-16 LoRA [18] (denoted as SD2+LoRA). Different from [53], we directly trained the LoRA model with a learning rate of 1×10^{-6} for 50 epochs on our training data.

Evaluation Metrics. We computed FID_{equ} and FID_{pole} , as described in Sec. 3.4 to effectively assess the quality of panoramic images. Note that FID_{equ} was calculated against 40K images and FID_{pole} was calculated against 20K images. To further validate our metrics and assess the quality of the generated images, we calculated an extra FID value similar to that used in previous methods [53]. More specifically, we cropped the panoramic images into 3 representative areas instead of obtaining 20 perspective images in calculating FID values. Specifically, we cropped the panoramic images into the center, cross-border and random patches with size 512×512 . Then, we calculated the corresponding FID value of them and obtained the FID_{cent} , FID_{bord} and FID_{rand} . By averaging these three FID values calculated against 10K images, we obtained FID_{avg} . Moreover, we reported FID, OmniFID [6], and Fréchet Auto-Encoder Distance (FAED) [53] to comprehensively evaluate the quality of the generated full panoramic images. On the other hand, to evaluate the alignment between text and generated panoramic images, we utilized the widely adopted CLIP score [16] (denoted

as CS).¹ Furthermore, we recorded the time for inference t_{sample} by averaging the time required to generate 10 images on an NVIDIA 4090 GPU for each method.

4.2. Comparison with Existing Methods

Quantitative Results. In Table 1, we report the quantitative results of our method and baselines for panoramic image generation from text descriptions. From this table, our method consistently achieves the lowest FID values, the highest CLIP scores, and the least sampling times. More importantly, among the top-performance models, our method maintains a comparatively lower parameter number. More specifically, the lower values of FID_{cent} , FID_{bord} , FID_{rand} , and FID_{avg} demonstrate that our SMGD model achieves the state-of-the-art performance within text-conditioned panoramic image generation². Moreover, the even lower FID_{equ} and FID_{pole} values indicate that our method well preserves the spherical geometry of panoramic images, compared to other baseline methods. Note that MVDiffusion [39] and SD2+LoRA achieved low FID_{avg} values but high FID_{pole} values. This may be due to the fact that images generated by MVDiffusion exhibited a limited field of view, while images produced by SD2+LoRA lacked the preservation of the spherical geometry in panoramic images. Furthermore, as shown in Table 1, our SMGD model achieves the fastest inference time among all comparison models, outperforming PanFusion [53], the second-best method in terms of image quality, by more than a factor of 10 in inference speed.

Qualitative Results. We randomly chose 2 text descriptions and report the generated panoramic images with corresponding real images from dataset in Fig. 5. As illustrated in the figure, the panoramic images generated by our SMGD model achieved the best subjective quality and the most accurate alignment with the text descriptions. More specifically, **improper spherical geometry**, including inconsistency across the border and an unnatural pixel distribution near the poles, was observed in panoramic images generated by SD2+LoRA, Text2Light, and Multi-Diffusion, due to their

¹The CLIP score was computed using the ViT-L/14 CLIP model, consistent with the CLIP model employed in our SMGD model.

²To further validate the robustness of our model, we conduct additional evaluations on the Structured3D [56] dataset in Supplementary Sec. A. We also provide additional qualitative results in Supplementary Sec. C.

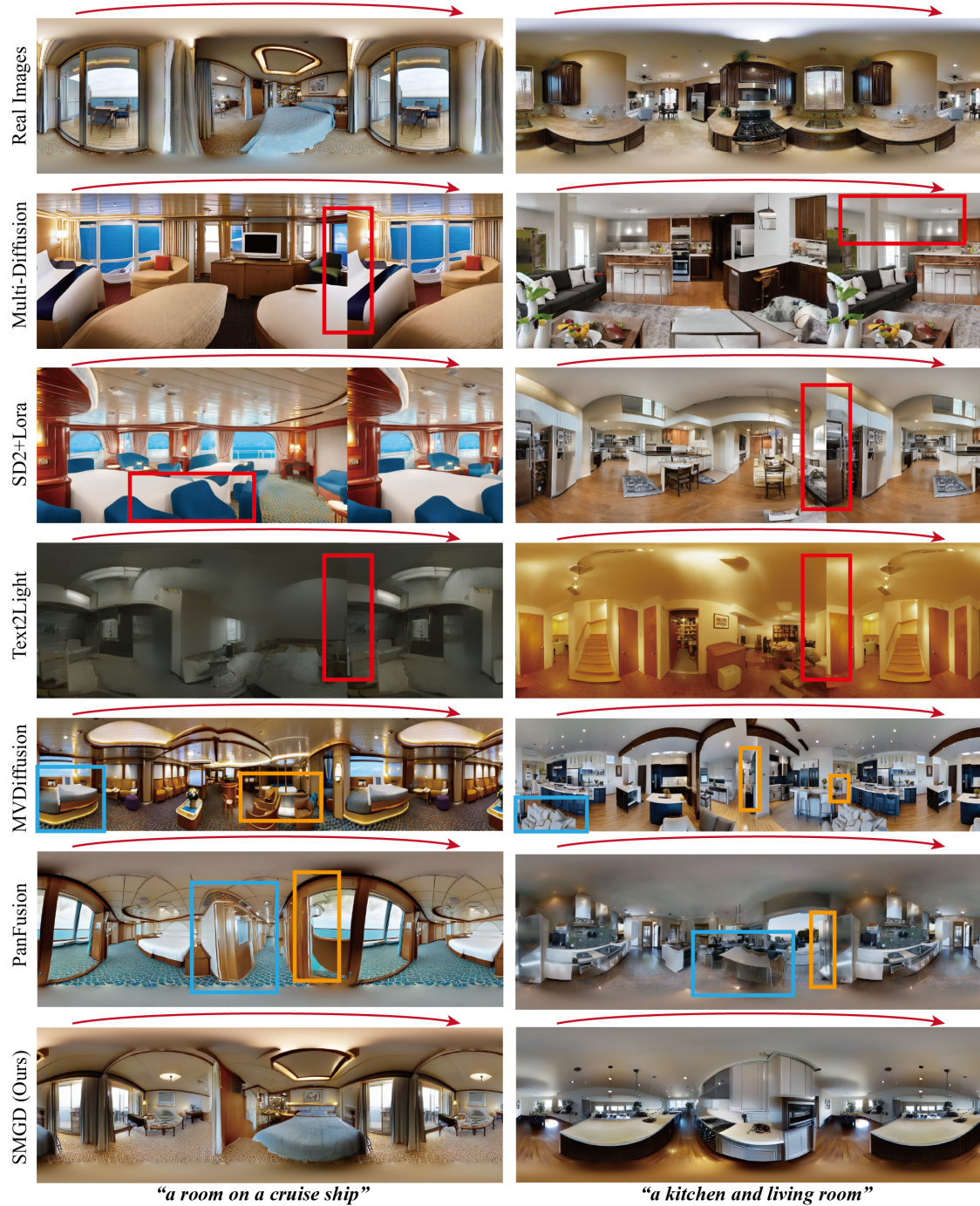


Figure 5. Qualitative results for text-conditioned panoramic image generation with real images from dataset displayed on the top row. We also show the copy from the left side of the panoramic image to the right to reveal the consistency across the whole image. Note that images generated by MVDiffusion [39] exhibit a limited field of view, resulting in images with a reduced height. We highlight the **improper spherical geometry**, **artifact regions** and **distorted objects and lines** with corresponding color of boxes, which are mitigated by our SMGD model.

neglect of spherical distortions. Panoramic images generated by Panfusion and MVDiffusion suffered from **artifact regions** and **distorted objects and lines** as highlighted with corresponding color boxes in Fig. 5. Moreover, images

generated by MVDiffusion exhibited a limited field of view. This limitation impeded the visual quality and completeness of the generated images, making them less suitable for fully immersive visual experiences. In contrast, our SMGD

Table 2. Ablation study for our SMGD model. We compare the performance achieved by adding individual components to the baseline model sequentially, including the spherical loss (denoted by $\mathcal{L}_{\text{SMSE}}$), the SMD blocks, the SME blocks, and the SPB modules. We represent the best numbers by **red color**.

| $\mathcal{L}_{\text{SMSE}}$ | SMD Blocks | SME Blocks | SPB | FID | FID _{avg} ↓ | FID _{cent} ↓ | FID _{bord} ↓ | FID _{rand} ↓ | FID _{equ} ↓ | FID _{pole} ↓ | CS ↑ |
|-----------------------------|------------|------------|-----|-------------|----------------------|-----------------------|-----------------------|-----------------------|----------------------|-----------------------|--------------|
| × | × | × | × | 23.11 | 20.57 | 20.34 | 20.91 | 20.45 | 15.35 | 34.87 | 0.187 |
| ✓ | × | × | × | 20.03 | 17.50 | 17.05 | 18.04 | 17.40 | 12.78 | 32.95 | 0.187 |
| ✓ | ✓ | × | × | 13.35 | 12.87 | 12.79 | 12.94 | 12.88 | 9.61 | 20.45 | 0.188 |
| ✓ | ✓ | ✓ | × | 12.36 | 11.42 | 11.46 | 11.60 | 11.19 | 9.21 | 21.70 | 0.186 |
| ✓ | ✓ | ✓ | ✓ | 9.85 | 9.70 | 9.77 | 9.62 | 9.71 | 7.88 | 21.05 | 0.189 |

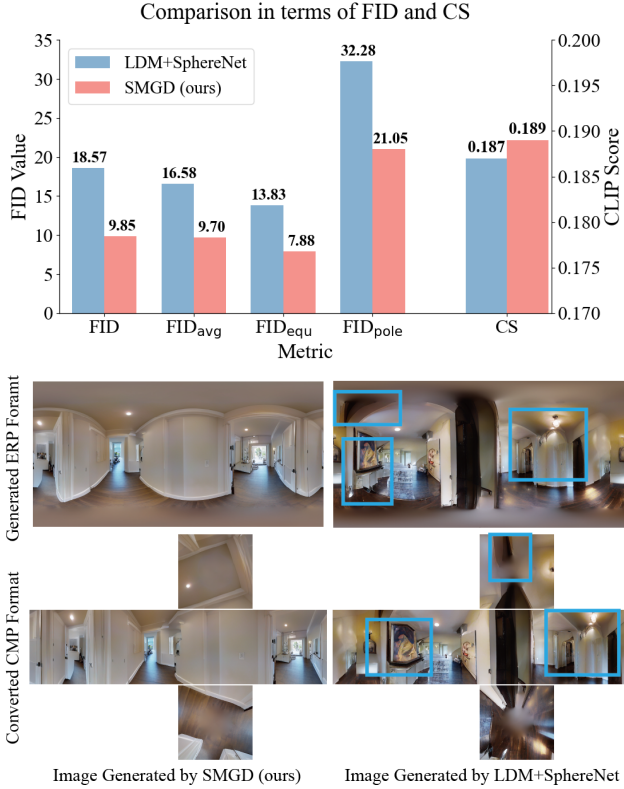


Figure 6. Comparison between LDM+SphereNet and the proposed SMGD model. Top: Bar plot of FID and CLIP score values achieved by these two methods. Bottom: Subjective results of panoramic images generated by the text prompt “a hallway in a house”. The comparison is conducted in both ERP and CMP formats, with **distorted objects and lines** highlighted.

model produced visually pleasing panoramic images with a well-defined overall structure and proper spherical geometry.

4.3. Ablation Study

In this section, we conducted ablation study to systematically evaluate the impact of core components in our SMGD model. More specifically, since the spherical loss and SMConv are important for handling spherical distortions and preserving spherical geometry in panoramic images, we analyzed performance changes by gradually adding the spherical loss $\mathcal{L}_{\text{SMSE}}$, the SMD blocks, the SME blocks and the SPB. As reported in Table 2, the involvement of SMD blocks

significantly enhanced performance, with SME blocks further contributing to the improvements. Moreover, both the spherical loss function and the SPB effectively improved the quality of generated panoramic images. Therefore, each module plays a vital role in optimizing the performance of the SMGD model.

Furthermore, to validate the effectiveness of SMConv compared to previous spherical convolution methods, we implemented LDM+SphereNet, which employs the panoramic VQGAN and a U-Net architecture identical to our SMUNet, whilst all SMConv layers were replaced by the convolution operations proposed in SphereNet [8]. The training and inference were kept consistent with those of our SMGD model. As shown in Fig. 6, we can conclude that the quality of panoramic images generated by LDM+SphereNet is limited, whereas our proposed SMGD model produces visually pleasing images with lower FID values and a higher CLIP score³. This demonstrates the effectiveness of using SMConv for advanced panoramic image generation.

5. Conclusion

In this paper, we have proposed a novel method named as spherical manifold diffusion (SMGD) model to address the challenges of spherical distortions associated with panoramic image generation. This was achieved by the spherical manifold convolution (SMConv) operation that is able to uniformly sample spherical content within each convolution layer. Different from existing spherical convolution operations applied in other tasks, the proposed SMConv operation was built upon the exponential mapping grounded in manifold theory, which is able to preserve lengths on the tangent plane and exhibit enhanced local properties. The overall training procedure of our SMGD model was also inspired by the spherical loss. Furthermore, we have proposed to calculate FID_{equ} and FID_{pole} for panoramic images in cubemap format to improve the evaluation on the quality of generated panoramic images. Experimental results have demonstrated that our method achieves superior performance in text-conditioned image generation, surpassing existing state-of-the-art methods in both the quality of generated images and computational efficiency.

³More qualitative comparison results between SMGD and LDM+SphereNet, as well as a detailed analysis, are provided in Supplementary Sec. B.

Acknowledgement

This work was supported by Natural Science Foundation of China (NSFC) under Grants 62450131, 62206011, 62231002 and 62401027, and Beijing Natural Science Foundation under Grant L223021.

References

- [1] Hao Ai, Zidong Cao, Haonan Lu, Chen Chen, Jian Ma, Pengyuan Zhou, Tae-Kyun Kim, Pan Hui, and Lin Wang. Dream360: Diverse and immersive outdoor virtual scene creation via transformer-based 360° image outpainting. *IEEE Transactions on Visualization and Computer Graphics*, 2024. [2](#)
- [2] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11441–11450, 2022. [2](#), [4](#), [5](#)
- [3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*, 2023. [2](#), [6](#)
- [4] 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. *arXiv preprint arXiv:1709.06158*, 2017. [5](#)
- [5] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. [2](#), [6](#)
- [6] Anders Christensen, Nooshin Mojab, Khushman Patel, Karan Ahuja, Zeynep Akata, Ole Winther, Mar Gonzalez-Franco, and Andrea Colaco. Geometry fidelity for spherical images. In *European Conference on Computer Vision*, pages 276–292. Springer, 2024. [5](#), [6](#)
- [7] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018. [2](#), [3](#)
- [8] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 518–533, 2018. [2](#), [3](#), [4](#), [8](#)
- [9] 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 *2022 International Conference on 3D Vision (3DV)*, pages 475–485. IEEE, 2022. [2](#)
- [10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. [1](#)
- [11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. [2](#), [5](#)
- [12] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and Jose J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, 5(2):1255–1262, 2020. [2](#), [3](#)
- [13] Jean Gallier. *Basics of Classical Lie Groups: The Exponential Map, Lie Groups, and Lie Algebras*, pages 367–414. Springer New York, New York, NY, 2001. [3](#)
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. [1](#)
- [15] Takayuki Hara, Yusuke Mukuta, and Tatsuya Harada. Spherical image generation from a single image by considering scene symmetry. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1513–1521, 2021. [2](#)
- [16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. [6](#)
- [17] 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. *Advances in neural information processing systems*, 30, 2017. [2](#)
- [18] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. [2](#), [6](#)
- [19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer International Publishing, 2016. [5](#)
- [20] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10124–10134, 2023. [1](#)
- [21] Renata Khasanova and Pascal Frossard. Geometry aware convolutional filters for omnidirectional images representation. In *International conference on machine learning*, pages 3351–3359. PMLR, 2019. [2](#)
- [22] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [6](#)
- [23] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36:50648–50660, 2023. [2](#)
- [24] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#)
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. [5](#)
- [26] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan:

- Generation by parts via conditional coordinating. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4512–4521, 2019. 5
- [27] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [28] Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14211–14219, 2024. 2
- [29] Thomas W Mitchel, Noam Aigerman, Vladimir G Kim, and Michael Kazhdan. Möbius convolutions for spherical cnns. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2, 3
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 2, 4
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 4, 5, 6
- [33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1
- [34] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pages 30105–30118. PMLR, 2023. 1, 2
- [35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6
- [36] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. *Advances in neural information processing systems*, 30, 2017. 2
- [37] Yu-Chuan Su and Kristen Grauman. Kernel transformer networks for compact spherical convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9442–9451, 2019. 2
- [38] Julius Surya Sumantri and In Kyu Park. 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*, pages 2386–2395, 2020. 2
- [39] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdifusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *ArXiv*, abs/2307.01097, 2023. 2, 6, 7
- [40] T. Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 23(176):88–93, 1975. 2, 3
- [41] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4933–4943, 2024. 2
- [42] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *The Twelfth International Conference on Learning Representations*, 2023. 2
- [43] Tao Wu, Xuewei Li, Zhongang Qi, Di Hu, Xintao Wang, Ying Shan, and Xi Li. Spherediffusion: Spherical geometry-aware distortion resilient diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6126–6134, 2024. 2
- [44] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020. 1
- [45] Yanyu Xu, Ziheng Zhang, and Shenghua Gao. Spherical dnns and their applications in 360 images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7235–7252, 2021. 2, 3
- [46] Jian-Ru Xue, Jian-Wu Fang, and Pu Zhang. A survey of scene understanding by event reasoning in autonomous driving. *International Journal of Automation and Computing*, 15(3):249–266, 2018. 1
- [47] 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 Transactions on Graphics (TOG)*, 41(4):1–10, 2022. 1
- [48] 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. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 650–660. IEEE, 2024. 1
- [49] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022. 2
- [50] Yan Ye, Elena Alshina, and Jill Boyce. Algorithm descriptions of projection format conversion and video quality metrics in 360lib. In *Joint Video Exploration Team of ITU-T SG*, 2018. 1, 5
- [51] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2
- [52] Li Yu, Yanjun Gao, Farhad Pakdaman, and Moncef Gabbouj. Panoramic image inpainting with gated convolution and contextual reconstruction loss. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4255–4259, 2024. 2, 5

- [53] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6347–6357, 2024. [1](#), [2](#), [6](#)
- [54] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10188–10198. IEEE, 2023. [2](#)
- [55] Qiang Zhao, Chen Zhu, Feng Dai, Yike Ma, Guoqing Jin, and Yongdong Zhang. Distortion-aware cnns for spherical images. In *IJCAI*, pages 1198–1204, 2018. [2](#), [3](#)
- [56] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. [6](#)
- [57] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17907–17917, 2022. [1](#)