DiffPano: Scalable and Consistent Text to Panorama Generation with Spherical Epipolar-Aware Diffusion

Weicai Ye[1,3,*] **Chenhao Ji**[2,*] **Zheng Chen**[4] **Junyao Gao**[2] **Xiaoshui Huang**[3]
Song-Hai Zhang[4] Wanli Ouyang[3] Tong He[3,✉] Cairong Zhao[2,✉] Guofeng Zhang[1,✉]
[1]State Key Lab of CAD&CG, Zhejiang University [2]Tongji University
[3]Shanghai AI Laboratory [4]Tsinghua Univerisity
maikeyeweicai@gmail.com jichenhao@tongji.edu.cn zhangguofeng@zju.edu.cn

Figure 1: **DiffPano allows scalable and consistent panorama generation (i.e. room switching) with given unseen text descriptions and camera poses.** Each column represents the generated multi-view panoramas, switching from one room to another.

Abstract

Diffusion-based methods have achieved remarkable achievements in 2D image or 3D object generation, however, the generation of 3D scenes and even 360° images remains constrained, due to the limited number of scene datasets, the complexity of 3D scenes themselves, and the difficulty of generating consistent multi-view images. To address these issues, we first establish a large-scale panoramic video-text dataset containing millions of consecutive panoramic keyframes with corresponding panoramic depths, camera poses, and text descriptions. Then, we propose a novel text-driven panoramic generation framework, termed DiffPano, to achieve scalable, consistent, and diverse panoramic scene generation. Specifically, benefiting from the powerful generative capabilities of stable diffusion, we fine-tune a single-view text-to-panorama diffusion model with LoRA on the established panoramic video-text dataset. We further design a spherical epipolar-aware multi-view diffusion model to ensure the multi-view consistency of the generated panoramic images.

*: Equal Contribution. ✉: Corresponding Author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0041

Extensive experiments demonstrate that DiffPano can generate scalable, consistent, and diverse panoramic images with given unseen text descriptions and camera poses. Code and dataset are available at https://zju3dv.github.io/DiffPano.

1 Introduction

Generating scenarios from the text descriptions that meet one's expectations is an imaginative and marvelous journey, which has many potential applications, such as VR roaming [62] for metaverse [67, 68, 69, 72], physical world simulation [2, 6, 14], embodied agents in scene navigation and manipulation [75], etc. With the advent of the AIGC era, several works [8, 14, 20, 29, 42, 43, 48, 51, 52] on object generation even scene generation has emerged. However, they generally only generate a series of perspective images, making it impossible to comprehensively simulate the entire environment for scene understanding [27, 19, 66] and reconstruction [3, 5, 65, 64, 30, 35, 34, 26, 21, 7, 12, 47, 49, 58]. Given this, some methods [70, 9] try to generate panoramas to solve these problems by taking advantage of the inherent characteristics of panoramic images, which can capture the surrounding environment with a single shot.

These methods can be roughly divided into four categories: 1) Directly single-view equirectangular projection (ERP) panorama generation [70]. However, its camera is immovable, making it incapable of scene exploration; 2) Multiple perspective views generation methods [20, 43, 48, 40] without considering multi-view panorama generation. 3) The inpainting solutions are based on the infinite expansion of a single perspective view, which lacks 3D awareness. Single scene optimization methods [53] with inpainting have no generalization ability and cost too much time for optimization. 4) Directly extending the multiple perspective views generation methods [43] to the ERP panorama, which is difficult to converge and results in poor multi-view consistency (see Fig. 5).

This paper aims to generate scalable and multi-view consistent panoramic images from text descriptions and camera poses (see Fig. 1) with many potential applications such as immersive VR roaming with unlimited scapes and preview for interior home design. However, achieving this goal is not trivial. To the best of our knowledge, there is currently a lack of rich and diverse panoramic datasets to meet the task of text-to-multi-view ERP panorama generation. To this end, we propose a novel panoramic video-text dataset and a generation framework suitable for the text-to-multi-view panorama generation task, advancing the development of this field. Specifically, we first establish a large-scale panoramic video-text dataset using Habitat Simulator [41] (see Sec. 3), which contains millions of panoramic keyframes and corresponding panoramic depths, camera poses, and text descriptions. Next, built upon the proposed dataset, we propose a generation framework for consistent multi-view ERP panorama generation (see Sec. 4), termed DiffPano. The DiffPano framework consists of a single-view text-to-panorama diffusion model (see Sec. 4.1) and a spherical epipolar-aware multi-view diffusion model (see Sec. 4.2). The single-view text-to-panorama diffusion model is obtained by fine-tuning the stable diffusion model [38] of perspective images using LoRA [18]. Considering that the single-view pano-based diffusion model cannot guarantee the consistency of generated multi-view panoramas with different camera poses, we derive a spherical epipolar constraint applicable to panoramic images, inspired by the perspective epipolar constraint. We then incorporated it as a spherical epipolar-aware attention module (see Sec. 4.2) into the multi-view panoramic diffusion model to ensure the multi-view consistency of the generated ERP panoramic images.

Since there are no related methods for comparison, we try to extend the MVDream method [43] to generate multi-view ERP panoramas. Trained and tested on the proposed panoramic video-text dataset, extensive experiments demonstrated that compared to the modified MVDream, our proposed multi-view panorama generation based on spherical epipolar-aware attention can generate more scalable and consistent panoramic images. Our method also demonstrates the generalization ability of the original diffusion model to generate satisfactory multi-view ERP panoramas with given unseen text descriptions and camera poses.

Our contributions can be summarized as follows: 1) To the best of our knowledge, we are the first to propose a scalable and consistent multi-view panorama generation task from text descriptions and camera poses. 2) We established a large-scale diverse and rich panoramic video-text dataset, which fosters the research of text-to-panoramic video generation. 3) We propose a novel text-driven panoramic generation framework with a spherical epipolar attention module, allowing scalable and consistent panorama generation with unseen text descriptions and camera poses.

2 Related work

2.1 Single-View Panorama Generation

Recently, latent diffusion model (LDM) methods have attracted widespread attention, and many single-view panorama generation works [70, 48, 74, 24, 60, 56, 33, 54, 62] have emerged, achieving remarkably impressive results. Among them, MVDiffusion [48] simultaneously generates eight fixed-viewpoint perspective images through a multi-view Correspondence-Aware diffusion model and stitches them together to produce a panorama. However, it cannot support the generation of top and bottom views, and the generated panorama resembles wide-angle images with an extensive field of view rather than true 360° images. Some methods [55, 13] solve this problem using equirectangular projection (ERP) and try to facilitate the interaction between the left and right sides during the panorama generation process to enhance the left-right continuity property inherent in ERP images. To address the domain gap between panorama and perspective images, PanFusion [70] proposed a novel dual-branch diffusion model that mitigates the distortion of perspective images projected on panoramas while providing global layout guidance. However, its more complex model architecture incurs longer inference times for panorama generation. In addition, PanFusion cannot be expanded as an effective pre-trained model to the multi-view panorama generation task due to its excessive network parameters. To strike a balance between computational complexity and ensuring left-right continuity of panoramas, our proposed single-view panorama-based stable diffusion model only requires fine-tuning with LoRA [18] to learn panoramic styles and achieve good edge continuity while maintaining higher generation speed and simpler architecture.

The existing single-view panorama generation methods cannot achieve scalable panorama generation. The core of our paper lies in the generation of multi-view consistent panoramic images, which we will introduce in Section 2.2. More importantly, the single-view panoramic image generated by previous methods mainly supports 3DoF roaming, while our method can generate multi-view panoramic images for 6DoF roaming, which can serve as the inputs for 360° Gaussian Splatting [23] or 360° NeRF [10, 11]. Our method also has a great potential value in 360° relightable novel view synthesis with the combination of 360° multi-view inverse rendering method [25].

2.2 Multi-View Image Generation

To the best of our knowledge, there is no work focusing on multi-view panorama generation. We review the existing works about multi-view generation for perspective images in this part.

Zero123 [29] laid the foundation for 3D object generation based on multi-view generation, while the pose-guided diffusion model [51, 40] explored consistent view synthesis of scenes. However, iteratively applying the diffusion model to generate individual views in the multi-view generation task may lead to poor multi-view consistency of generated images due to accumulated errors. To generate high-quality multi-view images simultaneously, some methods [31, 43, 57, 32] modify the UNets in the diffusion model into a multi-branch form and achieve the effect of generating consistent multi-view images through the interaction between different branch.

Currently, most multi-view generation tasks focus on generating multi-view perspective images of single objects [43, 61, 63, 28, 46, 45] or scenes [51, 15], while minimal research has been conducted on multi-view panorama generation. Narrow FoV (field-of-view) drawbacks of perspective images lead to the fact that the existing generation methods can only generate a very local region of the scene at a time. Our work focuses on the task of exploring the generation of 360° images from multiple different viewpoints. Due to the camera projection difference between panoramic and perspective images, achieving consistency in multi-view panoramas is challenging. It is impossible to directly apply the existing epipolar attention module [51, 20] to multi-view panoramas. We strive to derive the spherical epipolar line formula for panoramic images and propose a spherical epipolar attention module to ensure the multi-view consistency of the generated panoramas.

2.3 Panoramic Dataset

Great progress in text to single-view panorama generation has been witnessed. However, text-to-multi-view panorama generation is still a blank slate. One of the main limitations of this task is the lack of suitable datasets. The common panoramic datasets used in single-view panorama generation consist of indoor HDR dataset [16], outdoor HDR dataset [71], HDR360-UHD dataset [9], Structured3D [73],

 https://doi.org/10.52202/079017-0041

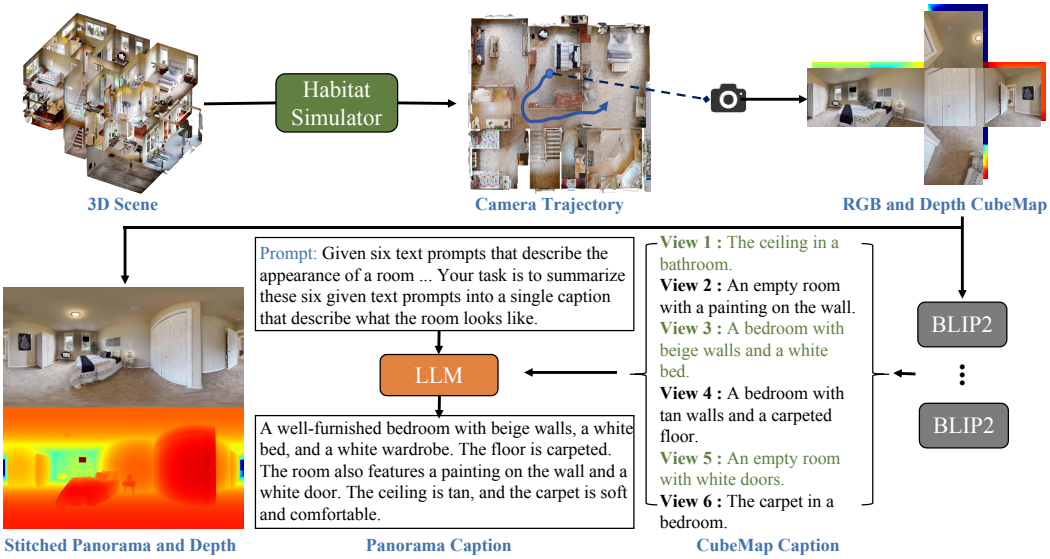

Figure 2: **Panoramic Video Construction and Caption Pipeline.**

Standford 2D-3D-S [1], and Matterport3D dataset [4], etc. Most of these datasets are relatively small in scale and only have single-view panoramas, which cannot support multi-view panorama generation, except Matterport3D [4]. In addition, the sky box images in Matterport3D [4] contain only sparse views. Although HM3D [37] provides the textured mesh of 1000 scenes, it lacks the corresponding text description for each view. To generate multi-view panoramas, we render cube maps at each viewpoint in the 3D meshes of HM3D, using the Habitat Simulator [41], and stitch them into panoramas. We generate complete text descriptions corresponding to the panoramas by using Blip2 [22] to create text descriptions for each face of the cube map separately, and then summarizing them using Llama2 [50]. In this way, we obtain a panoramic video-text dataset that includes camera poses, corresponding panoramas, and text descriptions of the panoramas, which facilitates subsequent multi-view panorama generation tasks.

3 Panoramic Video-Text Dataset

Due to the lack of high-quality panorama-text datasets, most text-to-panorama generation tasks require researchers to construct their own datasets. The dataset constructed in PanFusion [70] suffers from blurriness at the top and bottom of the panoramic images, and the corresponding text descriptions are not precise enough. To address these issues, we utilize the Habitat Simulator [41] to randomly select positions within the scenes of the Habitat Matterport 3D (HM3D) [37] dataset and render the six-face cube maps. These cube maps are then interpolated and stitched together to form panoramas so we can obtain panoramas with clear tops and bottoms. To generate more precise text descriptions for the panoramas, we first use BLIP2 [22] to generate corresponding text descriptions for each obtained cube map, and then employ Llama2 [50] to summarize and receive accurate and complete text descriptions. Furthermore, the Habitat Simulator allows us to render images based on camera trajectories within the HM3D scenes, enabling the creation of a dataset that simultaneously includes camera poses, panoramas, and corresponding text descriptions. This dataset will be utilized in the multi-view panorama generation (see Sec. 4.2).

4 Proposed Method: DiffPano

DiffPano is capable of generating multi-view consistent panoramas conditioned on camera viewpoints and textual descriptions, as illustrated in Fig. 1. In this section, we first introduce our single-view panorama stable diffusion in Sec. 4.1. We then elaborate on how to extend single-view panorama generation to multi-view consistent panorama generation by leveraging the spherical epipolar attention module in Sec. 4.2.

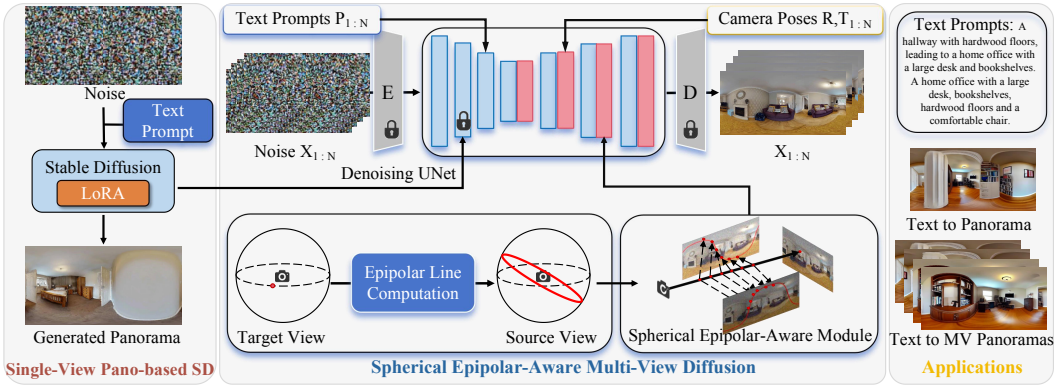

Figure 3: **DiffPano Framework.** The DiffPano framework consists of a single-view text-to-panorama diffusion model and a spherical epipolar-aware multi-view diffusion model. It can support text to single-view panorama or multi-view panorama generation.

4.1 Single-View Panorama-Based Stable Diffusion

A straightforward way to generate a single-view panorama from text is to train a text-to-panorama diffusion model from scratch with a large number of text-panorama pairs, which is both time-consuming and computationally expensive. However, stable diffusion [38] leverages a vast amount of perspective images and their corresponding textual descriptions as training data, endowing it with excellent prior knowledge of perspective images and strong text understanding capabilities. An economical and effective way for panorama generation is to fine-tune the trained perspective diffusion model with a few text-panorama pairs. To this end, panorama generation from text can be regarded as a style transfer of images generated by stable diffusion, converting them from perspective style to panoramic style, and requiring them to satisfy the left-right continuity property of panoramas.

LoRA-based fine-tuning Diffusion models for text-to-image generations possess excellent prior knowledge of 2D images and strong text comprehension capabilities. We aim to preserve these abilities of the model while fine-tuning it to generate images in the style of panoramas. We employ the Low-Rank Adaptation (LoRA) [18] fine-tuning method, which has been previously used in large language models. Compared to full fine-tuning, LoRA is faster and requires fewer computational resources. In our approach, we freeze all the parameters of the original Stable Diffusion model and add trainable layers to the UNet component using the LoRA fine-tuning method. We then train the model using our custom-created panorama-text dataset. To improve the left-right continuity of the generated images, we perform data augmentation on the panorama training dataset by randomly concatenating a portion of the right side of the panorama to the left side. Experiments demonstrate that the panoramas generated using this method exhibit satisfactory left-right continuity.

4.2 Spherical Epipolar-Aware Multi-View Diffusion

Built upon our proposed single-view panorama stable diffusion in Sec. 4.1, we extend the single-view diffusion model to a multi-view diffusion model with a spherical epipolar-aware attention module to generate multi-view scalable and consistent panoramas.

Spherical Epipolar-Aware Attention Module Epipolar attention was proposed in [20, 51] to ensure consistency between generated multi-view perspective images. However, due to the differences in imaging methods between perspective and panoramic views, existing epipolar attention cannot be directly used for panoramic views. To overcome this challenge, we derived the epipolar line for panoramic images in the equirectangular projection (ERP), and the specific proof process is provided in Appendix A. Equation 14 shows the mathematical form of the spherical epipolar line in ERP images. The spherical epipolar line is visualized in the spherical epipolar-aware attention module of Fig. 3. We extend the principle of epipolar attention to panoramic images to implement the spherical epipolar-aware attention module. Given a pixel $\mathbf{p}$ in the target view, we calculate its corresponding spherical coordinates $\mathbf{p}_{sphere}$ based on the spherical projection process:

https://doi.org/10.52202/079017-0041

$$\begin{aligned}\theta &= (0.5 - \frac{x_{pix}}{W}) \cdot 2\pi \\ \phi &= (0.5 - \frac{y_{pix}}{H}) \cdot \pi,\end{aligned} \tag{1}$$

where x_{pix} and y_{pix} are the pixel coordinates of $\mathbf{p}$, θ and ϕ are its corresponding spherical coordinates and W and H are the resolutions of panorama. We then transform the spherical coordinate system to the Cartesian coordinate system to obtain the camera coordinates p_{camera} corresponding to $\mathbf{p}$:

$$\begin{aligned}x_{cam} &= \cos(\phi) \cdot \sin(\theta) \\ y_{cam} &= \sin(\phi) \\ z_{cam} &= \cos(\phi) \cdot \cos(\theta).\end{aligned} \tag{2}$$

The camera coordinates of $\mathbf{p}_{cam}$ are converted to world coordinates $\mathbf{p}_{world}$ through the camera's pose matrix. This allows us to compute the ray from the camera center to $\mathbf{p}_{world}$ in the world coordinate system and construct the spherical epipolar attention module.

Given N feature maps $F = F_i | 1 \leq i \leq N$ corresponding to N panoramic images and their respective camera pose matrices, we implement cross-attention between different views through the spherical epipolar-aware module. During the generation process, each feature map in F can be considered as the target view, and the K nearest views are selected from the remaining features as reference views.

For each feature point in the target view feature map, we uniformly sample S points on the ray between the feature point and the camera. All sampled points are reprojected onto the feature maps of the K reference views, and the corresponding feature values are obtained through interpolation. We denote the features in the target view as q, and the features of all sampled points in the reference views as k and v. The cross-attention is then constructed using these features.

Let p_i be a feature point in the target view feature map F_t, and $p_{i,j} | 1 \leq j \leq S$ be the S uniformly sampled points on the ray between p_i and the camera center. We reproject these points onto the K reference view feature maps $F_{r_k} | 1 \leq k \leq K$ to obtain the corresponding feature values $f_{i,j,k} | 1 \leq j \leq S, 1 \leq k \leq K$. The query q_i, key k_i, and value v_i for the cross-attention mechanism are defined as follows:

$$q_i = F_t(p_i), \quad k_i = f_{i,j,k} | 1 \leq j \leq S, 1 \leq k \leq K, \quad v_i = f_{i,j,k} | 1 \leq j \leq S, 1 \leq k \leq K. \tag{3}$$

The cross-attention output o_i for the feature point p_i is computed as:

$$o_i = Attention(q_i, k_i, v_i) = softmax(\frac{q_i k_i^T}{\sqrt{d}}) v_i, \tag{4}$$

where d is the dimension of the query and key vectors.

Positional Encoding To enhance the model's understanding of 3D spatial relationships between different views, we follow EpiDiff [20] to employ the positional encoding method from Light Field Networks (LFN) [44]. In the world coordinate system, let p_i be a pixel in the target view, and o_i and d_i be the origin and direction of the ray between p_i and the camera center, respectively. The Plücker coordinates r_i of the ray are computed as:

$$r_i = (o_i \times d_i, d_i). \tag{5}$$

For each sampled point $p_{i,j}$ on the ray, its corresponding spherical depth $z_{i,j}$ is transformed using a harmonic transformation to get $\gamma_z(z_{i,j})$. Similarly, the Plücker coordinates r_i are transformed as $\gamma_r(r_i)$.The positionally encoded features $\gamma_r(r_i)$ and $\gamma_z(z_{i,j})$ are then concatenated to obtain the combined positional encoding $\gamma(r_i, z_{i,j})$:

$$\gamma(r_i, z_{i,j}) = [\gamma_r(r_i), \gamma_z(z_{i,j})]. \tag{6}$$

The combined positional encoding $\gamma(r_i, z_{i,j})$ is then concatenated with the feature maps F_t and $F_{r_k} | 1 \leq k \leq K$ to obtain the enhanced feature representations $\hat{F}t$ and $\hat{F}r_k | 1 \leq k \leq K$:

$$\hat{F}t(p_i) = [F_t(p_i), \gamma(r_i, z_{i,j})], \quad \hat{F}r_k(p_{i,j}) = [F_{r_k}(p_{i,j}), \gamma(r_i, z_{i,j})], \tag{7}$$

where $[\cdot, \cdot]$ denotes concatenation. These enhanced feature representations are then used in the cross-attention mechanism to improve the model's understanding of 3D spatial relationships between different views.

Two-Stage Training The main difference between panoramic images and perspective images is that panoramic images contain 360° content of the surroundings, while perspective images only contain content from a given viewpoint. Therefore, when the camera only rotates or translates by a small amount, the corresponding content in the panoramic image hardly changes. To make the generated multi-view panoramic images better match each corresponding text, we divide the training into two stages. In the first stage, we use the selected dataset with almost no change in image content (small camera movement) for training, which enhances the effect of the spherical epipolar-aware attention module. In the second stage, we increase the camera movement distance between each viewpoint and train with images that generate new content, improving the model's ability to understand text based on changes in perceived spatial location while ensuring multi-view consistency and enhancing scalability.

5 Experiment

Dataset We leverage the Habitat Simulator [41] to render a panoramic video dataset based on the Habitat Matterport 3D (HM3D) dataset [37]. The pipeline of dataset rendering and captioning is shown in Sec. 3. After post-processing such as dataset filtering, we constructed 8,508 panorama-text pairs as training sets for single-view panorama generation. For multi-view panorama generation, we constructed 19,739 multi-view panorama-text pairs with nearly identical image content and 18,704 multi-view panorama-text pairs with different image contents as training sets. For specific details regarding the dataset, please refer to Appendix B.

Implementation Details In the multi-view panorama generation, we simultaneously generate $N = 4$ panoramas from different viewpoints. Within the spherical epipolar-aware attention module, we consider the two nearest views to the target view as reference views, i.e., $K = 3$, and sample $S = 10$ points along each ray. We conducted separate training for 100 epochs on datasets with almost identical image content and datasets with varying image content. Please refer to Appendix C for further implementation details.

Evaluation Metrics To evaluate the performance of our proposed single-view panorama-based stable diffusion model, we employ three commonly used metrics: Fréchet Inception Distance (FID) [17], Inception Score (IS) [39], and CLIP Score (CS) [36]. FID measures the similarity between the distribution of generated panoramas and the distribution of real images. IS assesses the quality and diversity of the generated panoramas. CS is utilized to evaluate the consistency between the input text and the generated panoramas. To further evaluate the consistency of multi-view panorama generation, we leverage the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [59] metrics. PSNR and SSIM quantify the pixel-level differences and structural similarity between the generated views respectively.

5.1 Single-View Panorama Generation

Baselines We evaluate the performance of our proposed method by comparing it with the following baseline approaches for text-to-panorama generation:

- *Text2Light* [9]is a two-stage approach that first generates a low-resolution panorama based on the input text, and then expands it to ultra-high resolution.
- *PanFusion* [70]is a dual-branch text-to-panorama model that aims to mitigate the distortion caused by projecting perspective images onto a panoramic canvas while providing global layout guidance.

Since the panoramic images generated by MVDiffusion [48] do not include top and bottom viewpoints, they are not complete panoramas. Therefore, we do not compare our method with MVDiffusion.

 https://doi.org/10.52202/079017-0041

Figure 4: **Text to Panorama Comparison between TextLight, PanFusion, and Ours.**

Table 1: Quantitative Panorama Comparisons with Baseline Methods

Method	FID↓	IS↑	CS↑	Inference time(s)↓
Text2Light	76.50	3.60	27.48	50.4
PanFusion	**47.62**	**4.49**	28.73	27.6
Pano-SD(Ours)	48.52	3.30	**28.98**	**5.1**

Quantitative Results Table 1 presents the quantitative results. The FID, IS, and CS values of Text2Light are from the PanFusion [70]. We trained our proposed Pano-SD using the same dataset as PanFusion [70] and re-evaluated the performance of PanFusion. From the results, it can be observed that our method slightly outperforms others in terms of CS value and achieves a significantly lower FID value compared to Text2Light, which is close to PanFusion [70]. Moreover, due to the simplicity and efficiency of our model architecture, our method has a substantial advantage in inference time, which can significantly improve the efficiency of subsequent multi-view generation tasks.

Qualitative Results The qualitative comparison results are shown in Figure 4. We compare our method with the two models trained on their respective datasets. The panoramas generated by Text2Light exhibit poor left-right consistency. On the other hand, the panoramas generated by PanFusion suffer from blurriness at the bottom and top regions, which affects the overall integrity of the panoramas. In contrast, our model is capable of generating panoramas with clear bottom and top regions and better left-right continuity. However, due to the quality of the dataset, the image quality may be slightly inferior.

5.2 Multi-View Panorama Generation

To the best of our knowledge, there is no method for multi-view panorama generation, the existing SOTA method MVDream for perspective images cannot be directly applied to multi-view panorama generation. To verify the validity of our proposed spherical epipolar-aware attention module, we adapted MVDream to the panorama generation task as a comparative baseline.

Specifically, we remove the spherical epipolar-aware attention module from our method and load the pre-trained LoRA layers from Pano-SD. We then convert the 3D self-attention in the UNet to the form used in MVDream [43]. Additionally, we transform the camera pose matrix into camera

Table 2: User Study of Text to Multi-view Panoramas

Method	Image quality↑	Image-text consistency↑	Multi-view consistency↑
MVDream	3.6	3.7	3.8
PanFusion	3.8	4.0	3.0
DiffPano(Ours)	**4.0**	**4.2**	**4.3**

Table 3: Ablation Study on the Number of Sampling Points and Reference Frames

	FID↓	IS↑	CS↑	PSNR↑	SSIM↑	Inference time(s)↓
$K = 1$, $S = 10$	73.30	3.40	22.14	29.99	0.69	30.06
$K = 2$, $S = 10$	66.02	3.34	22.92	32.04	0.79	33.12
$K = 3$, $S = 10$	69.89	3.58	22.76	32.29	0.81	35.79
$K = 4$, $S = 6$	68.39	3.57	22.74	33.32	0.86	26.61
$K = 4$, $S = 8$	67.30	3.54	22.66	33.00	0.84	32.23
$K = 4$, $S = 10$	65.98	3.37	22.59	33.39	0.87	37.91
$K = 4$, $S = 12$	62.79	3.26	22.65	32.89	0.83	43.72

embeddings through a 2-layer MLP and add it as a residual to the time embeddings. The qualitative comparison of the experiment is shown in Fig 5. Under the same training iteration, DiffPano significantly outperforms MVDream [43] in terms of multi-view consistency. Our method can maintain consistency in the details of multi-view images, while MVDream can only achieve a certain level of similarity in the overall images. Even compared to MVDream with twice the number of training iterations, our method still performs better in terms of consistency.

User Study We collected 20 text prompts and recruited nearly 50 volunteers to evaluate text-to-multi-view panoramic image generation. Evaluation metrics of multi-view ERP panorama generation include the quality of multi-view panoramic images, multi-view consistency, and the consistency between the text and multi-view panoramas. Experimental results on Tab. 2 show that our method can generate multi-view panoramic images with better quality, higher text and image similarity, and more consistent multi-view images, compared with MVDream [43] and PanFusion [70]. See more qualitative results in Fig. 6 to Fig. 10.

5.3 Ablation Study

Figure 5: **Comparisons with MVDream.** DiffPano can generate more consistent multi-view panoramas. "MVDream×2" denotes MVDream is trained with twice iteration number relative to our method.

https://doi.org/10.52202/079017-0041

Table 4: Ablation Study of One-Stage vs Two-Stage Training

	FID↓	IS↑	LPIPS↓	PSNR↑	SSIM↑
One-stage	82.92	4.52	0.0454	31.64	0.74
Two-stage	74.08	3.13	0.0610	31.54	0.73

Number of Reference Views and Sample Points To explore the influence of varying numbers of reference frames and sampling points on model performance, we generate 5-frame multi-view panoramas simultaneously, and compute the FID, IS, and CS metrics for the first frame of each group to assess the quality of the generated panorama under different quantities of reference frames and sampling points. Furthermore, we set the same camera pose for the first and last frames of each generated panorama group, and calculate the PSNR and SSIM values between these two frames to evaluate the model's multi-view consistency. As shown in Table. 3, increasing the number of reference frames and sampling points can improve the quality of generated panoramas to a certain extent, but the changes in image diversity and consistency with text remain marginal. With an increasing number of reference frames and sampling points, the model's consistency exhibits improvement, however, when the number of sampling points becomes excessive (S=12), the multi-view consistency of the model diminishes.

One-Stage vs Two-Stage We conduct ablation experiments on one-stage and two-stage training. Table. 4 shows that the two-stage method will obtain the higher FID values. The IS of the two-stage training method is lower, and the diversity is reduced to a certain extent, which is slightly worse than the one-stage training. However, it should be noted that the images after one-stage training will have ghosting, but the two-stage will not.

6 Conclusion

We have proposed the panoramic video-text dataset and panorama generation framework with spherical epipolar-aware attention for text-to-single-view or multi-view panorama generation. Extensive experiments demonstrate that our method can achieve scalable, consistent, and diverse multi-view panoramas.

Limitation and Future Work Although our method demonstrates the ability to generate consistent multi-view panoramas under the same setting as the training phase, it is important to note that as the number of frames increases during inference, the model tends to hallucinate content.

Exploring the use of video diffusion models to improve the consistency of generated multi-view panoramas is a promising direction. Longer panoramic videos are expected to be realized based on the generated panoramas as conditions.

7 Acknowledgements

This work was partially supported by the National Key Research and Development Program of China (No. 2023YFF0905104) and National Natural Science Foundation of China (Nos. 61932003, 62076184 and 62473286).

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *https://openai.com/research/video-generation-models-as-world-simulators*, 2024.

[3] Weiwei Cai, Weicai Ye, Peng Ye, Tong He, and Tao Chen. Dynasurfgs: Dynamic surface reconstruction with planar-based gaussian splatting. *arXiv preprint arXiv:2408.13972*, 2024.

[4] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676. IEEE Computer Society, 2017.

[5] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *arXiv preprint arXiv:2406.06521*, 2024.

[6] Junyi Chen, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Where am i and what will i see: An auto-regressive model for spatial localization and view prediction. *arXiv preprint arXiv:2410.18962*, 2024.

[7] Junyi Chen, Weicai Ye, Yifan Wang, Danpeng Chen, Di Huang, Wanli Ouyang, Guofeng Zhang, Yu Qiao, and Tong He. Gigags: Scaling up planar-based 3d gaussians for large scene surface reconstruction. *arXiv preprint arXiv:2409.06685*, 2024.

[8] Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024.

[9] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven HDR panorama generation. *ACM Trans. Graph.*, 41(6):195:1–195:16, 2022.

[10] Zheng Chen, Yan-Pei Cao, Yuan-Chen Guo, Chen Wang, Ying Shan, and Song-Hai Zhang. Panogrf: Generalizable spherical radiance fields for wide-baseline panoramas. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[11] Zheng Chen, Chen Wang, Yuan-Chen Guo, and Song-Hai Zhang. Structnerf: Neural radiance fields for indoor scenes with structural hints. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15694–15705, 2023.

[12] Xiao Cui, Weicai Ye, Yifan Wang, Guofeng Zhang, Wengang Zhou, Tong He, and Houqiang Li. Streetsurfgs: Scalable urban street surface reconstruction with planar-based gaussian splatting. *arXiv preprint arXiv:2410.04354*, 2024.

[13] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *CoRR*, abs/2311.13141, 2023.

[14] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

[15] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

[16] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.

[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.

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[19] Chenxi Huang, Yuenan Hou, Weicai Ye, Di Huang, Xiaoshui Huang, Binbin Lin, Deng Cai, and Wanli Ouyang. Nerf-det++: Incorporating semantic cues and perspective-aware depth supervision for indoor multi-view 3d detection. *arXiv preprint arXiv:2402.14464*, 2024.

[20] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, and Lu Sheng. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. *CoRR*, abs/2312.06725, 2023.

 https://doi.org/10.52202/079017-0041

[21] Hai Li, Weicai Ye, Guofeng Zhang, Sanyuan Zhang, and Hujun Bao. Saliency guided subdivision for single-view mesh reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 1098–1107. IEEE, 2020.

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[23] Longwei Li, Huajian Huang, Sai-Kit Yeung, and Hui Cheng. Omnigs: Omnidirectional gaussian splatting for fast radiance field reconstruction using omnidirectional images. *CoRR*, abs/2404.03202, 2024.

[24] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, and Zhiwen Fan. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024.

[25] Zhen Li, Lingli Wang, Mofang Cheng, Cihui Pan, and Jiaqi Yang. Multi-view inverse rendering for large-scale real-world indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12499–12509. IEEE, 2023.

[26] Jiuming Liu, Guangming Wang, Weicai Ye, Chaokang Jiang, Jinru Han, Zhe Liu, Guofeng Zhang, Dalong Du, and Hesheng Wang. Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15109–15119, 2024.

[27] Jiuming Liu, Ruiji Yu, Yian Wang, Yu Zheng, Tianchen Deng, Weicai Ye, and Hesheng Wang. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*, 2024.

[28] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024.

[29] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023.

[30] Xiangyu Liu, Weicai Ye, Chaoran Tian, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Coxgraph: multi-robot collaborative, globally consistent, online dense reconstruction system. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8722–8728. IEEE, 2021.

[31] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *CoRR*, abs/2309.03453, 2023.

[32] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. *CoRR*, abs/2310.15008, 2023.

[33] 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.

[34] Yuhang Ming, Minyang Xu, Xingrui Yang, Weicai Ye, Weihan Wang, Yong Peng, Weichen Dai, and Wanzeng Kong. Viper: Visual incremental place recognition with adaptive mining and lifelong learning. *arXiv preprint arXiv:2407.21416*, 2024.

[35] Yuhang Ming, Weicai Ye, and Andrew Calway. idf-slam: End-to-end rgb-d slam with neural implicit mapping and deep feature tracking. *arXiv preprint arXiv:2209.07919*, 2022.

[36] 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.

[37] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022.

[39] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[40] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *CoRR*, abs/2310.17994, 2023.

[41] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[42] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *CoRR*, abs/2310.15110, 2023.

[43] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *CoRR*, abs/2308.16512, 2023.

[44] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Frédo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *NeurIPS*, pages 19313–19325, 2021.

[45] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025.

[46] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[47] Shengji Tang, Weicai Ye, Peng Ye, Weihao Lin, Yang Zhou, Tao Chen, and Wanli Ouyang. Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction. *arXiv preprint arXiv:2410.06245*, 2024.

[48] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023.

[49] Ziyu Tang, Weicai Ye, Yifan Wang, Di Huang, Hujun Bao, Tong He, and Guofeng Zhang. Nd-sdf: Learning normal deflection fields for high-fidelity indoor reconstruction. *arXiv preprint arXiv:2408.12598*, 2024.

[50] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

 https://doi.org/10.52202/079017-0041

[51] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16773–16783. IEEE, 2023.

[52] Chen Wang, Jiatao Gu, Xiaoxiao Long, Yuan Liu, and Lingjie Liu. GECO: generative image-to-3d within a second. *CoRR*, abs/2405.20327, 2024.

[53] Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. Perf: Panoramic neural radiance field from a single panorama. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024.

[54] 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.

[55] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *WACV*, pages 4921–4931. IEEE, 2024.

[56] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. *arXiv preprint arXiv:2308.14686*, 2023.

[57] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *CoRR*, abs/2312.02201, 2023.

[58] Yifan Wang, Di Huang, Weicai Ye, Guofeng Zhang, Wanli Ouyang, and Tong He. Neurodin: A two-stage framework for high-fidelity neural surface reconstruction. *arXiv preprint arXiv:2408.10178*, 2024.

[59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[60] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Ipo-ldm: Depth-aided 360-degree indoor rgb panorama outpainting via latent diffusion model. *arXiv preprint arXiv:2307.03177*, 2023.

[61] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024.

[62] 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.

[63] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7079–7088, 2024.

[64] Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, and Guofeng Zhang. IntrinsicNeRF: Learning Intrinsic Neural Radiance Fields for Editable Novel View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[65] Weicai Ye, Xinyue Lan, Shuo Chen, Yuhang Ming, Xingyuan Yu, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Pvo: Panoptic visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, June 2023.

[66] Weicai Ye, Xinyue Lan, Ge Su, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Hybrid tracker with pixel and instance for video panoptic segmentation. *arXiv preprint arXiv:2203.01217*, 2022.

[67] Weicai Ye, Hai Li, Tianxiang Zhang, Xiaowei Zhou, Hujun Bao, and Guofeng Zhang. Superplane: 3d plane detection and description from a single image. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 207–215. IEEE, 2021.

[68] Hailin Yu, Youji Feng, Weicai Ye, Mingxuan Jiang, Hujun Bao, and Guofeng Zhang. Improving feature-based visual localization by geometry-aided matching. *arXiv preprint arXiv:2211.08712*, 2022.

[69] Hailin Yu, Weicai Ye, Youji Feng, Hujun Bao, and Guofeng Zhang. Learning bipartite graph matching for robust visual localization. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 146–155. IEEE, 2020.

[70] 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*, 2024.

[71] Jinsong Zhang and Jean-François Lalonde. Learning high dynamic range from outdoor panoramas. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4519–4528, 2017.

[72] Tianxiang Zhang, Chong Bao, Hongjia Zhai, Jiazhen Xia, Weicai Ye, and Guofeng Zhang. Arcargo: Multi-device integrated cargo loading management system with augmented reality. In *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 341–348. IEEE, 2021.

[73] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, volume 12354 of *Lecture Notes in Computer Science*, pages 519–535. Springer, 2020.

[74] Haiyang Zhou, Xinhua Cheng, Wangbo Yu, Yonghong Tian, and Li Yuan. Holodreamer: Holistic 3d panoramic world generation from text descriptions. *arXiv preprint arXiv:2407.15187*, 2024.

[75] Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Point cloud matters: Rethinking the impact of different observation spaces on robot learning. *arXiv preprint arXiv:2402.02500*, 2024.

 https://doi.org/10.52202/079017-0041

A Spherical Epipolar Line Computation

Given the pixel coordinates of a point in the panorama of the target view i, the corresponding epipolar line in the panorama of the source view j can be calculated. First, according to Eq. (1), the camera coordinates $\mathbf{p}$ in the target view i are computed. Based on the relative pose $\left\{\mathbf{R}^{i\to j}, \mathbf{T}^{i\to j}\right\}$, $\mathbf{p}'$ in the source view j are calculated:

$$\mathbf{p}' = \mathbf{R}^{i\to j}\mathbf{p} + \mathbf{T}^{i\to j}. \tag{8}$$

Simultaneously, the coordinates $\mathbf{o}'$ of the camera origin $\mathbf{o} = (0, 0, 0)^T$ in the target view i projected onto the source view j are computed:

$$\mathbf{o}' = \mathbf{R}^{i\to j}\mathbf{o} + \mathbf{T}^{i\to j}. \tag{9}$$

Then, we need to find the plane L containing the three points $\mathbf{p}'$, $\mathbf{o}$, and $\mathbf{o}'$ in the camera coordinate system of the source view j. The intersection of this plane with the spherical surface is the desired epipolar line. The equation corresponding to plane L is:

$$AX + BY + CZ + D = 0. \tag{10}$$

Substituting the coordinates of the three points into the equation yields the coefficients A, B, C, and D:

$$\begin{aligned} A &= \frac{z_{\mathbf{o}'} \cdot y_{\mathbf{p}'} - z_{\mathbf{p}'} \cdot y_{\mathbf{o}'}}{x_{\mathbf{p}'} \cdot y_{\mathbf{o}'} - x_{\mathbf{o}'} \cdot y_{\mathbf{p}'}} \cdot C \\ B &= \frac{z_{\mathbf{o}'} \cdot x_{\mathbf{p}'} - z_{\mathbf{p}'} \cdot x_{\mathbf{o}'}}{y_{\mathbf{p}'} \cdot x_{\mathbf{o}'} - y_{\mathbf{o}'} \cdot x_{\mathbf{p}'}} \cdot C \\ D &= 0. \end{aligned} \tag{11}$$

To simplify the representation of the equation for plane L, we introduce new coefficients a_1 and a_2:

$$\begin{aligned} a_1 &= \frac{z_{\mathbf{o}'} \cdot y_{\mathbf{p}'} - z_{\mathbf{p}'} \cdot y_{\mathbf{o}'}}{x_{\mathbf{p}'} \cdot y_{\mathbf{o}'} - x_{\mathbf{o}'} \cdot y_{\mathbf{p}'}} \\ a_2 &= \frac{z_{\mathbf{o}'} \cdot x_{\mathbf{p}'} - z_{\mathbf{p}'} \cdot x_{\mathbf{o}'}}{y_{\mathbf{p}'} \cdot x_{\mathbf{o}'} - y_{\mathbf{o}'} \cdot x_{\mathbf{p}'}}, \end{aligned} \tag{12}$$

the equation of plane L can be simplified to:

$$a_1 X + a_2 Y + Z = 0. \tag{13}$$

According to Eq.1 and Eq.2, the epipolar line equation in the source view j can be obtained:

$$y_{pix} = H \cdot \left[\arctan\left(\frac{a_1 \sin\left(\frac{2\pi x_{pix}}{W}\right) - \cos\left(\frac{2\pi x_{pix}}{W}\right)}{a_2} \right) / \pi + 0.5 \right], \tag{14}$$

where x_{pix} and y_{pix} are the corresponding pixel coordinates.

B Experiment Details

Dataset Processing In single-view panorama generation, we select 100 scenes from HM3D [37] and render cube maps from random viewpoints within each scene, which are then stitched into panoramas through interpolation. However, due to the imperfect quality of the corresponding scene meshes, which may have missing parts, we filter out images with a high proportion of zero-depth values based on their corresponding depth maps, ultimately creating a dataset of 8,508 panoramas.

Table 5: Quantitative Perspective Images Comparisons with Baseline Methods

Method	FID↓	IS↑
MVDiffusion	188.16	1.82
PanFusion	213.19	**2.61**
Pano-SD(Ours)	**164.75**	1.96

In the multi-view panorama generation task, we render images based on camera trajectories across scenes. To accommodate the first stage of training, we select camera trajectories with minimal changes. In contrast, for the second stage, the dataset comprises panoramas with more significant camera displacements between consecutive frames. Each training dataset consists of multiple panoramas, their corresponding pose matrices, and text descriptions. To construct a multi-view panorama dataset with nearly identical image content, we select the first few frames from each camera trajectory, as the camera movement between these frames is minimal and they are captured within the same scene. For datasets where the image content changes, the construction needs to be performed across the entire camera trajectory. By projecting the panorama from the subsequent frame onto the viewpoint of the previous frame using spherical projection, we compare the pixel value differences between the two. If more than 40% of the pixel values differ, we consider new content to have been generated between the two frames. After filtering, there are 19,739 data sets with nearly identical image content and 18,704 data sets with differing image content.

Compare with Baseline Methods

- $Text2Light$ [9]: In the quantitative analysis, the FID, IS, and CS values are referenced from the results reported in PanFusion[70], which were obtained by testing on the same dataset after training. For the qualitative analysis, we directly employ their jointly trained model on both indoor and outdoor datasets to generate LDR (low dynamic range) panoramas based on text descriptions.

- $PanFusion$ [70]: To compare the quality of the generated panoramas, we also train the model using panoramas stitched from MP3D skybox images. We employ the same dataset split and text descriptions as in [70], which includes 9,820 panoramas for training and 1,092 for evaluation. In the qualitative analysis, we use our own dataset constructed from HM3D[37]. The panoramas in our dataset do not suffer from blurriness at the bottom and top regions, but the overall image quality may be slightly lower.

Compare with MVDream We adopt the method from MVDream[43] by formatting the $B \times N \times H \times W \times C$ tensor input to the self-attention module as $B \times NHW \times C$. This allows the model to simultaneously integrate features from all viewpoints during the self-attention computation, thereby improving the consistency among the generated multi-view panoramas. However, since MVDream is designed for multi-view image generation of objects, the camera pose matrices used in their method only contain rotation information without translation. In contrast, when generating camera embeddings, we use pose matrices that include both rotation and translation information, which increases the learning difficulty for the model. This may be one of the reasons why using this method does not yield particularly ideal results.

Transform Panoramas to Perspective Views We converted the generated panoramas into perspective images and conducted quantitative comparisons, shown in Tab. 5. Experiments show that our method achieves the lowest FID, while our method is higher than MVDiffusion in IS and slightly lower than PanFusion. It should be noted that MVDiffusion directly generates perspective images and then stitches them into panoramas. It is not in the ERP format and does not have the top and bottom parts. Our panorama generation speed is faster.

https://doi.org/10.52202/079017-0041

B.1 Applications

B.1.1 Text to Single-View Panorama

B.1.2 Text to Multi-View Panoramas

DiffPano can generate panoramic video frames with large camera pose spans and multi-view consistency based on diverse textual descriptions, thus achieving the effect of text-to-panoramic video, as shown in the Fig.11.

Figure 6: **Qualitative Comparisons of Text to Panoramic Videos.** Ours vs MVDream vs PanFusion.

Figure 7: **Qualitative Comparisons of Text to Panoramic Videos.** Ours vs MVDream vs PanFusion.

C Network Architecture and Training Details

Network Architecture Our LoRA-based single-view panorama generation model produces panoramas with a resolution of 512×1024. In the multi-view panorama generation approach, we generate continuous frames of panoramas with a resolution of 256×512. Generating multiple frames of 512×1024 resolution panoramas simultaneously would consume a significant amount of computational resources. Moreover, our experiments reveal that generating multi-frame high-resolution panoramas requires an exceptionally long training time to improve the quality of the generated images. The network architecture of our multi-view panorama generation model is shown in Table 6 and Table7.

Training Details We fine-tuned the Stable Diffusion v1.5 model using the LoRA method for single-view pano-based synthesis. The training was conducted on 6 A100 GPUs with 80GB memory for 100 epochs (approximately 6.5 hours), with a learning rate of 1e-4 and a batch size of 6. For

Table 6: Network architecture of DiffPano-1

	Layer	Output	Additional Inputs
(1)	Latent Map	$4 \times 32 \times 64$	
(2)	Conv.	$320 \times 32 \times 64$	
	CrossAttnDownBlock1		
(3)	ResBlock	$320 \times 32 \times 64$	Time emb.
(4)	AttnBlock	$320 \times 32 \times 64$	Prompt emb.
(5)	ResBlock	$320 \times 32 \times 64$	Time emb.
(6)	AttnBlock	$320 \times 32 \times 64$	Prompt emb.
(7)	DownSampler	$320 \times 16 \times 32$	
	CrossAttnDownBlock2		
(8)	ResBlock	$640 \times 16 \times 32$	Time emb.
(9)	AttnBlock	$640 \times 16 \times 32$	Prompt emb.
(10)	ResBlock	$640 \times 16 \times 32$	Time emb.
(11)	AttnBlock	$640 \times 16 \times 32$	Prompt emb.
(12)	DownSampler	$640 \times 8 \times 16$	
	CrossAttnDownBlock3		
(13)	ResBlock	$1280 \times 8 \times 16$	Time emb.
(14)	AttnBlock	$1280 \times 8 \times 16$	Prompt emb.
(15)	ResBlock	$1280 \times 8 \times 16$	Time emb.
(16)	AttnBlock	$1280 \times 8 \times 16$	Prompt emb.
(17)	DownSampler	$1280 \times 4 \times 8$	
	DownBlock		
(18)	ResBlock	$1280 \times 4 \times 8$	Time emb.
(19)	ResBlock	$1280 \times 4 \times 8$	Time emb.
	MidBlock		
(20)	ResBlock	$1280 \times 4 \times 8$	Time emb.
(21)	**EAModule**	$\mathbf{1280 \times 4 \times 8}$	
(22)	AttnBlock	$1280 \times 4 \times 8$	Prompt emb.
(23)	ResBlock	$1280 \times 4 \times 8$	Time emb.
	UpBlock		
(24)	ResBlock	$1280 \times 4 \times 8$	(19), Time emb.
(25)	ResBlock	$1280 \times 4 \times 8$	(18), Time emb.
(26)	ResBlock	$1280 \times 4 \times 8$	(17), Time emb.
(27)	**EAModule**	$\mathbf{1280 \times 4 \times 8}$	
(28)	UpSampler	$1280 \times 8 \times 16$	
	CrossAttnUpBlock1		
(29)	ResBlock	$1280 \times 8 \times 16$	(16), Time emb.
(30)	AttnBlock	$1280 \times 8 \times 16$	Prompt emb.
(31)	ResBlock	$1280 \times 8 \times 16$	(14), Time emb.
(32)	AttnBlock	$1280 \times 8 \times 16$	Prompt emb.
(33)	ResBlock	$1280 \times 8 \times 16$	(12), Time emb.
(34)	AttnBlock	$1280 \times 8 \times 16$	Prompt emb.
(35)	**EAModule**	$\mathbf{1280 \times 8 \times 16}$	
(36)	UpSampler	$1280 \times 16 \times 32$	

https://doi.org/10.52202/079017-0041

Figure 8: **Qualitative Comparisons of Text to Panoramic Videos.** Ours vs MVDream vs PanFusion.

Figure 9: **Qualitative Comparisons of Text to Panoramic Videos.** Ours vs MVDream vs PanFusion.

multi-view panoramas generation, we conducted training on 8 80G A100 GPUs with a batch size of 1 and a learning rate of 1e-5. Each GPU utilized approximately 50% of its memory. The two-stage training process involved 100 epochs for each stage, with a total training time of approximately 5 days.

D Societal Impact

Since our method can achieve scalable, consistent, and diverse multi-view panoramas, it has many potential applications, such as unlimited room roaming in VR, interior design preview, embodied intelligent robot exploration, etc.

https://doi.org/10.52202/079017-0041

Figure 10: **Qualitative Comparisons of Text to Panoramic Videos.** Ours vs MVDream vs PanFusion.

Figure 11: **Qualitative Results of Text to Panoramic Videos.** DiffPano can generate scalable and consistent panorama videos.

https://doi.org/10.52202/079017-0041

"A remodeled garage with a door open, displaying a cluttered space filled with various items on the walls. A well-lit hallway with a door and various furniture. "

"A living room with white walls and a white door. The room is furnished with red leather couches and a TV mounted on the wall. A spacious room with hardwood floors, a kitchen and dining area, and an old-fashioned touch."

"A cozy room with wooden floors and a sliding glass door. The room is well-lit with a sink and wooden furniture. A cozy bedroom with wooden floors anda large bed. The room is furnished with a rug."

"A living area with a high ceiling, a light fixture, and a kitchenette. The room is furnished with a dining table and chairs and a painting on the wall, featuring a living room and dining room combination, and a white couch."

Figure 12: **Qualitative Results of Text to Panoramic Videos.** DiffPano can generate scalable and consistent panorama videos.

"A home office with a large desk, bookshelves, hardwood floors and a comfortable chair. A hallway with hardwood floors, leading to a home office with a large desk and bookshelves."

"A bedroom with a bed, and a dresser, featuring a window with curtains and blinds and a TV on top of the dresser.A cozy bedroom with a bed, dresser, and TV, surrounded by a brown carpet."

"A cozy living room with a purple couch, a fireplace, and a rug. The ceiling has a light on it. A living room with a white staircase, wooden floors, a couch, and an open door that leads to another room."

"A small room with a wooden door and a bed, featuring a white wall with a picture. A spacious room with a carpeted floor, a hallway leading to a bedroom with wooden doors and a comfortable bed."

Figure 13: **Qualitative Results of Text to Panoramic Videos.** DiffPano can generate scalable and consistent panorama videos.

Table 7: Network architecture of DiffPano-2

	Layer	Output	Additional Inputs
		CrossAttnUpBlock2	
(37)	ResBlock	$640 \times 16 \times 32$	(11), Time emb.
(38)	AttnBlock	$640 \times 16 \times 32$	Prompt emb.
(39)	ResBlock	$640 \times 16 \times 32$	(9), Time emb.
(40)	AttnBlock	$640 \times 16 \times 32$	Prompt emb.
(41)	ResBlock	$640 \times 16 \times 32$	(7), Time emb.
(42)	AttnBlock	$640 \times 16 \times 32$	Prompt emb.
(43)	**EAModule**	$\mathbf{640 \times 16 \times 32}$	
(44)	UpSampler	$640 \times 32 \times 64$	
		CrossAttnUpBlock3	
(45)	ResBlock	$320 \times 32 \times 64$	(6), Time emb.
(46)	AttnBlock	$320 \times 32 \times 64$	Prompt emb.
(47)	ResBlock	$320 \times 32 \times 64$	(4), Time emb.
(48)	AttnBlock	$320 \times 32 \times 64$	Prompt emb.
(49)	ResBlock	$320 \times 32 \times 64$	(2), Time emb.
(50)	AttnBlock	$320 \times 32 \times 64$	Prompt emb.
(51)	**EAModule**	$\mathbf{320 \times 32 \times 64}$	
(52)	GroupNorm	$320 \times 32 \times 64$	
(53)	SiLU	$320 \times 32 \times 64$	
(54)	Conv.	$4 \times 32 \times 64$	

https://doi.org/10.52202/079017-0041

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims stated in the abstract and introduction accurately reflect the contributions and scope of our paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the information for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-0041

Answer: [Yes]

Justification: We will open source the code and data after the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe all the implementation details in the Experiment chapter.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our method requires a large amount of graphics cards and resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the graphic card used and training time in Experiment chapter.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that our research adheres to all aspects of the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-0041

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not generate any pretrained models or image generators that may be misused.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have ensured that all assets used in our research are properly credited to their original creators.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0041