MultiViewPano: A Generalist Approach to 360-degree Panorama Generation

Anonymous ICCV submission

Paper ID *****

Abstract

We propose MultiViewPano*, a training-free framework for 360° panorama generation from one or more arbitrarily posed input images. Our approach leverages a pretrained multi-view diffusion model to synthesize novel views along a virtual camera trajectory, which are then fused using a custom pose-aware stitcher. Unlike prior methods that require fixed field-of-view inputs or task-specific fine-tuning,* MultiViewPano *supports flexible camera poses and generalizes across diverse scenes. Our experiments demonstrate that our method achieves competitive visual fidelity compared to state-of-the-art approaches, while offering greater flexibility and simplicity.*

1. Introduction

360° panorama generation aims to create a coherent panorama from one or more input views. This task overlaps with Novel View Synthesis (NVS), which generates images from novel viewpoints. Imagining regions beyond the input image requires inferring scene geometry, occluded content, and preserving geometric consistency.

360° panoramic images play a crucial role in applications such as virtual reality, realistic material design, and scene reconstruction [7]. Unlike standard rectilinear images, panoramas capture a full horizontal Field of View (FoV) from a specific vantage point, providing rich geometric and contextual information about the surrounding environment. Existing panoramic datasets are small and lack scene diversity compared to standard vision datasets.

Existing techniques rely on fine-tuning diffusion models on these datasets, to learn a specific panorama projection format (e.g., cubemap or equirectangular). This training limits generalization. For example, CubeDiff produces high quality cubemap panoramas but requires inputs at exactly 90° FoV [17].

Outpainting approaches such as PanoDiffusion [29] can extend to multiple inputs, but assume all crops originate from a single camera location. This is an impractical requirement for real-world captures.

(A) Our method (single- & multi-view input)

Multiple arbitrarily posed inputs

Generated 360° panorama

(B) Existing methods

Fixed 90° FOV input

Generated 360° panorama

Figure 1. Existing methods focus on generating panoramas from a single image, assumed to have a 90° FoV. Our method extends to multiple arbitrarily posed inputs and FoV. In example A, the input images are from the ScanNet++ dataset [31]. Example B uses a perspective crop from the SUN360 dataset [30]. Both panoramas shown here were generated using our method.

Multi-view diffusion models, trained on video and view-consistent image datasets, naturally inherit spatial and temporal priors. These models learn to respect viewpoint coherence, parallax, and object permanence, properties that are critical for panorama stitching but absent in single-view models.

We present a training-free framework for 360° panorama

generation that leverages a pretrained multi-view diffusion model. Our method samples virtual camera trajectories to densely cover the scene from overlapping viewpoints, and fuses the generated views via a custom pose-aware stitching pipeline. This design avoids task-specific fine-tuning and supports single or multi-image inputs with arbitrary FoV and camera poses. Panoramas can be synthesized from any desired position within the scene, as illustrated in Figure 1. Our key contributions are:

1. We analyze the limitations of existing panorama-generation approaches and motivate a training-free alternative.
2. We introduce *MultiViewPano*, a novel pipeline that accepts one or more input views, generates additional frames, and stitches them into a 360° panorama, using our novel pose-aware stitcher.
3. We achieve competitive visual quality and robustness across diverse datasets, on par with state-of-the-art methods.

2. Related Work

2.1. Panorama Generation

Panorama synthesis can be broadly divided into *Text-to-Panorama* [5, 9, 27, 32] and *Image-to-Panorama* [8, 17, 29, 33]. Most existing methods fine-tune Stable Diffusion to predict an equirectangular panorama via progressive outpainting or multi-view generation, but this strategy often produces noticeable geometric distortions [33]. CubeDiff [17] alleviates these artifacts by generating the six perspective faces of a cubemap, better matching the inductive biases of the pretrained model. Likewise, MVDiffusion [26] generates eight perspective views using a correspondence-aware attention architecture. The views can then be stitched into a full 360° panorama, albeit with a restricted vertical FoV.

2.2. Novel View Synthesis

Stable Virtual Camera (SEVA) stands apart as a generalist diffusion model; it accepts any number of input images with unrestricted camera intrinsics and extrinsics, and directly samples novel views at user-specified poses. SEVA achieves state-of-the-art consistency and fidelity across diverse benchmarks [16]. Other notable contributions are GEN3C [23] and CAT3D [11]. In this paper we use SEVA, but this component of the proposed pipeline is interchangeable with any multi-view conditioned model.

2.3. Image stitching

Image stitching refers to the process of combining multiple overlapping images into a single seamless representation. Brown and Lowe's seminal work [3] established the foundation for modern image-stitching techniques by introducing an algorithm for the automatic alignment and blending of overlapping images. The proposed pipeline relies on Scale-Invariant Feature Transform (SIFT) [20] for matching keypoints between images and Random Sample Consensus (RANSAC) [10] for estimating homographies.

Once image positions on the stitching canvas are determined, subsequent processing involves gain compensation to equalize exposure and color discrepancies, followed by blending. A widely adopted technique is multi-band blending [4], in which each input image is decomposed into multiple spatial-frequency bands and then recombined using spatially varying weights.

Recent approaches have proposed neural image stitching methods [18, 24], which train end-to-end networks to align and blend image pairs. However, their pairwise training paradigm limits scalability to large image sets, making them unsuitable for constructing full equirectangular panoramas.

2.4. Neural Radiance methods

NeRF [21] introduces a neural radiance field approach that learns a continuous 3D scene representation via a neural network. From this representation, full equirectangular panoramas can be rendered directly. When camera poses are known, NeRF-style methods offer an alternative to classic stitching, acting as interpolators to create seamless equirectangular projections.

Building on classical light field approaches [13, 19] and recent advances in neural light field representations [1, 25], Neural Light Spheres (NLS) [6] propose a compact spherical representation that implicitly stitches and re-renders panoramic frame sequences. The NLS can re-render a sequence of frames and also generate higher FoV images of the scene. As the authors denote in their paper, this method constitutes an implicit image-stitching approach.

3. MultiViewPano

We propose *MultiViewPano*, a training-free and flexible pipeline for 360° panorama generation from one or more arbitrarily posed input images. As shown in Figure 2, the method comprises two stages: (1) frame synthesis using SEVA, and (2) a pose-aware image stitching module.

3.1. Stable virtual Camera

To generate a full 360° panorama with SEVA, we must first specify a *set of camera poses* that cover the 360 scene. We consider two simple pose sets:

1. **Pure panoramic rotation:** A ring of poses at a fixed spatial location, each differing only in latitudinal angle, spanning 360°. No translational offset is applied.
2. **Panorama circle:** Similar to 1. but camera poses are placed radially outward on a small circle.

These two pose sets yield complete 360° horizontal coverage, although with a restricted vertical FoV.

Figure 2. **MultiViewPano** overview: from a 90°–cropped input image, we generate a virtual camera trajectory around the scene center, synthesize multiple views with SEVA, and stitch them into a complete 360-degree panorama.

SEVA exhibits two key limitations when sampling poses:

1. **Multi-View Consistency.** When only a single input image is provided, the synthesized frames often exhibit visual discontinuities. While additional input views improve consistency, residual artifacts may persist.
2. **Pose Consistency.** If the requested poses are spatially distant from one another, results can look visually consistent, yet the poses themselves do not align on a single viewing sphere.

Consequently, to minimize multi-view and pose inconsistencies, we restrict our evaluation to those two pose sets.

3.2. Spherical Mapping

Each rectilinear frame is first converted to unit-sphere polar coordinates using the pinhole camera model centered at the origin.

$$\mathbf{d}_{\text{cam}} = \frac{K^{-1}\,[x,\ y,\ 1]^{\mathsf{T}}}{\|K^{-1}\,[x,\ y,\ 1]^{\mathsf{T}}\|_2}, \qquad \mathbf{d}_{\text{world}} = R\,\mathbf{d}_{\text{cam}}, \quad (1)$$

where K is the intrinsics matrix containing the focal lengths and principal point and R is the camera rotation from frame metadata. Longitude θ and latitude φ are then

$$\theta = \operatorname{atan2}(d_x, d_z), \qquad \varphi = \arcsin(d_y), \quad (2)$$

which map to equirectangular pixel coordinates

$$u = (\pi + \theta)/(2\pi)\,W,\ v = (\varphi + \pi/2)/\pi\,H \quad (3)$$

Forward bilinear splatting accumulates each source pixel into its four neighbouring (u, v) bins.

3.3. Camera-Pose based Stitcher

Given the mapped frames, our camera-pose based stitcher proceeds in two steps:

Seam finding. For every overlapping frame pair we compute an L_2 colour cost in the overlap and extract the minimum cost *vertical seam* per connected component via dynamic programming.

Feather blending. Seams partition the canvas into disjoint regions. For each seam we linearly blend a symmetric $\pm w$-pixel band,

$$I(u, v) = \alpha\, I_i(u, v) + (1 - \alpha)\, I_j(u, v) \quad (4)$$

where,

$$\alpha = \tfrac{1}{2} + \tfrac{d_i - d_j}{2w} \quad (5)$$

after a simple per-channel gain match in the overlaps. This pose-aware blending removes visible discontinuities while remaining fast and robust to parallax.

4. Experiments

In this section, we present our experimental findings.

4.1. Experimental setup

Datasets: We evaluate on two standard panorama benchmarks, Laval Indoor [12] and SUN360 [30]. For each, we randomly select 1000 equirectangular images as our test size.

Evaluation metrics. We use Fréchet Inception Distance (FID) [14] and Kernel Inception Distance (KID) [2] to evaluate generation quality. FID measures the distance between feature distributions of generated and real images. KID computes the squared MMD using a polynomial kernel.

Baselines: We compare against CubeDiff, PanoDiffusion, and OmniDreamer. Following CubeDiff's protocol, we extract 10 random 90° FOV perspective crops from both generated and ground-truth panoramas, evaluating on 1000 images per dataset to ensure a fair comparison.

Stitching algorithms: We first run two off-the-shelf stitchers (Hugin and OpenCV's `Stitcher` [15], [22], [28]), which do not take camera poses, then our custom, pose-based stitcher.

4.2. Ablation Study

Full results are provided in the supplementary material. In brief, we varied SEVA's key hyperparameters and found the following optimal configuration: using our custom stitcher, classifier-free guidance weight of 5, camera-scale of 0.1, with a *panorama circle* trajectory on Laval Indoor and a pure panoramic rotation on SUN360.

4.3. Quantitative evaluation

In Table 1 the results of our quantitative evaluation on Laval indoor and SUN360 bechmarks can be seen. Our method outperforms previous outpainting approaches and has comparable results with the current SOTA CubeDiff, despite being training-free.

Figure 3. Qualitative comparison on two example scenes. For each scene, we show the full generated panorama (the central 90° region denotes the input image) followed by two 90° rectilinear views of synthesized areas. Results for baseline methods are taken directly from the CubeDiff paper [17].

4.4. Qualitative evaluation

In Figure 3 the qualitative results of our method in comparison to the other methods can be seen. Similar to the quantitative, the qualitative results surpass PanoDiffusion and OmniDreamer and are on par with CubeDiff.

	LAVAL Indoor		**SUN360**	
	FID ↓	KID ($\times 10^2$)↓	FID ↓	KID ($\times 10^2$)↓
OmniDreamer	71.0	5.17	92.3	8.89
PanoDiffusion	58.6	4.08	52.9	3.51
CubeDiff	11.7	0.47	27.4	1.35
MultiViewPano	17.5	0.83	28.2	1.25

Table 1. Comparison of FID and KID on Laval Indoor and SUN360 in the single image to panorama setting. Baseline results are taken directly from the CubeDiff paper [17].

5. Conclusion

We have presented **MultiViewPano**, a training-free framework that combines SEVA view synthesis with a novel camera-pose based stitcher to produce 360° panoramas from one or more arbitrarily posed inputs. By sampling simple pose sets and mapping each rectilinear frame onto the unit sphere, we generate consistent overlapping views. Our custom lightweight stitcher then finds optimal seams and applies feathered blending to remove discontinuities.

Future work will focus on:

1. Exploring full coverage camera trajectories to extend vertical FoV and mitigate the inconsistencies that arise with it.
2. Developing a learning based approach for stitching, inspired by NeRF.
3. Establishing a benchmark for evaluating multi-view to panorama generation.

Overall, this work demonstrates how pretrained multi-view diffusion models with a camera-aware stitcher, can form a general, flexible foundation for panorama generation.

References

[1] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19819–19829, 2022. 2

[2] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations (ICLR)*, 2018. Proposed Kernel Inception Distance (KID). 3

[3] Matthew Brown and David G. Lowe. Automatic Panoramic Image Stitching using Invariant Features. *International Journal of Computer Vision*, 74(1):59–73, 2007. 2

[4] Peter J. Burt and Edward H. Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics*, 2(4):217–236, 1983. 2

[5] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2Light: Zero-Shot Text-Driven HDR Panorama Generation, 2023. arXiv:2209.09898 [cs]. 2

[6] Ilya Chugunov, Amogh Joshi, Kiran Murthy, Francois Bleibel, and Felix Heide. Neural Light Spheres for Implicit Image Stitching and View Synthesis. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[7] Lucas da Silveira, Carlos Esteves, Stefan Borko, Thomas Funkhouser, and Paul Guerrero. 3d scene geometry estimation from 360° imagery: A survey. *arXiv preprint arXiv:2401.09252*, 2024. 1

[8] 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)*, page 475–485. IEEE, 2022. 2

[9] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 Degree Panoramic Image Generation based on Diffusion Models, 2023. 2

[10] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[11] 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, 2024. 2

[12] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xing Shen, Emanuele Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. In *ACM Transactions on Graphics (TOG)*, page 176, 2017. 3

[13] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1996)*, pages 43–54. ACM, 1996. 2

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017. Introduced Fréchet Inception Distance (FID). 3

[15] Hugin Contributors. *Hugin - Panorama photo stitcher*. SourceForge, 2024. https://hugin.sourceforge.io/. 3

[16] Jensen, Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable Virtual Camera: Generative View Synthesis with Diffusion Models, 2025. 2

[17] Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. CUBEDIFF: REPURPOSING DIFFUSION-BASED IMAGE MODELS FOR PANORAMA GENERATION. 2025. 1, 2, 4

[18] Minsu Kim, Jaewon Lee, Byeonghun Lee, Sunghoon Im, and Kyong Hwan Jin. Implicit Neural Image Stitching, 2024. arXiv:2309.01409 [cs]. 2

[19] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1996)*, pages 31–42. ACM, 1996. 2

[20] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[21] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, 2020. arXiv:2003.08934 [cs]. 2

[22] OpenCV Contributors. *OpenCV: Images stitching*, 2025. Version 4.13, Open Source Computer Vision Library. 3

[23] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control, 2025. 2

[24] Dae-Young Song, Geonsoo Lee, HeeKyung Lee, Gi-Mun Um, and Donghyeon Cho. Weakly-Supervised Stitching Network for Real-World Panoramic Image Generation, 2022. 2

[25] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8269–8279, 2022. 2

[26] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. MVDiffusion++: A Dense High-resolution Multi-view Diffusion Model for Single or Sparse-view 3D Object Reconstruction, 2024. 2

[27] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models, 2023. 2

[28] Lukas Alexander Weber et al. Automatic stitching of fragmented construction plans of hydraulic structures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2025. Used for development of OpenStitching package, GitHub repository: https://github.com/OpenStitching/stitching. 3

[29] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. PANODIFFUSION: 360-DEGREE PANORAMA OUT- PAINTING VIA DIFFUSION. 2024. 1, 2

[30] Jia Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 3

[31] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1

[32] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. (PanFusion):Taming Stable Diffusion for Text to 360° Panorama Image Generation, 2024. 2

[33] Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jianfang Hu, and Wei-Shi Zheng. Panorama Generation From NFoV Image Done Right. 2025. 2