AnyStitch: Generative Large Parallax Image Stitching

ZIXUAN CHEN, The Chinese University of Hong Kong, China
LINGEN LI, The Chinese University of Hong Kong, China
JUNHAO ZHUANG, Tsinghua University, China
RUIKANG LI, The Chinese University of Hong Kong, China
TIANFAN XUE*, The Chinese University of Hong Kong, Shanghai AI Laboratory, CPII under InnoHK, China

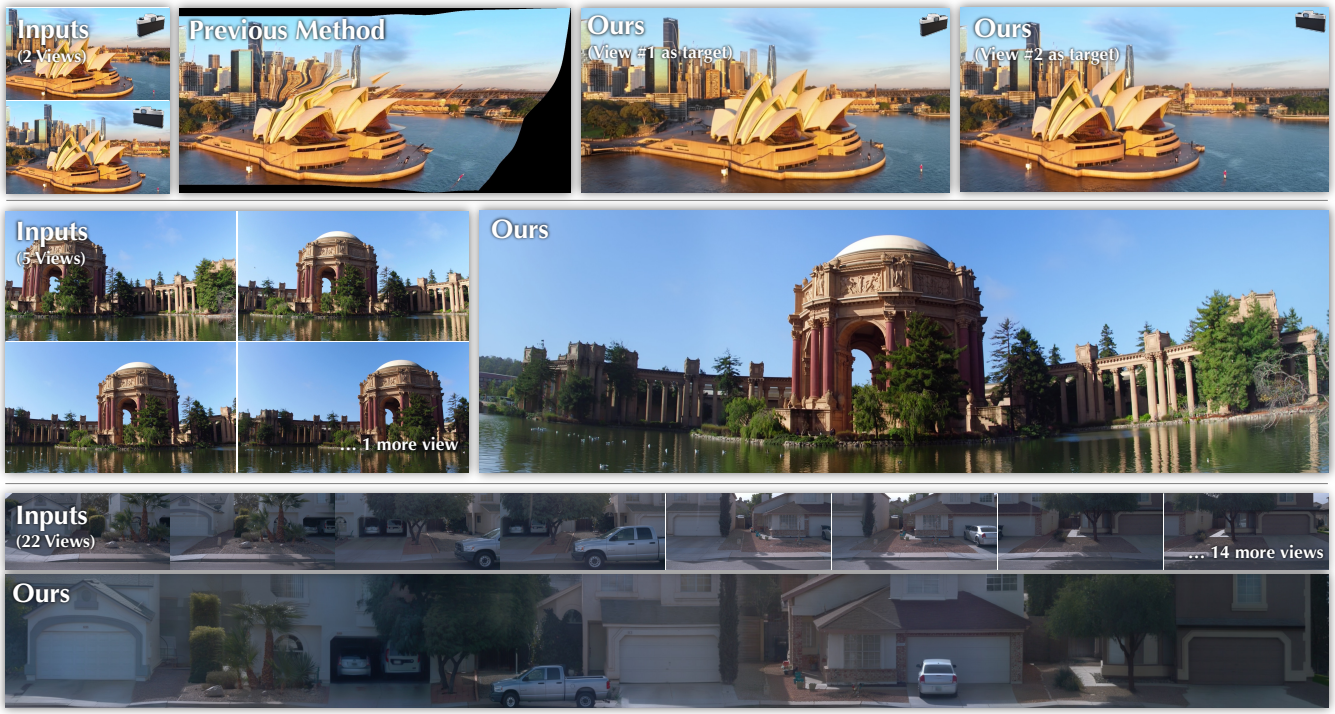


Fig. 1. Panoramas synthesized by AnyStitch across three distinct scenes with varying projection types. **Top**: Drone aerial perspective (rectilinear projection). **Middle**: Casual walking-tour captures [Knapitsch et al. 2017] involving both camera translation and pure rotation (cylindrical projection). **Bottom**: Street-view sweeps [Sun et al. 2020] featuring pure camera translation (parallel projection). AnyStitch seamlessly generates panoramas at flexible resolutions from both two-view and multi-view captures.

Wide field-of-view panorama stitching is important for computational photography and immersive visual applications, but often faces some challenges like severe geometric disparities. The typical "warp-and-blend" stitching technique only works with limited parallax and fails on large-parallax scenes, due to its inherent restriction to 2D geometric transformations that cannot resolve underlying 3D structures. To break this limitation, we present AnyStitch, a generative framework that can merge multiple reference images with large parallax into a seamless panorama. Unlike previous methods, we formulate the stitching problem as a reference-based image generation process, which naturally avoids stitching artifacts. Specifically, we propose a Hierarchical Projected Rotary Positional Encoding (HPRoPE) based on a 3D feed-forward model to establish accurate geometric relationships between the inputs and the panoramic canvas while preserving fidelity. Given that the resolution of the generated image varies with the number of inputs, we design a spatial autoregressive generation strategy to generate visually coherent panoramas at flexible resolutions. Finally, to enable the supervised fine-tuning of the proposed model, we construct a paired dataset utilizing 3D Gaussian Splatting (3DGS) scenes. Experiments demonstrate that AnyStitch can seamlessly stitch a variable number of inputs (like 22 views in teaser) with large camera motion, support flexible control of target

*Corresponding author.

Authors' Contact Information: Zixuan Chen, Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, China, zxchen3350@gmail.com; Lingen Li, Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, China, lgli@link.cuhk.edu.hk; Junhao Zhuang, Tsinghua University, Beijing, China, zhuangjh23@mails.tsinghua.edu.cn; Ruikang Li, Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, China, 1155231343@link.cuhk.edu.hk; Tianfan Xue, Multimedia Laboratory, The Chinese University of Hong Kong, Shanghai AI Laboratory, CPII under InnoHK, Hong Kong, China, tfxue@ie.cuhk.edu.hk.

This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 1557-7368/2026/12-ART264
https://doi.org/10.1145/3842587

view and projection type, and handle simple dynamic scenes. Project page: https://cecret3350.github.io/AnyStitch/.

CCS Concepts: • **Computing methodologies → Computational photography**.

Additional Key Words and Phrases: Image stitching, computational photography, generative model

ACM Reference Format:
Zixuan Chen, Lingen Li, Junhao Zhuang, Ruikang Li, and Tianfan Xue. 2026. AnyStitch: Generative Large Parallax Image Stitching. *ACM Trans. Graph.* 45, 6, Article 264 (December 2026), 13 pages. https://doi.org/10.1145/3842587

1 Introduction

How to capture an image with a wide field-of-view (FOV) that exceeds the camera capability has always been a fundamental problem in computer vision. To achieve this, many image stitching techniques are proposed to merge multiple inputs into a large seamless panorama. Typically, these methods follow a two-stage paradigm. They first warp the input images using homographies [Brown and Lowe 2007; Lee and Sim 2020; Li et al. 2019, 2017; Liao and Li 2019a; Zaragoza et al. 2013] or optical flow [Jia et al. 2023; Jin et al. 2025], and then fuse the aligned images to form the final panorama. Despite all recent advances, these 2D warping paradigms are intrinsically limited to scenes with small parallax. Once large parallax occurs, even when there are still substantial overlapping regions between the input views, previous methods fail severely, as shown in the top scene of Fig. 1.

Due to this limitation, users should keep their cameras as stationary as possible, which restricts the flexibility of capture and the diversity of applicable scenarios. Imagine a casual "walk-and-shoot" scenario around a large architectural landmark or along a long street, as shown in Fig. 1. The continuous camera translation introduces massive parallax across multiple views. In such cases, previous methods are unable to disentangle the overlapping 3D structures. Therefore, in this work, we study the following question: *can we seamlessly merge multiple inputs with a large parallax to create a coherent visual representation?*

Inspired by the remarkable capability of state-of-the-art image generation models to synthesize coherent visual content, we answer this question from a generative perspective. Unlike traditional image stitching methods, we formulate the stitching problem as a reference-based image generation and propose a novel framework, **AnyStitch**. In our design, the final stitched image is generated by taking multiple large-parallax captures as references. Because the output is generated rather than simply blended, this approach inherently circumvents the persistent seams and ghosting artifacts that plague traditional methods.

However, directly applying image generation models to image stitching by taking multiple inputs as conditions is far from a trivial task due to the following reasons. First, existing generative models typically assume that the reference latent shares the same spatial positional encoding as the noisy latent. By contrast, in image stitching, the output resolution is substantially larger than the input, often scaling up by 5 to 10 times with a dramatically altered aspect ratio (*e.g.*, from a standard 2:1 input to a 10:1 panoramic output, as shown in Fig. 1). As a result, the input condition occupies only a small, localized region within the expansive final canvas. Relying on standard positional encodings forces the model to depend exclusively on attention mechanisms to infer spatial correspondences. This compromises geometric accuracy and leads to noticeable stitching errors, as shown in Fig. 11. In addition, it is computationally infeasible for generation models to generate the entire stitched image in a single pass [Li et al. 2026]. Second, fine-tuning a generative model is inherently a supervised process, whereas previous learning-based stitching methods depend on unsupervised training. There is a critical absence of a dataset containing large-parallax inputs paired with geometrically consistent ground truth.

To address these challenges, we design AnyStitch, a novel stitching model combines the geometric perception with generative model. First, to deal with large parallax, we move beyond the traditional 2D warping paradigm to 3D modeling. While 2D-based methods typically collapse under large parallax due to their restrictive assumption of a planar scene, a 3D feed-forward model [Lin et al. 2025; Wang et al. 2025, 2024, 2026] enables more robust 3D geometric grounding [Jia et al. 2025; Zhu et al. 2025]. This shift allows the model to perceive the underlying 3D structure of the scene, effectively disentangling overlapping foreground and background elements that are otherwise inseparable in 2D pixel space. Specifically, we utilize 3D model to estimate depth and camera pose of each input. Then, we can flexibly define a target view and projection type. With these geometric priors, we are able to determine the resolution of the final canvas and establish the precise spatial mapping of each input onto it.

To ensure this geometric information is properly used in a generative stitching model, we further introduce specialized mechanisms to generate the final panorama. A simple solution is to perform a hard projection of inputs in the pixel space, but it may suffer from severe structure distortions and information loss when depth estimation are inaccurate. Instead, we formulate the spatial mapping as a geometric prior. Specifically, we propose a Hierarchical Projected Rotary Positional Encoding (HPRoPE) that unifies the positional encodings of the input references and the generated image. Within HPRoPE, we allocate varying projection precisions across different attention heads to simultaneously ensure strict geometric accuracy and preserve content fidelity. As shown in Fig. 11 and Fig. 12, this prevents the structural distortions inherent to pixel-space projection. Second, to handle variable-resolution outputs and circumvent computational constraints, we introduce a patch-wise spatial autoregressive generation strategy. We generate the stitched panorama patch-by-patch using a sliding window, utilizing previously generated tokens as memory context to ensure consistency. Additionally, a relative positional encoding is integrated during this autoregressive process to prevent extrapolation artifacts. Finally, we bridge the data gap by constructing a synthetic dataset called AnyStitchSyn, utilizing 3D Gaussian Splatting (3DGS) scenes, rendering large-parallax inputs paired with precise ground truths to enable robust supervised fine-tuning.

Fig. 1 demonstrates that AnyStitch can seamlessly integrate a variable number of large-parallax captures (like 20+ views shown in bottom) while offering flexible, user-defined control over target views and projection types, ultimately generating a visually coherent panorama at flexible resolutions.

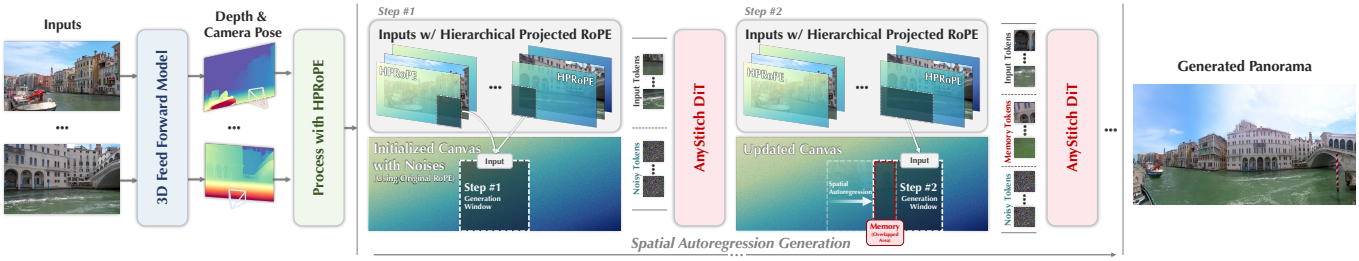


Fig. 2. The pipeline overview of AnyStitch. We first extract depths and camera poses via a 3D feed-forward model to determine the final size of panoramic canvas and compute the Hierarchical Projected RoPE (HPRoPE), establishing precise geometric conditioning. Subsequently, a Diffusion Transformer (DiT) synthesizes the panorama using a spatial autoregressive generation strategy. By processing the canvas patch-by-patch and incorporating previously decoded regions as memory tokens, our framework seamlessly merges large-parallax inputs into a visually coherent panorama at flexible resolutions.

2 Related Work

Image stitching. Traditional image stitching frameworks rely on geometric features to establish correspondences, evolving from basic global homography alignments [Brown and Lowe 2007] to more flexible spatially-varying warps [Lee and Sim 2020; Li et al. 2019, 2017; Liao and Li 2019a; Zaragoza et al. 2013]. To further mitigate the unnatural distortions introduced by these flexible deformations, subsequent approaches incorporated structure-preserving constraints, such as global similarity priors [Chen and Chuang 2016; Du et al. 2022; Liao and Li 2019b] and line-point consistency [Gao et al. 2011; Jia et al. 2021], to maintain the integrity of salient geometric structures.

With the success of Convolutional Neural Networks (CNNs) in other computer vision tasks, learning-based methods have emerged to extract high-level semantic features via neural networks. Recent learning-based paradigms have progressed from directly regressing global projective parameters [Nie et al. 2021] to predicting sophisticated local mesh deformations [Nie et al. 2023, 2026] and dense bidirectional optical flow fields [Jia et al. 2023; Jin et al. 2025]. More recently, some methods have introduced generative models for image stitching. SRStitcher [Xie et al. 2024b] uses a pre-trained generative model to refine coarse fusion and rectangling results obtained from 2D-aligned inputs. GPIS [Tuli et al. 2025] further fine-tunes a generative model for each scene using 2D pre-aligned inputs and generates the panorama patch by patch with global canvas coordinates. Nevertheless, these methods either manipulate images through 2D pixel-space warping or rely on 2D pre-alignment, limiting their robustness under large parallax.

Beyond 2D warping, geometry-based methods construct panoramas through 3D reconstruction. Instant 3D Photography [Hedman and Kopf 2018] uses captured depth to align RGB-D frames in 3D and fuse them into a panorama that supports parallax-aware rendering. More recently, feed-forward 3D reconstruction models (*e.g.*, DUSt3R [Wang et al. 2024], VGGT [Wang et al. 2025], Depth Anything 3 [Lin et al. 2025], and Pi3 [Wang et al. 2026]) have demonstrated efficient and generalizable reconstruction even from sparse image inputs. LiftProj [Jia et al. 2025] and PIS3R [Zhu et al. 2025] build on these models [Wang et al. 2025, 2024] to lift the inputs into 3D before projecting them onto the target view. These approaches improve geometric alignment, but their pixel-space rendering remains sensitive to errors in the estimated geometry.

Generative novel view synthesis. Generative novel-view synthesis combines learned priors with geometric guidance to synthesize content beyond the observed views. Context-aware layered depth inpainting [Shih et al. 2020] synthesizes color and depth in disoccluded regions for novel-view rendering. GenWarp [Seo et al. 2024] augments generative novel-view synthesis with geometric warping signals. ViewCrafter [Yu et al. 2024] conditions video diffusion on point-based 3D cues. GEN3C [Ren et al. 2025] renders a 3D cache along a target camera trajectory to guide video generation. These methods primarily target novel-view image or video synthesis, whereas AnyStitch integrates multiple casually captured inputs into a single panoramic canvas.

Geometric positional encoding. Geometric positional encodings incorporate camera and projection geometry into token positions, enabling attention to reason about spatial correspondences across views. PRoPE [Li et al. 2025] encodes relative projective transformations between complete camera frustums, incorporating both camera intrinsics and extrinsics into multi-view attention. RayRoPE [Wu et al. 2026] associates each image patch with a camera ray and predicts a point along the ray, whose query-frame projection defines a geometry-adaptive positional encoding for multi-view attention. PE-Field [Bai et al. 2026] and MosaicMem [Yu et al. 2026] further introduce geometry-aware positional encodings into generative models. They both establish source-to-target correspondence without explicitly warping the source RGB content, instead assigning RoPE to source tokens according to their reprojected coordinates in the target view. Unlike these methods, HPRoPE is specifically tailored to multi-image stitching with a dual-precision projection design, where separate attention heads assign input tokens to panorama coordinates using dense-depth and median-depth projections, respectively.

3 Methodology

Given a variable number of casually captured input images $\{I_i\}_{i=1}^{N}$, a user-specified target view V_{target} represented by intrinsic and extrinsic matrix $(\mathbf{K}_{tgt}, \mathbf{T}_{tgt})$, and a user-specified projection type $\mathcal{P}$ (*e.g.*, rectilinear, cylindrical, and equirectangular projection), our

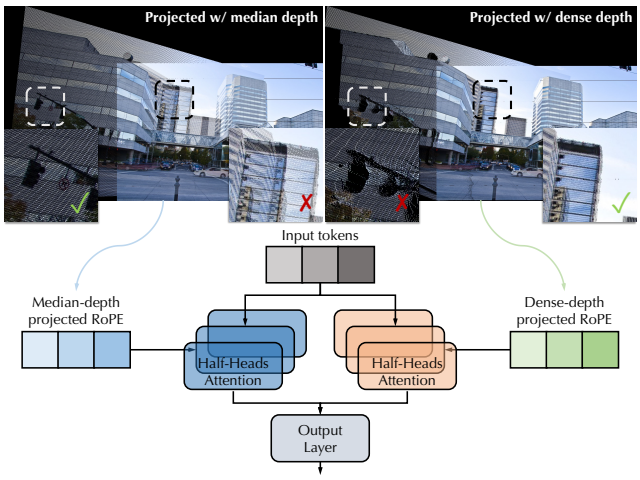


Fig. 3. Illustration of HPRoPE. Dense-depth projection captures fine-grained geometry but can lose information around inaccurate depth boundaries, while median-depth projection provides coarse yet stable alignment that better preserves content fidelity. HPRoPE evenly assigns the two projected RoPE variants to separate halves of the attention heads and the output layer is trained to smartly extract and combine their outputs.

goal is to stitch these inputs into a seamless panorama I_{out}. Previous methods typically warp and blend the inputs, making them highly sensitive to large parallax. In contrast, to address the challenging alignment of input images under motion parallax, we use a feed-forward 3D reconstruction model to estimate the camera pose and depth map for each input image. Using the estimated geometry, we map each input patch to its corresponding location on the output canvas and encode this correspondence by modifying its RoPE, without explicitly warping the input content. Then we finetune the pretrained generative model with these virtually warped patches as condition, to ensure they can be naturally stitched together.

In addition, we also devise a spatial autoregressive generation strategy including memory mechanisms and relative positional encoding to enable large panorama generation with a variable number of inputs. The details of each module are described below.

3.1 Hierarchical Projected RoPE

Standard diffusion transformers [Peebles and Xie 2023] rely on 2D RoPE to model spatial relationships. This standard RoPE fundamentally assumes that the reference conditions and the generated latent represent the same scene layout and share an identical, uniform spatial grid. However, image stitching fundamentally violates this assumption. In large-parallax stitching scenarios, casually captured inputs $\{I_i\}_{i=1}^{N}$ arbitrarily occupy localized, unaligned regions within a substantially larger panoramic canvas. Standard 2D RoPE fails to establish the correct spatial coordinates for input references, leading to severe spatial misalignment during the generative process.

To establish accurate spatial correspondences, we leverage a 3D feed-forward model. By explicitly modeling 3D geometry through depth and camera poses, it naturally overcomes the inherent limitations of traditional 2D transformations (*e.g.*, homography or optical flow) in large-parallax scenes, providing accurate geometric priors [Jia et al. 2025; Zhu et al. 2025].

Specifically, we first employ Depth Anything 3 (DA3) [Lin et al. 2025] to estimate the intrinsic matrix $\mathbf{K}_i$, extrinsic matrix $\mathbf{T}_i$ and a corresponding dense depth map D_i. With these estimations, we can project each original spatial RoPE (u_i, v_i) from the input image plane into a unified 3D space, and then project to the panorama canvas via the given target view V_{target} and the projection type $\mathcal{P}$ [Yu et al. 2026]:

$$(u_i', v_i') = \Pi_{\mathcal{P}} \left(\mathbf{K}_{tgt} \mathbf{T}_{tgt} \mathbf{T}_i^{-1} \mathbf{K}_i^{-1} (u_i, v_i, D_i) \right), \tag{1}$$

where $\Pi_{\mathcal{P}}(\cdot)$ denotes the project operation of $\mathcal{P}$.

While Eq. 1 provides a rigorous geometric alignment, it heavily relies on the accuracy of the estimated dense depth map D_i. In casually captured large-parallax scenes, depth estimation inevitably suffers from local errors, particularly around occlusion boundaries and fine structures. Strictly enforcing such fine-grained but noisy 3D priors across all attention layers can overly constrain the generative process, occasionally producing local distortions or spatial offsets in the generated panorama.

To mitigate this issue, we propose Hierarchical Projected RoPE (HPRoPE). As shown in Fig. 3, instead of relying exclusively on the dense depth D_i, we introduce a robust auxiliary planar projection driven by the scalar median depth of the input view, denoted as $\bar{D}_i = \text{median}(D_i)$. This alternative spatial mapping is formulated as:

$$(\bar{u}_i', \bar{v}_i') = \Pi_{\mathcal{P}} \left(\mathbf{K}_{tgt} \mathbf{T}_{tgt} \mathbf{T}_i^{-1} \mathbf{K}_i^{-1} (u_i, v_i, \bar{D}_i) \right). \tag{2}$$

This dual-branch design offers obvious advantages, as shown in Fig. 3. The dense-depth projection (Eq. 1) provides an accurate geometric layout, while the median-depth projection (Eq. 2) treats the scene as a simple 3D plane, providing a stable baseline that is unaffected by local depth noise.

Finally, we downscale the coordinates and obtain $(u_{l,i}, v_{l,i}) = (u_i'/s, v_i'/s)$ and $(\bar{u}_{l,i}, \bar{v}_{l,i}) = (\bar{u}_i'/s, \bar{v}_i'/s)$ for latent space, where s denotes the downsampling factor from image space to latent space. By integrating the input index t and the text index c, we construct two sets of RoPE coordinates for each latent token: the dense-depth coordinates $(t_i, u_{l,i}, v_{l,i}, c_i)$ and the planar median-depth coordinates $(t_i, \bar{u}_{l,i}, \bar{v}_{l,i}, c_i)$. To empower the self-attention layer to process these dual geometric representations concurrently, we introduce a head-wise coordinate assignment mechanism. Given an attention layer with H heads, we partition them into two disjoint subsets: $\mathcal{H}_{dense}$ and $\mathcal{H}_{median}$. The heads are divided evenly between the two subsets, such that $|\mathcal{H}_{dense}| = |\mathcal{H}_{median}| = H/2$. For a specific head h, the applied rotary positional encoding is explicitly conditioned on its designated subset:

$$\text{RoPE}_h(\mathbf{x}_{h,i}) = \begin{cases} \mathcal{F}_{rot}(\mathbf{x}_{h,i} \mid t_i, u_{l,i}, v_{l,i}, c_i), & \text{if } h \in \mathcal{H}_{dense} \\ \mathcal{F}_{rot}(\mathbf{x}_{h,i} \mid t_i, \bar{u}_{l,i}, \bar{v}_{l,i}, c_i), & \text{if } h \in \mathcal{H}_{median} \end{cases} \tag{3}$$

where $\mathbf{x}_{h,i}$ represents the query or key feature of token i in the h-th head, and $\mathcal{F}_{rot}(\cdot)$ denotes the standard rotary transformation function utilizing the given base coordinates. By assigning these complementary coordinates to separate attention heads, HPRoPE enables each attention layer to jointly exploit fine-grained geometric alignment and a robust global layout. Moreover, it introduces no

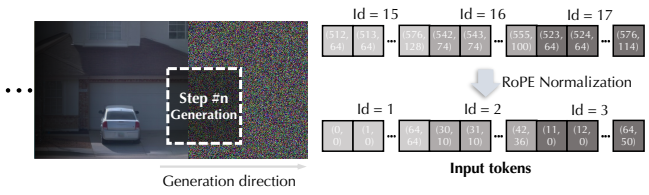


Fig. 4. Relative RoPE in each step of spatial autoregressive generation.

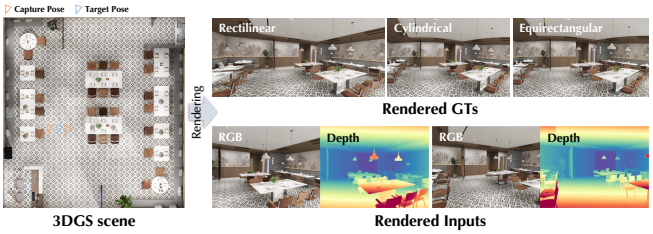


Fig. 5. Rendering examples of our synthetic dataset.

additional parameters and negligible computational overhead, since it only applies precomputed coordinates to existing attention heads.

3.2 Spatial Autoregressive Generation

Unlike most image processing and generation tasks, the output size of image stitching is decided by the number and spatial arrangement of captured inputs. While state-of-the-art DiTs support variable sizes, their generation quality degrades severely when extrapolated beyond a limited resolution range (most modern DiTs struggle to generate output larger than 4K, which is even not sufficient for image stitching). Instead of attempting to denoise this massive panorama in a single computationally prohibitive and unstable pass, we progressively generate content patch by patch on the final canvas.

First, after projecting all input views using the specified target view and projection type, we define the spatial extent of the target canvas as the minimum bounding rectangle enclosing all valid dense-depth projected coordinates $\{(u'_i, v'_i)\}$. This bounding rectangle determines the field of view and resolution of the output panorama. Based on this size, we initialize the panorama as pure Gaussian noise. Specifically, as shown in Fig. 2, we start the generation from the spatial center of the panorama, processing local patches of size $L \times L$ and progressively expanding outwards with a stride of S. The traversal proceeds column by column from the central column, while patches within each column are generated from the center toward the bottom and top. In each patch-wise generation, the input to our AnyStitch DiT is constructed by concatenating three distinct types of tokens: 1) noisy tokens cropped from the pure noisy panorama in the current $L \times L$ spatial window, 2) input tokens which are routed into the current generation window according to whether its dense-depth projected coordinate (u'_i, v'_i) falls within the patch, 3) memory tokens that sampled from the overlapping regions of the previously generated panoramic patches, ensuring consistent generation between adjacent patches (see Fig. 13).

Moreover, we observe that DiTs' extrapolation failure in large panorama generation typically manifests in two-folds: spatial dimension and sequential dimension. On the one hand, when the panoramic canvas size significantly exceeds the resolution encountered during training, the absolute spatial frequencies in RoPE shift out of the stable distribution, leading to structural collapse. On the other hand, as the number of input views increases, the sequential indices t_i assigned to input tokens similarly encounter extrapolation issues. To address these challenges, we introduce a relative positional encoding mechanism, as illustrated in Fig. 4. Specifically, within each local generation context, we jointly shift both branches of spatial RoPE by the shared minimum coordinate along each spatial axis and shift the input sequence indices by their minimum value. This operation ensures that both spatial and sequential frequencies consistently remain within a stable, pre-trained range, regardless of the global canvas size or the total number of input images. By circumventing these two types of extrapolation failures, AnyStitch achieves robust panorama generation on a large canvas.

3.3 AnyStitchSyn Dataset

Most existing learning-based image stitching frameworks follow UDIS's unsupervised paradigms [Nie et al. 2021], primarily focusing on feature matching and warping. However, this approach inherently conflicts with the DiTs, which fundamentally depend on high-fidelity ground truths to learn the distribution. Furthermore, the image stitching community has long suffered from a critical absence of training datasets that provide casually captured, large-parallax inputs paired with geometrically consistent panoramic grount truths. To bridge this gap, we introduce the AnyStitchSyn Dataset, a synthetic dataset tailored for large-parallax stitching.

Leveraging the photorealistic rendering capabilities and explicit geometric control of 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023], we can synthesize inputs from views we want, introducing large disparities. Moreover, we can also synthesize perfectly aligned ground truths for corresponding inputs. We visualize the capturing process within 3DGS scenes in Fig. 5. We first randomly sample a camera pose to serve as the initial capture view. To simulate natural hand-held camera movements and induce parallax, we translate the camera by a random distance and inject rotational perturbations to obtain the second capture view. The corresponding RGB images and dense depth maps for both input views are then rendered.

To construct the panoramic ground truth, we define the target camera pose as the interpolated average of the input views. Three distinct types of panoramic ground truths from this target pose are rendered, *i.e.*, rectilinear, cylindrical and equirectangular. For the rectilinear ground truth, we directly render the scene using standard perspective projection with an enlarged field-of-view (FOV). For the cylindrical and equirectangular ground truths, we first construct an intermediate environmental cubemap by rendering six orthogonal perspective images from the target pose. Next, we formulate dense 3D viewing rays based on continuous cylindrical and spherical (longitude-latitude) coordinates, respectively. By projecting these specific rays onto the cubemap faces and performing bilinear sampling, we effectively obtain seamless cylindrical and equirectangular panoramas.



Fig. 6. Visual comparisons on AnyStitchSyn test set. Our method uses DA3-estimated geometry.

Table 1. Full-reference comparison on the AnyStitchSyn test set. AnyStitch uses camera poses and depth maps estimated by DA3 rather than the ground-truth camera poses and rendered depth maps available from the 3DGS scenes. For each method, metrics are computed over the non-black region shared by its output and the ground truth.

Method	AnyStitchSyn$^{\text{w/ DA3 geometry}}$	
	SSIM ↑	LPIPS ↓
UDIS++ [Nie et al. 2023]	0.7471	0.5255
SRStitcher [Xie et al. 2024b]	0.7165	0.5408
RopStitch [Nie et al. 2026]	0.7487	0.5052
AnyStitch	**0.8061**	**0.3200**

Last, we utilize Eq. 1 to explicitly project the input views onto the chosen ground truth canvas. By computing the minimum bounding rectangle that tightly encloses these projected coordinates, we crop the full panorama to yield the final, perfectly aligned ground truth tailored for the given inputs.

We synthesize our dataset using 1000 scenes from InteriorGS dataset[SpatialVerse Research Team 2025]. We utilize the first 950 scenes for training, randomly generating 1 to 3 data groups per scene to collect a total of 2,192 training pairs. Crucially, each training pair includes two input views alongside corresponding ground-truth panoramas rendered in three different projection models (The effectiveness analysis of this mixed-projection data is provided in supplementary material). For evaluation, we extract 100 test pairs from the remaining 50 scenes and adopt rectilinear projection for ground truth.

4 Experiment

4.1 Experimental Setting

Implementation details. We implement AnyStitch built upon FLUX.2-klein-9B [Labs 2025], a state-of-the-art Diffusion Transformer (DiT) designed for image generation and editing. To adapt the powerful pre-trained generative prior encapsulated in this model to the large-parallax image stitching task, we fine-tune the network on our AnyStitchSyn Dataset. Specifically, we employ Low-Rank Adaptation (LoRA) [Hu et al. 2022] and set LoRA rank to 32. The model is fine-tuned for 6 epochs with a batch size of 4. We optimize the network using the Adam optimizer with a constant learning rate set to 1×10^{-4}. Training takes approximately 8 hours on 4 NVIDIA A800 equivalent GPUs. At inference, we use a patch size L of 1024 and a stride S of 768, corresponding to a 25% overlap, together with 40 denoising steps. More training details are available in the supplementary.

Evaluation metrics. Beyond conventional full-reference metrics commonly used in image processing (*e.g.*, PSNR, SSIM [Wang et al. 2004] and LPIPS [Zhang et al. 2018]), evaluating large-parallax image

Table 2. Quantitative comparison on two large-parallax datasets. † denotes our reproduction of a PIS3R/LiftProj-style pipeline.

Method	PTIS dataset SIQS ↑	RISMR dataset SIQS↑
SPW [Liao and Li 2019a]	7.26	7.11
LPC [Jia et al. 2021]	7.49	7.64
UDIS++ [Nie et al. 2023]	7.63	6.69
SRStitcher [Xie et al. 2024b]	8.00	7.88
PIS3R/LiftProj-style† [Jia et al. 2025; Zhu et al. 2025]	7.91	8.36
GPIS [Tuli et al. 2025]	6.06	–
RopStitch [Nie et al. 2026]	5.57	4.57
AnyStitch	**9.40**	**9.71**

stitching also requires assessing semantic and perceptual coherence. To this end, we employ two reference-free metrics: the Stitching Image Quality Score (SIQS) and Multi-Image Comparative Quality Score (MICQS) [Xie et al. 2024a]. SIQS leverages the advanced reasoning capabilities of large multimodal models [Bai et al. 2025] to grade the stitched panoramas across five critical perceptual dimensions: seam visibility, brightness transition, geometric distortion, clarity, and the presence of abnormal content. Meanwhile, MICQS utilizes large multimodal models as an impartial judge to evaluate and compare the stitched panoramas based on the aforementioned five perceptual dimensions. In addition, we also conduct human subjective evaluation. Specifically, for each dataset, we randomly select 20 scenes and invite 51 users to participate. During the evaluation, users are asked to perform pairwise comparisons between the stitching results of our model and a randomly selected baseline method.

Comparison methods and benchmarks. Since many recent advancements in image stitching focus on two-view inputs, we first evaluate our approach by comparing it with existing two-view methods, including traditional methods SPW [Liao and Li 2019a], LPC [Jia et al. 2021] and learning-based methods UDIS++ [Nie et al. 2023], SRStitcher [Xie et al. 2024b], PIS3R [Zhu et al. 2025], LiftProj [Jia et al. 2025], GPIS [Tuli et al. 2025] and RopStitch [Nie et al. 2026]. Notably, similar to our approach, SRStitcher and GPIS also leverage generative models for image stitching, while PIS3R and LiftProj employ 3D-based models to establish geometry-aware correspondences under large parallax, making them highly relevant competitors for evaluating the superiority of our method. Since the official implementations of PIS3R and LiftProj were unavailable, we reproduce their shared reconstruction-and-projection pipeline by estimating a point cloud with VGGT, projecting it to the target view, and applying a generative restoration model fine-tuned on AnyStitchSyn. We denote this reproduction as PIS3R/LiftProj-style† rather than reporting it as an official result. We conduct experiments on our AnyStitchSyn test set and two challenging public benchmarks: the PTIS dataset [Zhang and Liu 2014] and the RISMR dataset [Herrmann et al. 2018]. These datasets are characterized by scenes exhibiting large parallax, making them ideal testing cases for assessing

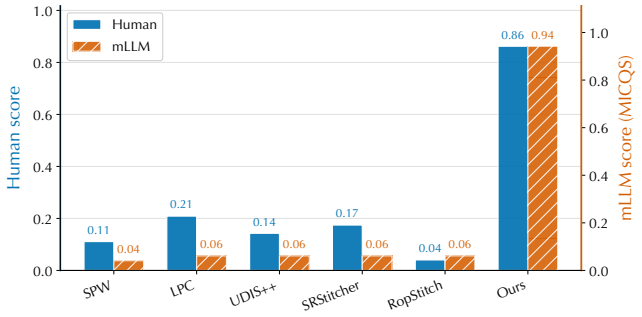


Fig. 7. User study and mLLM evaluation on two large-parallax datasets.

the robustness of different image stitching algorithms. Additionally, we compare our method against the state-of-the-art multi-view image stitching method [Du et al. 2022] on standard classical scenes and our custom large-parallax scenes, evaluating the robustness of AnyStitch with multiple inputs.

4.2 Comparison results

Two-view comparison. We first evaluate on the AnyStitchSyn test set, which provides pixel-aligned ground-truth panoramas, using full-reference metrics SSIM and LPIPS. We compare AnyStitch with UDIS++, SRStitcher, and RopStitch. For a fair evaluation, AnyStitch estimates the input depth and camera poses using Depth Anything 3 (DA3) [Lin et al. 2025], rather than using the rendered ground-truth geometry available in the synthetic dataset. To prevent irregular black boundaries from biasing the metrics, for each method, we compute the metrics over the non-black region shared by its output and the ground truth. As reported in Tab. 1, AnyStitch achieves the best result on both metrics, indicating that our method better handles large parallax while preserving content fidelity. The qualitative comparisons in Fig. 6 further show that AnyStitch produces fewer structural distortions and better preserves the scene content relative to the ground truth.

We then perform a reference-free evaluation on the two real-world large-parallax benchmarks. We summarize the quantitative comparison results in Tab. 2. Our AnyStitch achieves the best performance, indicating that our method outputs the stitched panorama with superior visual quality and the most plausible semantic coherence. Specifically, our method outperforms the generative method SRStitcher in a large margin on both datasets in terms of SIQS score (9.40/9.71 *v.s.* 8.00/7.88). Beyond individual image quality assessment, Fig. 7 compares our method with the baselines through human and mLLM-based MICQS evaluations. These results are consistent with the quantitative findings in Tab. 2.

The qualitative comparison results are visualized in Fig. 8 and Fig. 9. Traditional and learning-based frameworks, such as LPC, UDIS++, and RopStitch, struggle with depth disparities. Because they rely on 2D homography or localized 2D deformation fields, and often forcibly warp non-planar regions to satisfy feature alignment, resulting in unnatural spatial stretching and severe geometric distortions, as evident on buildings and the temple pillar in Fig 9. Furthermore, as they fail to disentangle complex 3D foreground-background



Fig. 8. Visual comparisons on RISMR [Herrmann et al. 2018] dataset.

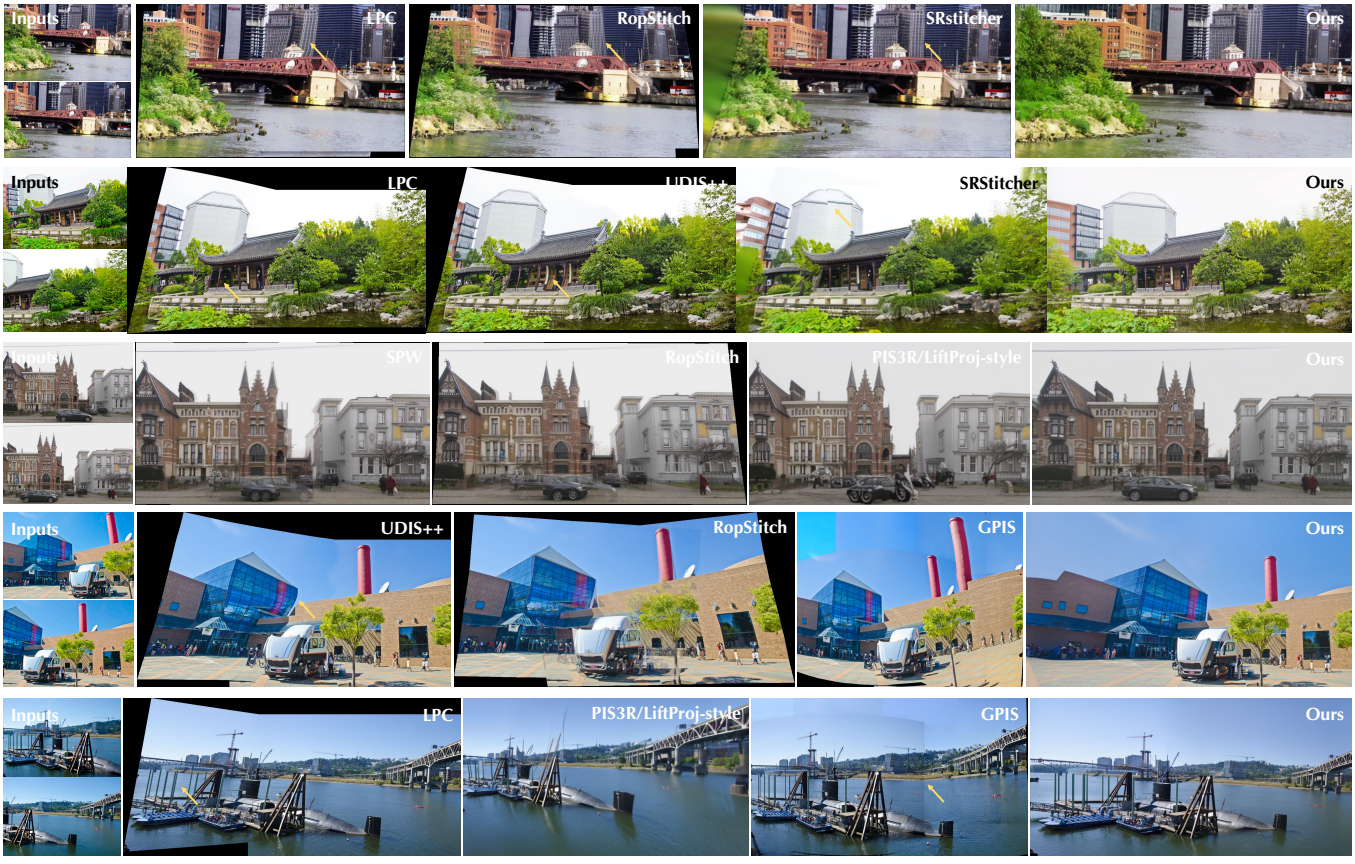


Fig. 9. Visual comparisons on the PTIS dataset [Zhang and Liu 2014].

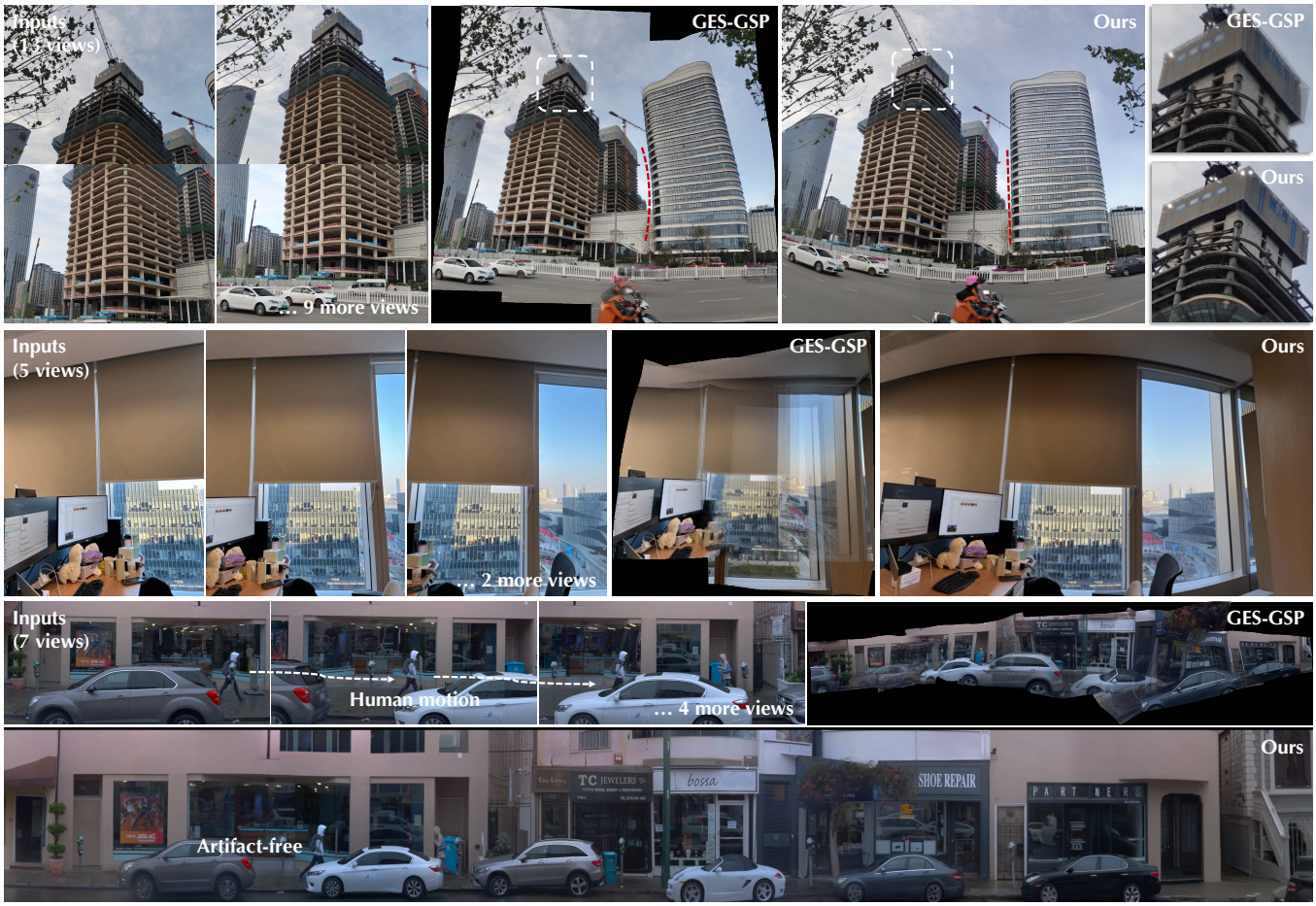


Fig. 10. Comparison on multi-view scenes. Top and middle scenes: equirectangular projection. Bottom scene: parallel projection.

relationships, these methods introduce severe ghosting artifacts on distinct foreground objects, such as people, moving cars and traffic cones in Fig 8. While SRStitcher and GPIS employ generative models to facilitate stitching and hallucinate missing regions, they lack explicit and precise 3D geometric guidance. Consequently, their results frequently exhibit structural distortions, visible seams, and blurry content, as we marked in Fig. 9. Although the PIS3R/LiftProj-style baseline uses explicit 3D geometry to guide stitching, its pixel-level projection can lead to severe content distortions.

In contrast, AnyStitch generates visually stunning, artifact-free panoramas that adhere to the real-world's structure. By grounding DiT with the precise hierarchical projected RoPE derived from explicit 3D geometry, our method perfectly aligns intricate architectural lines (see Fig. 9) without structural degradation. Moreover, AnyStitch can correctly handle the large parallax between foreground and background, rendering artifact-free panoramas (see Fig. 8).

Multi-view comparison. Fig. 10 presents a qualitative comparison between AnyStitch and the state-of-the-art multi-view method GES-GSP [Du et al. 2022] on three scenarios with increasing camera translation and parallax. In the top scene, which is a simple case from nearly without camera translation. However, GES-GSP still suffers from overlay artifacts and structure distortion. In the middle and bottom scene, where severe parallax occurs, GES-GSP degrades remarkably. In contrast, our AnyStitch is robust to both tiny-parallax and large-parallax scenes, producing visually pleasing panoramas. Moreover, as shown in the bottom case, AnyStitch effectively handles simple isolated human motion, enabled by the full attention mechanism of the DiT architecture.

4.3 Ablation Study

To validate the effectiveness of our proposed architectural designs, we conduct comprehensive ablation studies on two key components, *i.e.*, Hierarchical Projected RoPE (HPRoPE) and our autoregressive generation strategy.

Hierarchical Projected RoPE. The accurate geometric perception capability of our model relies on the injection of spatial correspondence. To investigate this, we compare three variants. The first variant (denoted as *baseline*) is the vanilla DiT, which is completely



Fig. 11. Visual comparisons of different geometric injection methods. Our method uses GT geometry.



Fig. 12. Visual effectiveness of using two complementary projected RoPE.

Table 3. Quantitative ablation study on HPRoPE. The geometry-conditioned variants use the ground-truth camera poses and rendered depth maps from the 3DGS scenes.

Model	AnyStitchSyn$^{\text{w/ GT geometry}}$ PSNR↑	SSIM↑
baseline	13.71	0.7646
baseline + projected inputs	20.72	0.8744
(Ours) baseline + HPRoPE	**21.39**	**0.8806**

devoid of any explicit geometric priors. The second variant (denoted as *baseline + projected inputs*) utilizes estimated 3D information to project raw inputs to the panoramic canvas before tokenization. This variant still employs standard 2D positional encodings. The final variant is our proposed HPRoPE (denoted as *baseline + HPRoPE*), which directly modifies the positional encodings to inject precise projection coordinates into the latent space. We conduct these evaluations on the test set of our AnyStitchSyn Dataset, measuring the performance using PSNR and SSIM. Unlike the full-reference comparison in Tab. 1, the geometry-conditioned variants in this controlled ablation use the ground-truth camera poses and rendered depth maps from the 3DGS scenes. The quantitative and qualitative results are presented in Tab. 3 and Fig. 11, respectively.

As shown, the baseline variant entirely depends on its inherent self-attention mechanism to infer spatial correspondences across the unorganized, large-parallax input views. As shown in Tab. 3 and Fig. 11, this unconstrained generation struggles significantly, yielding poor performance (13.71 PSNR and 0.7646 SSIM). The second variant attempts to mitigate this by introducing explicit spatial modeling in the image space. Although it forces the model to synthesize a structurally correct panorama (significantly boosting PSNR to 20.72), pixel-imperfect depth will cause the content distortion in the projected inputs. Consequently, it suffers from noticeable information loss and drift, as indicated by Fig. 11. Instead, our proposed method resolves these issues by explicitly feeding the geometric mapping into the attention mechanism via Hierarchical Projected RoPE. As a result, our method achieves the highest quantitative performance (21.39 PSNR and 0.8806 SSIM). Fig. 11 also demonstrates that our approach can effectively avoid content loss and distortion compared to performing projection in image space.

In-depth analysis of HPRoPE. We further investigate the impact of the introduction of two distinct depth projection, *i.e.*, accurate dense depth and the median depth. As illustrated in Fig. 12, relying exclusively on either single depth projection produces suboptimal results. On the one hand, without RoPE projected by dense depth, it fails to generate results with accurate 3D geometry, leading to severe content change. On the other hand, using RoPE projected by dense depth as the sole geometric guidance may also cause the content change due to the inaccurate depth estimation, especially in the transition regions between the foreground and the background. As shown in the zoomed-in region of Fig. 12, the statue on the roof is severely distorted. Our proposed hierarchical projected RoPE resolves this dilemma by simultaneously feeding both geometric cues into disjoint subsets of attention heads. This dual-precision mechanism balances geometrical alignment and the fidelity preservation, ensuring the generation of a complete statue in Fig 12 under the view change. We provide an additional detailed analysis under imperfect geometry in supplementary material.

Next, we evaluate the effectiveness of our patch-wise autoregressive generation mechanism for flexible panoramas generation.

Memory tokens. We first investigate the critical role of memory tokens during the progressive patch generation process. As shown in Fig. 13, generating adjacent spatial patches without the injection of memory tokens leads to the overlay artifacts. This failure occurs because, in the absence of memory tokens, the DiT processes each local generation window independently, lacking the contextual awareness of the previously generated regions. By concatenating memory tokens sampled from the overlapping areas, our method



Fig. 13. Visual effectiveness of introducing memory tokens.

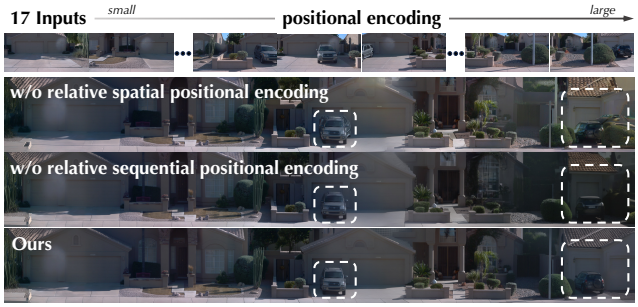


Fig. 14. Visual effectiveness of relative positional encoding.

seamlessly anchors the newly generated latent features to the existing global canvas, ensuring the overall consistency when generating a large panorama.

Relative positional encoding. Second, we investigate the necessity of the relative positional encoding when processing a large number of views (*e.g.*, 17 inputs captured in a long street, as shown in Fig. 14). In such scenarios, absolute positional coordinates inevitably grow beyond the distribution observed during DiT training, triggering extrapolation collapse. When the relative spatial normalization is disabled (*w/o relative spatial positional encoding*), the spatial frequencies shift out of the stable pre-trained range. As highlighted in the rightmost dashed box, this spatial extrapolation failure causes the network to generate noticeable artifacts. Similarly, disabling the relative normalization for the input sequence index (*w/o relative sequential positional encoding*) leads to sequential extrapolation failure. By integrating these components, we constrain all positional encodings within a stable, trained range. As demonstrated in the bottom row of Fig. 14, our AnyStitch is able to process a variable number of inputs and output seamless contents across a large canvas.

Window size and stride. Last, we evaluate four generation-window sizes while fixing the overlap ratio at 25%. All timings reported in Tab. 4 use 40 denoising steps and are measured on panoramas from AnyStitchSyn test set with an average resolution of 3337 × 1586. Although the 1536/1152 setting yields the highest PSNR and the 2048/1536 setting yields the highest SSIM, the default 1024/768 setting provides the best overall trade-off between reconstruction quality and inference time.

Table 4. Ablation of generation-window and stride sizes. Runtime is measured on panoramas with an average resolution of 3337 × 1586.

Window/stride	AnyStitchSyn$^{\text{w/ GT geometry}}$		
	PSNR ↑	SSIM ↑	Time ↓
512/384	20.13	0.8597	7.5 min
1024/768	21.39	0.8806	**7.2 min**
1536/1152	**21.59**	0.8795	13.6 min
2048/1536	21.33	**0.8816**	10.6 min

4.4 More Applications

As a generative stitching framework, AnyStitch can move far beyond mere image stitching. As shown in Fig. 15, it can generalize to cross-time stitching, HDR fusion of misaligned exposures, and consistent image outpainting, further highlighting the robustness and potential of our method.

5 Limitations

Although our experiments demonstrate AnyStitch's strong potential for casual-photography panorama stitching, several limitations remain. First, applying AnyStitch to measurement-critical applications, such as mapping or satellite imagery, may pose some fidelity risks due to its generative nature. Second, as a 40-step diffusion model, AnyStitch is slower than conventional warping-based methods: generating a 2K panorama takes approximately 5 minutes and requires about 40 GB of GPU memory on one NVIDIA A800 equivalent GPU. Finally, although the method exhibits some tolerance to isolated moving subjects, complex crowded scenes may still produce motion-related duplication and ghosting artifacts, as shown in Fig. 16. Reducing the number of sampling steps through distillation and incorporating dynamic-scene training data are promising directions for future work.

6 Conclusion

In this paper, we present AnyStitch, a novel generative framework that tackles the highly challenging problem of large-parallax image stitching. By reformulating the stitching pipeline from traditional 2D pixel-space warping to a reference-based generative process, we effectively circumvent the persistent ghosting, seams, and structural tearing. To overcome the spatial misalignment issues inherent in standard diffusion models, we introduce the Hierarchical Projected Rotary Positional Encoding (HPRoPE). This mechanism injects 3D geometric priors directly into the latent space, utilizing varying attention-head precisions to balance rigorous structural constraints with high-fidelity content generation. Furthermore, we design a patch-wise spatial autoregressive strategy that bypasses the computational resolution limits of generative models. To train our model, we synthesize large-parallax inputs with corresponding ground truths from 3DGS data. With these features, AnyStitch enables the seamless integration of a variable number of casual captures into visually coherent panoramas. Experiments demonstrate that AnyStitch consistently outperforms both existing two-view methods and multi-view frameworks.

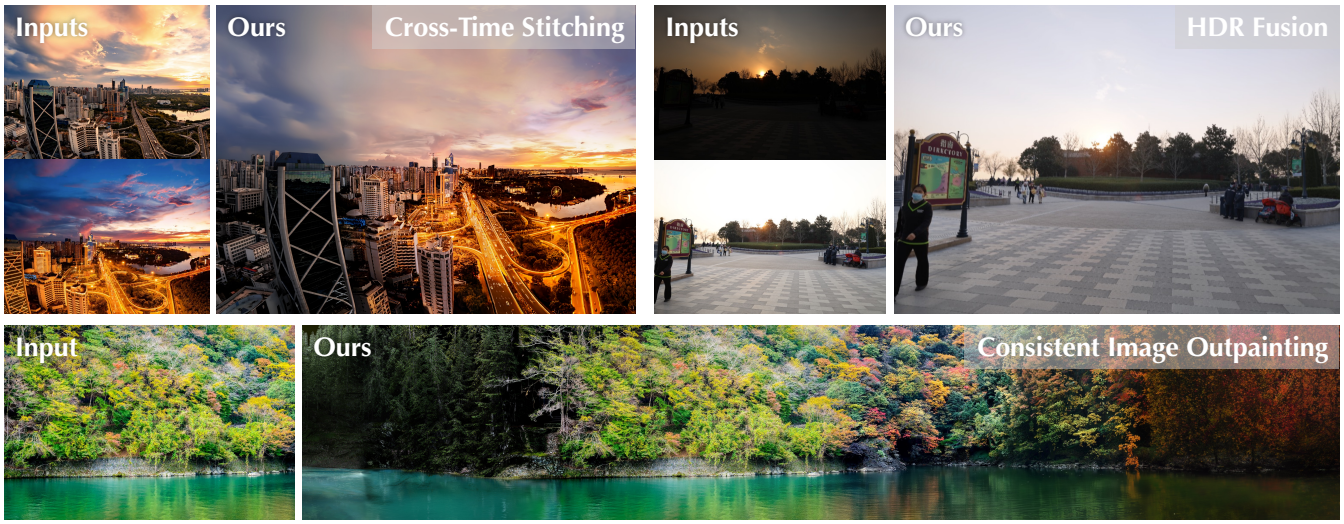


Fig. 15. More applications of our AnyStitch.



Fig. 16. A failure case in a crowded dynamic scene. Given eight input views captured at different times, moving pedestrians may be duplicated or appear ghosted in the stitched panorama, as highlighted in the zoomed-in region.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2025YFE0201300) and the Research Grants Council of Hong Kong under the Early Career Scheme (ECS) (No. 24209224). Additional support was provided by the CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund (No. 2025A0505000053). The work was also supported in part by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. We also thank Bohan Zhou, Jin Dai, Chen Yang, and Shaoyong Zhang for their help with this project.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631* (2025).

Yunpeng Bai, Haoxiang Li, and Qixing Huang. 2026. Positional Encoding Field. In *International Conference on Learning Representations*.

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

Yu-Sheng Chen and Yung-Yu Chuang. 2016. Natural image stitching with the global similarity prior. In *European conference on computer vision*. Springer, 186–201.

Peng Du, Jifeng Ning, Jiguang Cui, Shaoli Huang, Xinchao Wang, and Jiaxin Wang. 2022. Geometric structure preserving warp for natural image stitching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3688–3696.

Junhong Gao, Seon Joo Kim, and Michael S Brown. 2011. Constructing image panoramas using dual-homography warping. In *CVPR 2011*. IEEE, 49–56.

Peter Hedman and Johannes Kopf. 2018. Instant 3D Photography. *ACM Transactions on Graphics* 37, 4, Article 101 (2018).

Charles Herrmann, Chen Wang, Richard Strong Bowen, Emil Keyder, Michael Krainin, Ce Liu, and Ramin Zabih. 2018. Robust image stitching with multiple registrations. In *Proceedings of the European conference on computer vision (ECCV)*. 53–67.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr* 1, 2 (2022), 3.

Qi Jia, Xiaomei Feng, Yu Liu, Xin Fan, and Longin Jan Latecki. 2023. Learning pixel-wise alignment for unsupervised image stitching. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1392–1400.

Qi Jia, ZhengJun Li, Xin Fan, Haotian Zhao, Shiyu Teng, Xinchen Ye, and Longin Jan Latecki. 2021. Leveraging line-point consistence to preserve structures for wide parallax image stitching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12186–12195.

Yuan Jia, Ruimin Wu, Rui Song, Jiaojiao Li, and Bin Song. 2025. LiftProj: Space Lifting and Projection-Based Panorama Stitching. *arXiv preprint arXiv:2512.24276* (2025).

Hengzhe Jin, Lang Nie, Chunyu Lin, Xiaomei Feng, and Yao Zhao. 2025. PixelStitch: Structure-Preserving Pixel-Wise Bidirectional Warps for Unsupervised Image Stitching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 28125–28134.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics* 36, 4 (2017), 1–13. doi:10.1145/3072959.3073599

Black Forest Labs. 2025. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2.

Kyu-Yul Lee and Jae-Young Sim. 2020. Warping residual based image stitching for large parallax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8198–8206.

Jing Li, Baosong Deng, Rongfu Tang, Zhengming Wang, and Ye Yan. 2019. Local-adaptive image alignment based on triangular facet approximation. *IEEE Transactions on Image Processing* 29 (2019), 2356–2369.

Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang. 2017. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Transactions on multimedia* 20, 7 (2017), 1672–1687.

Lingen Li, Guangzhi Wang, Xiaoyu Li, Zhaoyang Zhang, Qi Dou, Jinwei Gu, Tianfan Xue, and Ying Shan. 2026. CubeComposer: Spatio-Temporal Autoregressive 4K 360° Video Generation from Perspective Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 32625–32635.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. 2025. Cameras as Relative Positional Encoding. In *Advances in Neural Information Processing Systems*, Vol. 38. 15984–16009.

Tianli Liao and Nan Li. 2019a. Single-perspective warps in natural image stitching. *IEEE transactions on image processing* 29 (2019), 724–735.

Tianli Liao and Nan Li. 2019b. Single-perspective warps in natural image stitching. *IEEE transactions on image processing* 29 (2019), 724–735.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. 2025. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025).

Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. 2021. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Transactions on Image Processing* 30 (2021), 6184–6197.

Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. 2023. Parallax-tolerant unsupervised deep image stitching. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7399–7408.

Lang Nie, Yuan Mei, Kang Liao, Xunqiu Xu, Chunyu Lin, and Bin Xiao. 2026. Robust image stitching with optimal plane. *IEEE Transactions on Visualization and Computer Graphics* (2026).

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6121–6132.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. 2024. GenWarp: Single Image to Novel Views with Semantic-Preserving Generative Warping. In *Advances in Neural Information Processing Systems*, Vol. 37. 80220–80243.

Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 2020. 3D Photography Using Context-Aware Layered Depth Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8025–8035.

Manycore Tech Inc. SpatialVerse Research Team. 2025. InteriorGS: A 3D Gaussian Splatting Dataset of Semantically Labeled Indoor Scenes. https://huggingface.co/datasets/spatialverse/InteriorGS.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2446–2454.

Mathieu Tuli, Kaveh Kamali, and David B Lindell. 2025. Generative Panoramic Image Stitching. *arXiv preprint arXiv:2507.07133* (2025).

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5294–5306.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20697–20709.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. 2026. π^3: Permutation-Equivariant Visual Geometry Learning. In *International Conference on Learning Representations*.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. 2026. RayRoPE: Projective Ray Positional Encoding for Multi-view Attention. *arXiv preprint arXiv:2601.15275* (2026).

Ziqi Xie, Xiao Lai, Weidong Zhao, Siqi Jiang, Xianhui Liu, and Wenlong Hou. 2024a. Modification takes courage: Seamless image stitching via reference-driven inpainting. *arXiv preprint arXiv:2411.10309* (2024).

Ziqi Xie, Weidong Zhao, Xianhui Liu, Jian Zhao, and Ning Jia. 2024b. Reconstructing the image stitching pipeline: Integrating fusion and rectangling into a unified inpainting model. *Advances in Neural Information Processing Systems* 37 (2024), 123812–123845.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, Animesh Garg, et al. 2026. MosaicMem: Hybrid Spatial Memory for Controllable Video World Models. *arXiv preprint arXiv:2603.17117* (2026).

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. ViewCrafter: Taming Video Diffusion Models for High-Fidelity Novel View Synthesis. *arXiv preprint arXiv:2409.02048* (2024).

Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. 2013. As-projective-as-possible image stitching with moving DLT. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2339–2346.

Fan Zhang and Feng Liu. 2014. Parallax-tolerant image stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3262–3269.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.

Muhua Zhu, Xinhao Jin, Chengbo Wang, Yongcong Zhang, Yifei Xue, Tie Ji, and Yizhen Lao. 2025. PIS3R: Very Large Parallax Image Stitching via Deep 3D Reconstruction. *arXiv preprint arXiv:2508.04236* (2025).