PanoFree: Tuning-Free Holistic Multi-view Image Generation with Cross-view Self-Guidance

Aoming Liu[1,2*], Zhong Li[1†], Zhang Chen[1†], Nannan Li[2], Yi Xu[1], and Bryan A. Plummer[2]

[1] OPPO US Research Center, Palo Alto, CA 94303, USA
[2] Boston University, Boston, MA 02215, USA

Abstract. Immersive scene generation, notably panorama creation, benefits significantly from the adaptation of large pre-trained text-to-image (T2I) models for multi-view image generation. Due to the high cost of acquiring multi-view images, tuning-free generation is preferred. However, existing methods are either limited to simple correspondences or require extensive fine-tuning to capture complex ones. We present PanoFree, a novel method for tuning-free multi-view image generation that supports an extensive array of correspondences. PanoFree sequentially generates multi-view images using iterative warping and inpainting, addressing the key issues of inconsistency and artifacts from error accumulation without the need for fine-tuning. It improves error accumulation by enhancing cross-view awareness and refines the warping and inpainting processes via cross-view guidance, risky area estimation and erasing, and symmetric bidirectional guided generation for loop closure, alongside guidance-based semantic and density control for scene structure preservation. In experiments on Planar, 360°, and Full Spherical Panoramas, PanoFree demonstrates significant error reduction, improves global consistency, and boosts image quality without extra fine-tuning. Compared to existing methods, PanoFree is up to 5x more efficient in time and 3x more efficient in GPU memory usage, and maintains superior diversity of results (2x better in our user study). PanoFree offers a viable alternative to costly fine-tuning or the use of additional pre-trained models.

Keywords: Tuning-free generation; Multi-view Image, Panorama

1 Introduction

Text-to-image (T2I) generation over multiple views for immersive scenes, like panorama generation, is a challenging task requiring coherence and diversity among many generated images (*e.g.*, [3, 8, 12, 16, 17, 21, 36, 37, 39, 48]). Early work using GANs or VAEs (*e.g.*, [4–6, 11, 24, 25, 47, 50]) have been replaced recently with diffusion-based models (*e.g.*, [1,9,10,22,23,46,49,51,52,54]), often leveraging Stable Diffusion [38]. State-of-the-art panorama generation methods

*Work partly completed during Aoming's internship at OPPO US Research Center.
†Corresponding Authors.

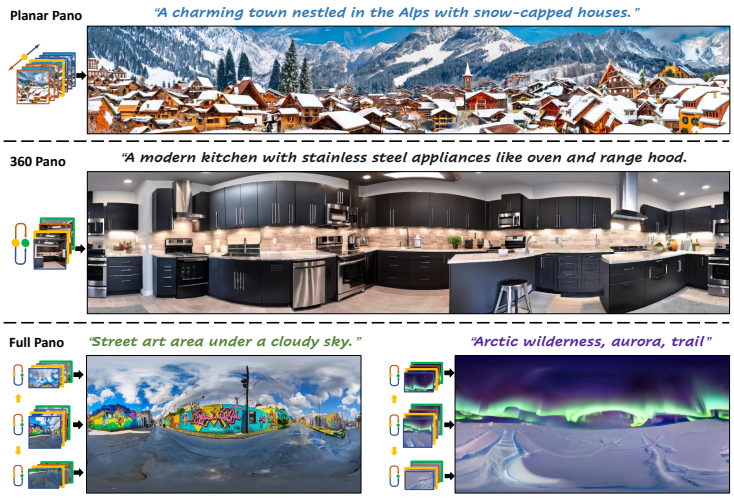


Fig. 1: PanoFree can generate multi-view images according to different types of correspondences without fine-tuning, and a natural application is tuning-free generation for different types of panoramas. We demonstrate this by generating three commonly used panoramas. Top: Planar Panorama; Middle: 360 Panorama; Bottom: Full Panorama.

use Joint Diffusion (*e.g.*, [1,22,46]), where parallel diffusion processes to generate multi-view images and enhancing global consistency by fusing latent or attention features based on cross-view correspondences. However, we find these methods can only generate some types of panoramas, limiting their scope (*e.g.*, [1, 22]), or require fine-tuning using expensive panorama datasets (*e.g.*, [22, 46]).

To address these challenges, we propose PanoFree a tuning-free multi-view image generation method using iterative warping and inpainting of perspective images to support diverse correspondences with low costs (see Fig. 1 for example generations). Iterative warping and inpainting of perspective images provide a means to the diverse multi-view correspondences required in panorama generation without additional fine-tuning [4, 11, 18]. However, recent work has overlooked these benefits due to accumulated errors causing suboptimal image quality [1, 22, 46]. We find that most accumulated errors from iterative warping can be attributed to the deficient conditions during generation. Specifically, conditioning solely on the previous image narrows cross-view awareness, leading to inconsistencies. Warping and inpainting can also propagate noise, *e.g.*, truncated objects or jagged edges. Additionally, the given conditions may be incomplete to meet specific requirements, such as ensuring 360-degree consistency for loop closure and maintaining correct spatial relationships for realistic scenes.

To address these issues, PanoFree expand cross-view awareness by conditioning the current view on multiple views with guided image synthesis techniques such as SDEdit [29]. Then, PanoFree estimates and erases the risky areas, regions likely containing noise, to reduce the noise introduced by warping and inpainting. In addition, PanoFree adopts a bidirectional generation path with a symmetrical conditioning strategy for loop closure. Lastly, PanoFree further

utilizes pseudo global guidance with region-specific semantic and density control to make scene structure more reasonable.

We evaluate PanoFree on three text-to-panorama generation tasks: Planar, 360°, and Full Spherical Panoramas. PanoFree effectively alleviating accumulated errors in sequential generation, and significantly improves image quality and global consistency (*e.g.* 31.6% better in FID). This enables PanoFree to have better (or at least comparable) results to the state-of-the-art [1, 22, 46], despite these methods either having narrower applications or requiring fine-tuning datasets. Specifically, PanoFree is up to 5x more efficient in time and 3x more efficient in GPU memory usage, and maintains superior diversity of results (2x better in our user study). Lastly, PanoFree is also highly flexible, enabling it to plug-in-and-play with various pre-trained T2I models and adapters.

Our contributions can be summarized as follows:

- We introduce PanoFree, a tuning-free multi-view image generation method applicable for various correspondences and pre-trained T2I models. Thus PanoFree can greatly reduce the data and fine-tuning costs for immersive scene generation tasks such as text-to-panorama generation.
- We provide a in-depth perspective of accumulated errors and identify the deficient conditions as the main causes. We further effectively rectify deficient conditions and alleviate accumulated errors with the cross-view guidance as well as risky area estimation and erasing in PanoFree.
- As far as we know, PanoFree is the first to achieve feasible tuning-free generation for 360° Panoramas and Full Spherical Panoramas.

2 Related Work

Diffusion Models [15, 19, 40, 42–45] are a popular framework for generative models. Early work required a long trajectory for sampling to produce high-quality samples [7, 45], before being sped up with advanced sampling techniques that also preserved generation quality [20, 26, 27, 41]. Latent Diffusion Models (LDMs) [33, 38] made these models more efficient by training in the latent space.

T2I Diffusion and Panorama Generation. Diffusion models are widely adopted for text-to-image (T2I) generation [31, 36, 38, 39]. Many downstream tasks used large pre-trained T2I diffusion models, like Stable Diffusion [38], to boost performance [4, 11, 18], including panorama generation [4, 11, 18]. These methods have largely supplanted GAN and VAE methods [4–6, 11, 24, 25, 32, 47, 50], with most recent work in panorama generation tasks using diffusion models [1, 9, 10, 22, 23, 46, 49, 51, 52, 54]. These diffusion-based panorama generation methods use joint diffusion to fuses multiple diffusion processes through latent or attention manipulation [1, 22, 46]. However, they are either limited to modeling simple correspondences or require extensive fine-tuning to model complex ones.

Guided Image Synthesis with Diffusion Models. It can be challenging to achieve satisfactory results solely relying on text guidance. Therefore, some prior work [28–30, 53] guide or control the generation results with reference images as fine-grained condition. ControlNet [53] and T2I-Adapter [30] are the most

commonly used methods to incorporate additional image conditions by adding extra image encoders, but they all require few-shot fine-tuning. SDEdit [29] achieves tuning-free guided image synthesis by adding noise to the guide image and then denoising it back to a real image using a pre-trained diffusion model.

3 Method

PanoFree targets the text-to-panorama generation task, which takes textual descriptions as guidance to create multi-view perspective images that can be stitched into a wide-angle, high-quality panorama. PanoFree generates multi-view images through sequential warping and inpainting steps, which typically results in acclimated errors due to deficit conditions (discussed in Sec. 3.1). Each component of Panofree is designed to minimize the effect of these various deficiencies. Specifically, Sec. 3.2 mitigates inconsistency using SDEdit-based cross-view guidance and Sec. 3.3 removes artifact-inducing content by estimating and erasing risky areas. At a higher level, PanoFree employs Bidirectional Generation with Symmetric Guidance for loop closure and error reduction (Sec. 3.4). Additionally, it applies guidance-based semantic and density control for scene structure preservation (Sec. 3.5). See Fig. 2 for an overview of our approach.

3.1 Deficient Conditions behind Accumulated Errors

In this section, we discuss the causes behind the deficit conditions in the iterative warping and inpainting process and reformulate the problem as conditional generation. Given the text prompt c_t, i-th view's image $\mathbf{x}_i$, warping function $\mathcal{W}$, transformation matrix of the projection from i-th view to the $(i+1)$-th view $\mathbf{P}_i^{i+1}$, and pre-trained T2I inpainting model $\mathbf{\Phi}_{inp}$, the warping and inpainting step to generate the $(i+1)$-th view can be denoted as:

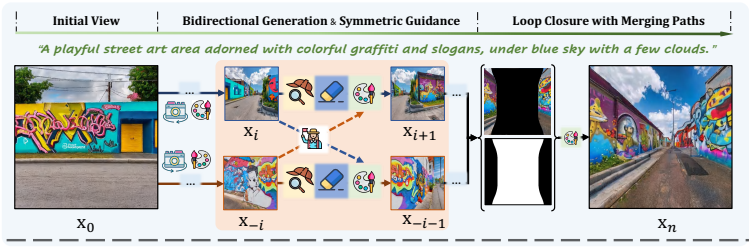
$$\hat{\mathbf{x}}_i, \mathbf{m}_i = \mathcal{W}(\mathbf{x}_i, \mathbf{P}_i^{i+1}); \quad \mathbf{x}_{i+1} = \mathbf{\Phi}_{inp}(\hat{\mathbf{x}}_i, \mathbf{m}_i, c_t), \tag{1}$$

where $\hat{\mathbf{x}}_i$ is the image warped from i-th view to $(i+1)$-th view and $\mathbf{m}_i$ is the masking indicating the area to inpaint. And we can simplify the warping and inpainting steps in the following conditional image generation form:

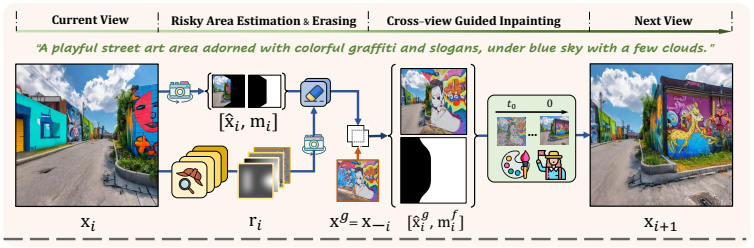
$$\mathbf{x}_{i+1} \sim \ q(\mathbf{x}|c_t, \mathbf{x}_i, \mathbf{P}_i^{i+1}). \tag{2}$$

However, during this generation process, we found conditions can become deficient. Major accumulated errors arise from three types of deficient conditions: Biased Conditions, Noisy Conditions, and Partial Conditions. See Sec. A.2 in the supplementary for detailed error illustrations.

Biased Conditions is the most obvious problem. In the above step, $\mathbf{x}_{i+1}$ is solely conditioned on $\mathbf{x}_i$, which biases the cross-view awareness heavily to i-th view. If $\mathbf{x}_i$ has deviated from the desired global distribution in certain aspects, then $\mathbf{x}_{i+1}$ is likely to continue deviating in the same direction, resulting in significant inconsistency. We also found that slight style and content shifts often accumulate in this way, leading to significant inconsistency between distant views.



(a) Overview of the generation and guidance framework.



(b) Illustration of a single warping and inpainting step.

Fig. 2: Overview of our PanoFree method, taking 360 Panorama Generation as an example. (a): At a framework level, PanoFree adopts two generation paths with opposite viewpoint translation or rotation. It enhances consistency by symmetrically selecting views from the other path as guidance to generate a new view (Sec. 3.4). Loop closure is ensured by merging these two paths. (b): In each warping and inpainting step, PanoFree reduces accumulated error by guiding the inpainting process with cross-view images (Sec. 3.2), along with estimating and erasing risky areas (Sec. 3.3).

Noisy Conditions mainly refer to $\mathbf{x}_i$ containing artifact-inducing contents. Existing artifacts in $\mathbf{x}_i$ could guide inpainting model to generate similar artifacts in $\mathbf{x}_{i+1}$ and propagate to every following view. Additionally, disjointed or distorted areas, jagged or sharp content, and objects truncated by edges in $\mathbf{x}_i$ are also highly risky to introduce artifacts in $\mathbf{x}_{i+1}$.
Partial Conditions refer to the conditions not containing all the necessary information to meet specific requirements. For example, if we follow Eq. (2) on the final view, we lack information about $\mathbf{x}_0$, making it impossible to generate image coherent with $\mathbf{x}_0$ to ensure loop closure. Additionally, using a single text prompt to generate all views within a full spherical panorama may lead to hallucinations, such as cities floating in the sky or underwater mountains.

3.2 Cross-View Guidance

To rectify the biased conditions discussed in Sec.3.1 and enlarge cross-view awareness, a natural idea is to let $\mathbf{x}_{i+1}$ conditioned on more views,

$$\mathbf{x}_{i+1} \sim q(\mathbf{x}|c_t, \mathbf{x}_i, \mathbf{P}_i^{i+1}, \mathbf{x}_1^g, ..., \mathbf{x}_m^g) \tag{3}$$

where $\mathbf{x}_1^g, ..., \mathbf{x}_m^g$ are selected from $\mathbf{x}_0, ..., \mathbf{x}_{i-1}$. This naturally results in a Guided Image Synthesis task form with self-generated images as guidance. Many existing methods can be adapted to implement our design, such as ControlNet [53] and T2I adapter [30]. To avoid relying on fine-tuning and reduce memory and time costs, we choose SDEdit [29], a training-free guided image synthesis approach, with a single guidance image $\mathbf{x}^g \in [\mathbf{x}_0, ..., \mathbf{x}_{i-1}]$.

Guided Image Synthesis using SDEdit. Given $\mathbf{x}^g$ as guidance, SDEdit establishes a Gaussian distribution using $\mathbf{x}^g$ as the expectation and the intermediate status at time t_0 in the reverse SDE process. The desired data distribution is obtained by gradually removing noise from $\mathbf{x}^g(t_0)$:

$$\mathbf{x}^g(t_0) \sim \mathcal{N}(\mathbf{x}^g, \sigma^2(t_0)\mathbf{I}); \quad \mathbf{x} \sim \text{SDEdit}(\mathbf{x}^g, t_0, \mathbf{\Phi}), \tag{4}$$

where $\mathbf{\Phi}$ denotes a generative model. In PanoFree, we use inpainting mask $\mathbf{m}_i$ to paste the guidance image $\mathbf{x}^g$ to the blank areas in the warped image $\hat{\mathbf{x}}_i$ and then use SDEdit in the inpainting process:

$$\hat{\mathbf{x}}_i^g = \mathbf{m}_i \cdot \mathbf{x}^g + (1 - \mathbf{m}_i) \cdot \hat{\mathbf{x}}_i; \quad \mathbf{x}_i \sim \text{SDEdit}(\hat{\mathbf{x}}_i^g, t_0, \mathbf{\Phi}_{inp}). \tag{5}$$

Since we only want to use additional guidance images to rectify the biased conditions rather than replicate the guidance image, we use $t_0 \in [0.9, 1.0)$ in practice. Meanwhile, we found that different selection of generation path and guidance image results in different generation qualities, and the optimal choice may vary for different tasks. We introduce a general selection effective for various tasks in Section 3.4 and provide an example of extending this technique to make scenes more realistic in specific scenarios in Section 3.5.

3.3 Risky Area Estimation and Erasing

To rectify the noisy conditions discussed in Sec.3.1 and eliminate the accumulation of artifacts, a natural idea is to detect and localize the artifact-inducing contents, and erase them. However, precise detection and localization often requires costly training. Thus, we turn to roughly estimate and erase the risky areas that are likely to contain artifact-inducing contents, based on indicators often associated with artifact: distances, color and smoothness. See Sec. A.2 in the supplementary for examples.

Risk Estimation based on Distances. We consider the distance from the center point of the initial view $\mathbf{x}_0$ and the distance to the edges. This is based on two priors: 1. The farther from the initial view, the more accumulated errors and the more likely to contain artifact-inducing contents. 2. Areas close to the edges are highly risky because truncated objects are mostly generated around the edges, and areas near the edges are often more severely distorted during warping. We use initial risk $\mathbf{r}_{init}$ to represent the risk estimated based on the distance from the initial view, and edge risk $\mathbf{r}_{edge}$ to represent the risk estimated based on the distance from edges. They are derived from the following:

$$\mathbf{r}_{init}(\mathbf{c}_i) = \mathcal{R}_p(\mathcal{D}_0(\mathbf{c}_i)); \quad \mathbf{r}_{edge}(\mathbf{c}_i) = \mathcal{R}_p(\mathcal{D}_\mathbf{e}(\mathbf{c}_i)), \tag{6}$$

where $\mathbf{c}_i$ represents the pixel coordinates of $\mathbf{x}_i$ within the panorama coordinate system, $\mathcal{D}_0$ measures the distance to the center point of the initial view along the generation path, $\mathcal{D}_\mathbf{e}$ measures the distance to all edges $\mathbf{e}$, and $\mathcal{R}_p$ is a scaling function. We use weighted euclidean distance for $\mathcal{D}_0$, Gaussian filters for $\mathcal{D}_\mathbf{e}$, and min-max normalization for $\mathcal{R}_p$.

Risk Estimation based on Color and Smoothness. After generating a view, we can predict the risk based on color and smoothness. This uses two priors: 1. Artifacts are often not smooth or distinct in color. 2. Salient areas with abrupt colors or unevenness are prone to causing artifacts. Color-based risk $\mathbf{r}_{color}$ and smoothness-based risk $\mathbf{r}_{smooth}$ are estimated in similar forms:

$$\mathbf{r}_{color}(\mathbf{x}_i) = \mathcal{R}_f(\mathcal{D}_c(\mathbf{x}_i)); \quad \mathbf{r}_{smooth}(\mathbf{x}_i) = \mathcal{R}_f(\mathcal{D}_s(\mathbf{x}_i)), \tag{7}$$

where $\mathcal{D}_c$ and $\mathcal{D}_s$ measures the abruptness of each pixel based on color and smoothness. When implementing them, we choose pixels with the same vertical coordinates across views, and calculate the "distances" of each pixel to the mean color and color gradient. Within $\mathcal{R}_f$, we applied Gaussian filtering after min-max normalization, as those estimated risks are usually noisy.

Erasing with Estimated Risks. With the estimated risks, we can erase the risky areas on the image warped to next view and the inpainting mask. Assume that we get inpainting mask for current view $\mathbf{m}_i$ and risks for previous view $\mathbf{r}_{i-1} = [\mathbf{r}^i_{i-1}, \mathbf{r}^e_{i-1}, \mathbf{r}^c_{i-1}, \mathbf{r}^s_{i-1}]$. The risks are combined linearly and new inpainting mask for current view can be obtained with:

$$\mathbf{m}^r_i = \mathcal{M}_r(\mathbf{m}_i, \mathcal{W}(\mathbf{r}_{i-1} \cdot \mathbf{w}, \mathbf{P}^i_{i-1})). \tag{8}$$

$\mathcal{M}_r$ is the risk-based remasking function, and $\mathbf{w}$ are user defined combination weights. We define $\mathcal{M}_r$ as thresholding the risk within the warped area.

Smoothing and Anti-aliasing. We note that the inpainting mask from risk-based erasing may not be smooth. Additionally, sharp and jagged edges on the inpainting mask can lead to artifacts. Therefore, we also employ fixed filtering $\mathcal{M}_f$, where Gaussian filtering and thresholding are used to smooth the mask and reduce sharp edges, while median filtering is used to reduce jagged edges. Then, we use the final inpainting mask for the combination with guidance, and the risky areas on the warped image are removed and regenerated.

$$\mathbf{m}^f_i = \mathcal{M}_f(\mathbf{m}^r_i); \quad \hat{\mathbf{x}}^g_i = \mathbf{m}^f_i \cdot \mathbf{x}^g + (1 - \mathbf{m}^f_i) \cdot \hat{\mathbf{x}}_i. \tag{9}$$

3.4 Bidirectional Generation with Symmetric Guidance

Bidirectional Generation. We begin by dividing a unidirectional generation path $\mathbf{x}_0 \rightarrow \mathbf{x}_1 ... \mathbf{x}_{2n}$ into two bidirectional generation paths $\mathbf{x}_0 \rightarrow \mathbf{x}_1 \rightarrow ... \rightarrow \mathbf{x}_n$ and $\mathbf{x}_{-n} \leftarrow ... \leftarrow \mathbf{x}_{-1} \leftarrow \mathbf{x}_0$. Typically, we would make these two generation paths symmetric. And we found this can reduce accumulated errors because the distance to the initial view is reduced in each direction. This consistently reduces artifacts, but may not reduce style and content inconsistency, as there may be different style/content shift in the two directions.

Loop Closure. To ensure loop closure, we can add a $(2n+1)$-th view as the "merging view" to merge the 2 generation paths by warping $\mathbf{x}_n$ and $\mathbf{x}_{-n}$ to the $(2n+1)$-th view and inpaint it. However, if the differences between the two paths are too large, $\mathbf{x}_{2n+1}$ may contain image tearing, failing to ensure loop closure. This is due to the partial conditions on each path: there is no information from the other path before merging. Therefore, we rectify the partial conditions by introducing awareness of the other path.

Symmetric Guidance. We introduce the awareness of the other path by selecting symmetric guidance images from the other path. Specifically, when generating $\mathbf{x}_{i+1}$, we will select $\mathbf{x}_{-i}$ as the guidance image. Thus, $\mathbf{x}_{i+1}$ will get the awareness of both paths as it is conditioned on $\mathbf{x}_i$ and $\mathbf{x}_{-i}$:

$$\mathbf{x}_{i+1} \sim \ q(\mathbf{x}|c_t, \mathbf{x}_i, \mathbf{x}_{-i}, \mathbf{P}_i^{i+1}) \tag{10}$$

We emperically found that bidirectional generation with symmetric guidance is not only effective in ensuring loop closure but also a universally applicable strategy to effectively reduce accumulated errors in various scenarios.

3.5 Aligning with Scene Structure Prior

When generating full spherical panoramas, we divide a spherical panorama into five parts: first, we generate a 360 panorama as the central part, then we expand upwards and downwards, and finally, we generate two images centered around the top and bottom poles to close up the entire spherical surface. During the expansion and closing stages, models often fail to align with scene structure priors due to partial conditions and generate artifacts.

Hallucination refers to the artifacts caused by mismatches between partial conditions and scene structure priors. For example, when generating a city scene, using the same prompt during the expansion and closing stages may result in a floating city in the sky or a city underwater. The most direct solution is to input a new prompt, but this would require additional manual effort, which is not ideal. So, we attempt to rectify the partial conditions by extracting scene structure priors and applying semantic and variance control from the initial view.

Prior Extraction. Although the pretrained T2I model may not align a full panorama with scene structure priors, it can align a single perspective view image with them. Therefore, we extract the scene structure prior from the initial view image $\mathbf{x}_0$ and incorporate it into the expansion process. For example, when generating the first view image in the upward expansion $\mathbf{x}_0^{ue}$, we use upper 1/3 part of the initial view image as guidance with resizing it to the size of $\mathbf{x}_0^{ue}$.

Semantic and Variance Tuning. When the give text prompt only describe part of the scene, we may want generated semantic contents less conditioned on the partial prompt and more conditioned on the prior images during expansion and closing. We achieve this by reducing guidance scale and widen the field of view. Meanwhile, we adjust the variance of the initial noise to avoid the color blocks caused by low guidance scale. Through experimentation, we've found that a combination of slightly high initial variance and low guidance scale can stably reduce hallucinations and color blocks during the expansion and closing stages.

4 Experiments

We evaluate the performance of PanoFree across three generation tasks: Planar Panorama Generation, 360 Panorama Generation, and Full Panorama Generation. However, note that we focus on planar panorama and 360° panorama generation, where the comparisons are more precise and consistent.

Implementation details. PanoFree is implemented using the publicly available Stable Diffusion code from Diffusers [34] based on the PyTorch framework. For the experiments in the main paper, we utilized the generation and inpainting models of Stable Diffusion (SD) v2.0 [38]. All experiments are conducted on a single NVIDIA RTX A6000 GPU. Further details and specific configurations can be found in the corresponding sections of the main paper and the supplementary.

Evaluation metrics. We introduce a more comprehensive set of evaluation metrics than prior work [1, 22, 46] covering five themes: image quality, global consistency, prompt capability, diversity, and resource consumption.

- *Image Quality* is measured with Fréchet Inception Distance (FID) [14], Kernel Inception Distance (KID) [2], which measure fidelity and diversity. FID and KID calculated between the views randomly cropped from the panorama and reference images generated by SD with the same prompts.
- *Global Consistency* is measured with Intra-LPIPS (IL) [55] used by SyncDiffusion [22], which is computed by cropping non-overlapping views from a panorama and computing the averaged LPIPS scores of all view pairs.
- *Prompt Capability* is measured via CLIP Score (CS) [13] by computing the text-image similarity of randomly cropped views of the panorama.
- *Panorama Diversity* is also measured by FID and KID. Additionally, we propose Cross-LPIPS (CS) [55]. Cross-LPIPS is computed across 2 panoramas generated with a same text with differents random seeds. We crop non-overlapping views from each panorama, and compute the averaged LPIPS scores of all view pairs where two views come from different panoramas.
- *Resource Consumption* includes time consumption, measured by the cumulative time cost of all diffusion processes to generate a single panorama, and peak GPU memory consumption, measured by the maximum GPU memory consumption during inference.

Evaluation Settings. Prior work either used arbitrary prompts [1, 22] or only focused on a single type of scene [46]. Instead, we consider 3 distinct scene types: indoor, street, and city scenes, and natural scenes. We obtained 100 prompts for each type from ChatGPT [35]. We use 10 random seeds per prompt for planar panorama and 360 panorama generation, and 3 different random seeds per prompt for full panorama generation (see supplementary for details).

User Study. For planar panorama generation and 360° panorama generation, we conducted four user studies for each task to further evaluate the global consistency, image quality, prompt compatibility, and diversity of the generated panoramas (see supplementary for details).

Fig. 3: Planar Panorama generation results. Compared to vanilla Sequential Generation and MultiDiffusion (MD) [1], PanoFree achieves superior global consistency and image quality. It is also comparable to SyncDiffusion (SYD) [22] in these aspects.

Table 1: Comparison of tuning free methods for Planar Panorama generation using Stable Diffusion [38]. We find PanoFree (PF) outperforms the state-of-the-art while having low computational requirements. Note that Cross-LPIPS and Intra-LPIPS are in 10^{-2} scale, KID is in 10^{-3} scale.

Method	Intra-LPIPS↓	Cross-LPIPS↑	FID↓	KID↓	CS↑	Time (s)↓	Memory (GB)↓
SG	70.40	71.03	24.91	4.33	26.68	25	3.2
MD [1]	68.48	69.92	21.16	3.50	27.89	95	5.8
SYD [22]	64.48	68.07	20.56	3.62	27.18	128	10.0
PF (Ours)	65.34	69.68	17.05	3.80	27.21	26	3.2

4.1 Planar Panorama Generation

Planar Panorama corresponds to the scene observed with camera translation along the focal plane in reality. This is a relatively simple task, as it only involves extending the image without considering more complex geometric changes.
Baselines. We have chosen 3 tuning-free baselines for comparison, *Vanilla Sequential Generation (SG)*, *MultiDiffusion (MD)* [1] and *SyncDiffusion (SYD)* [22]. Additional details are in the supplementary.
Results. The quantitative and qualitative evaluations are shown in Table 1 and Fig. 3, respectively. Below we compare PanoFree to each baseline.

- *Compared with vanilla Sequential Generation,* PanoFree significantly enhances image quality and global consistency, demonstrating its effectiveness in reducing accumulated errors. Moreover, PanoFree does not compromise diversity or have a significant effect on GPU time and memory overhead.
- *Compared with MultiDiffusion,* PanoFree has significant advantages in image quality and global consistency. Meanwhile, its time and GPU memory overhead is only 26% and 55% that of MultiDiffusion, respectively.
- *Compared with SyncDiffusion*, PanoFree achieves comparable performance in global consistency and image quality. Although SyncDiffusion performs better in consistency, it requires introducing additional models for latent optimization. This leads PanoFree's time overhead to be 20% of SyncDiffusion and GPU memory overhead to be 32% of SyncDiffusion.

Fig. 4: Diversity comparison on Planar Panorama Generation task. Each group is generated using the same text with different random seeds. Compared to MultiDiffusion (MD) [1] and SyncDiffusion (SYD) [22], PanoFree achieves superior diversity.

Table 2: User study results of Planar Panorama Generation. 15 questions are used for each evaluation item and answered by 5 Amazon MTurk workers.

	Consistency (%)	Quality (%)	Prompt Compatibility (%)	Diversity (%)
SYD [22]	52.7	46.0	41.3	35.3
PF (ours)	47.3	54.0	58.7	64.7

The Loss of Diversity with Joint Diffusion. When using different random seeds with the same prompt, methods using Joint Diffusion exhibit reduced diversity in their results. In contrast, our PanoFree method can better maintain diversity (see cross-LPIPS scores in Table 1). Additionally, we believe this is the source of PanoFree's gains over MultiDiffusion and SyncDiffusion in FID.

This diversity issue becomes particularly apparent when given some underspecified prompts. Therefore, we generated a "underspecified set" consisting of 20 short and blurry prompts to demonstrate this issue. For each prompt, we used 20 different random seeds. We demonstrate the diversity differences qualitatively in Fig. 4. Please refer to supplementary for quantitative analysis.

User Study. The results in Table 2 clearly show that human evaluators believe PanoFree produces more diverse panoramas and demonstrates better compatibility with prompts than SyncDiffusion [22]. Additionally, both methods exhibit similar levels of global consistency and image quality.

4.2 360 Panorama Generation

Due to the distortion caused by equirectangular projection, generating 360-degree panoramas is more challenging than planar panorama generation. Vanilla sequential generation tends to produce many artifacts, significantly decreasing image quality. Moreover, MultiDiffusion [1] and SyncDiffusion [22] cannot be directly used for generating 360-degree panoramas. As far as we know, PanoFree is the first implementation of training-free 360-degree panorama generation.

Baselines. We used 2 baselines for comparison: *Vanilla Sequential Generation (SG)* and *MVDiffusion (MVD)*. Additional details are in the supplementary.

Results. The quantitative and qualitative evaluations are shown in Table 3 and Fig. 5 respectively. Below we compare PanoFree to each baseline.

Table 3: Comparison of 360° Panorama generation methods using Stable Diffusion [38]. We find PanoFree (PF) still outperforms the state-of-the-art while having low computational requirements. Note that Cross-LPIPS and Intra-LPIPS are in 10^{-2} scale, KID is in 10^{-3} scale.

Method	Intra-LPIPS↓	Cross-LPIPS↑	FID↓	KID↓	CS↑	Time (s)↓	Memory (GB)↓
SG	70.62	73.06	32.28	7.90	26.35	21	3.2
MVD [46]	67.71	70.07	37.89	8.76	26.27	110	6.9
PF (ours)	68.62	72.67	25.84	7.48	26.51	22	3.2

Fig. 5: 360° Panorama generation results. Compared to vanilla Sequential Generation (SG), PanoFree achieves superior global consistency and image quality. It is also comparable to MVDiffusion (MVD) [46] in these aspects.

- *Compared with vanilla Sequential Generation,* PanoFree significantly enhances image quality and global consistency. Specifically, vanilla sequential generation creates artifacts with complex optical geometry transformations, severely impacting image quality. However, PanoFree effectively recovers image quality by estimating and erasing risky areas, minimizing artifact propagation.
- *Compared with MVDiffusion,* PanoFree achieves comparability in image quality and global consistency, yet significantly outperforms in terms of time, GPU memory overhead, and diversity. Particularly, MVDiffusion is significantly worse than PanoFree in terms of FID and KID scores, even underperforming vanilla Sequential Generation. This is partly due to the inevitable bias of MVDiffusion's generated results towards the training dataset, resulting in larger discrepancies compared to those produced by Stable Diffusion. Visually, MVDiffusion also exhibits a noticeable lack of generation diversity. As depicted in Fig. 6, given a prompt, results generated with different random seeds show minimal variation in both content and style.

User Study. The user study results in Table 4 show that human evaluators believe PanoFree also produces more diverse 360° panoramas compared with MVDiffusion [46]. And PanoFree demonstrates better global consistency. Both methods exhibit similar levels of image quality and prompt comparability.

4.3 Full Spherical Panorama Generation

PanoFree is also the first to achieve feasible tuning-free generation for Full Spherical Panoramas. However, it's hard to conduct meaningful comparisons due the

Fig. 6: Diversity comparison on 360° Panorama Generation task. Each group is generated using the same text with different random seeds. Compared to MVDiffusion (MVD) [46], PanoFree achieves superior diversity.

Table 4: User study results of 360° Panorama Generation. 15 questions are used for each evaluation item and answered by 5 Amazon MTurk workers.

	Consistency (%)	Quality (%)	Prompt Compatibility (%)	Diversity (%)
MVD [46]	40.7	48.7	47.3	33.3
PF (ours)	59.3	51.3	52.6	66.6

lack of tuning-free generation baseline methods or those with strong out-of-scope generation capabilities for Full Spherical Panoramas generation task. Thus, we conduct qualitative evaluation as well as comparison with vanilla sequential generation by showcasing generated results in Fig. 7. Vanilla sequential generation exhibit more artifacts as distortion increases. Additionally, partial conditioning issue mentioned in Sec. 3.5 causes hallucinations. And PanoFree still could effectively reduce artifacts and hallucinations. Note that we start both methods from 360° panoramas generated by PanoFree into full spherical panoramas, otherwise vanilla sequential generation will perform even worse.

4.4 Ablation Study

Tab. 5 contains an ablation study that sequentially integrates each PanoFree component. We evaluate consistency (Intra-LPIPS) and image quality (FID) with 30% of the prompts from Sec. 4.1 & Sec. 4.2 for both planar and 360° panorama. We show that cross-view guidance provides the strongest benefit, followed by distance and edge-based risky area erasing. These components effectively reduce image tearing and visual chaos. Color and smoothness-based erasing have a smaller impact, likely due noise in these low-level features. Qualitative results are in Sec. B of the supplementary.

5 Conclusion

We present PanoFree, a tuning-free multi-view image generation that supports an extensive array of correspondences. PanoFree improves error accumulation

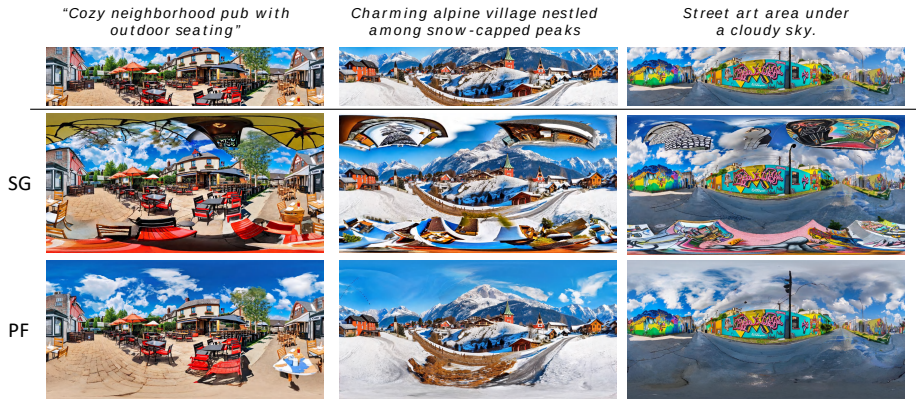


Fig. 7: Full Spherical Panorama generation results. Vanilla Sequential Generation (SG) tends to generate hallucinations due to partial conditions, while PanoFree effectively mitigates this issue.

Table 5: Quantitative ablation of PanoFree components using 30% of prompts from Sec. 4.1 & 4.2. Intra-LPIPS (10^{-2}) and FID show cross-view guidance offers the most benefit, followed by distance and edge-based risky area erasing. Color and smoothness-based erasing have minimal impact.

Task	Method	Intra-LPIPS↓	FID↓
Planar	None	70.84	24.75
	+ CG	66.63	20.63
	+ Dist	65.87	18.21
360°	None	71.34	33.47
	+ CG	69.21	27.38
	+ Dist & Edge	69.03	26.69
	+ Color & Smooth	68.94	26.51

by enhancing cross-view awareness and refining the warping and inpainting processes through cross-view guidance, risky area estimation and erasing, and symmetric bidirectional guided generation for loop closure, alongside guidance-based semantic and density control for scene structure preservation. PanoFree is evaluated on various panorama types—Planar, 360°, and Full Spherical Panoramas. PanoFree demonstrates significant error reduction, improved global consistency, and image quality across different scenarios without extra fine-tuning. Compared to existing methods, PanoFree is up to 5x more efficient in time and 3x more efficient in GPU memory usage, and maintains superior diversity of results (2x better in our user study). Moreover, PanoFree can be extended to texture generation for 3D models. We intend to explore these possibilities in future research.

Limitations. A limitation of our work is that we are unable to generate scenes beyond the capability of the pre-trained T2I model. Therefore, we rely on large pre-trained T2I models to ensure the broad application scope. And when provided with text descriptions beyond the capability range of the pre-trained T2I models, the generated results may not match the text.

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
2. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. ArXiv **abs/1801.01401** (2018), https://api.semanticscholar.org/CorpusID:3531856
3. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
4. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. arXiv preprint arXiv:2303.11396 (2023)
5. Chen, Z., Wang, G., Liu, Z.: Text2light: Zero-shot text-driven hdr panorama generation. ACM Transactions on Graphics (TOG) **41**(6), 1–16 (2022)
6. Cheng, Y.C., Lin, C.H., Lee, H.Y., Ren, J., Tulyakov, S., Yang, M.H.: Inout: Diverse image outpainting via gan inversion. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11421–11430 (2021), https://api.semanticscholar.org/CorpusID:232478397
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
9. Fang, C., Hu, X., Luo, K., Tan, P.: Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. arXiv preprint arXiv:2310.03602 (2023)
10. Feng, M., Liu, J., Cui, M., Xie, X.: Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. ArXiv **abs/2311.13141** (2023), https://api.semanticscholar.org/CorpusID:265351889
11. Fridman, R., Abecasis, A., Kasten, Y., Dekel, T.: Scenescape: Text-driven consistent scene generation. arXiv preprint arXiv:2302.01133 (2023)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
13. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. ArXiv **abs/2104.08718** (2021), https://api.semanticscholar.org/CorpusID:233296711
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Neural Information Processing Systems (2017), https://api.semanticscholar.org/CorpusID:326772
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) (2020)
16. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**(47), 1–33 (2022)
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
18. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. arXiv preprint arXiv:2303.11989 (2023)

19. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. arXiv preprint arXiv:2206.00364 (2022)
20. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. arXiv preprint arXiv:2206.00364 (2022)
21. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34**, 852–863 (2021)
22. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via synchronized joint diffusions. Advances in Neural Information Processing Systems **36** (2024)
23. Li, J., Bansal, M.: Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. ArXiv **abs/2305.19195** (2023), https://api.semanticscholar.org/CorpusID:258967291
24. Lin, C.H., Chang, C.C., Chen, Y.S., Juan, D.C., Wei, W., Chen, H.T.: Cocogan: Generation by parts via conditional coordinating. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4511–4520 (2019), https://api.semanticscholar.org/CorpusID:90262507
25. Lin, C.H., Lee, H.Y., Cheng, Y.C., Tulyakov, S., Yang, M.H.: Infinitygan: Towards infinite-pixel image synthesis. In: International Conference on Learning Representations (2021), https://api.semanticscholar.org/CorpusID:238419701
26. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In: The International Conference on Learning Representations (2023)
27. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. arXiv preprint arXiv:2206.00927 (2022)
28. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: Repaint: Inpainting using denoising diffusion probabilistic models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11451–11461 (2022), https://api.semanticscholar.org/CorpusID:246240274
29. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2021), https://api.semanticscholar.org/CorpusID:245704504
30. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. ArXiv **abs/2302.08453** (2023), https://api.semanticscholar.org/CorpusID:256900833
31. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
32. Oh, C.H., Cho, W., Park, D., Chae, Y., Wang, L., Yoon, K.J.: Bips: Bimodal indoor panorama synthesis via residual depth-aided adversarial learning. ArXiv **abs/2112.06179** (2021), https://api.semanticscholar.org/CorpusID:245123664
33. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: International Conference on Computer Vision. pp. 4195–4205 (2023)
34. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022)

35. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Language models are unsupervised multitask learners (2019)
36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
37. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems (NeurIPS) (2022)
40. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. PMLR (2015)
41. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
42. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems (NeurIPS) (2019)
43. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
44. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems **33**, 12438–12448 (2020)
45. Song, Y., Sohl-Dickstein, J.N., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. ICLR (2021)
46. Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. ArXiv **abs/2307.01097** (2023), `https://api.semanticscholar.org/CorpusID:259316427`
47. Teterwak, P., Sarna, A., Krishnan, D., Maschinot, A., Belanger, D., Liu, C., Freeman, W.T.: Boundless: Generative adversarial networks for image extension. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10520–10529 (2019), `https://api.semanticscholar.org/CorpusID:201106503`
48. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
49. Voynov, A., Hertz, A., Arar, M., Fruchter, S., Cohen-Or, D.: Anylens: A generative diffusion model with any rendering lens (2023)
50. Wang, G., Yang, Y., Loy, C.C., Liu, Z.: Stylelight: Hdr panorama generation for lighting estimation and editing. In: European Conference on Computer Vision (2022), `https://api.semanticscholar.org/CorpusID:251196614`
51. Wang, H., Xiang, X., Fan, Y., Xue, J.H.: Customizing 360-degree panoramas through text-to-image diffusion models. ArXiv **abs/2310.18840** (2023), `https://api.semanticscholar.org/CorpusID:264590753`
52. Wu, T., Zheng, C., Cham, T.J.: Panodiffusion: 360-degree panorama outpainting via diffusion (2023), `https://api.semanticscholar.org/CorpusID:259360663`
53. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)

54. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10188–10198 (2023), https://api.semanticscholar.org/CorpusID:257834007
55. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 586–595 (2018), https://api.semanticscholar.org/CorpusID:4766599