

# TwinDiffusion: Enhancing Coherence and Efficiency in Panoramic Image Generation with Diffusion Models

Teng Zhou<sup>a</sup> and Yongchuan Tang<sup>a,\*</sup>

<sup>a</sup>College of Computer Science and Technology, Zhejiang University

**Abstract.** Diffusion models have emerged as effective tools for generating diverse and high-quality content. However, their capability in high-resolution image generation, particularly for panoramic images, still faces challenges such as visible seams and incoherent transitions. In this paper, we propose TwinDiffusion, an optimized framework designed to address these challenges through two key innovations: the Crop Fusion for quality enhancement and the Cross Sampling for efficiency optimization. We introduce a training-free optimizing stage to refine the similarity of adjacent image areas, as well as an interleaving sampling strategy to yield dynamic patches during the cropping process. A comprehensive evaluation is conducted to compare TwinDiffusion with the prior works, considering factors including coherence, fidelity, compatibility, and efficiency. The results demonstrate the superior performance of our approach in generating seamless and coherent panoramas, setting a new standard in quality and efficiency for panoramic image generation.

## 1 Introduction

Over the past few years, diffusion models [15, 25, 36] have demonstrated their creativity in generation tasks. They define a pair of forward and reverse Markov chains to learn the data distribution, which bypasses the limitations in other types of generative models like GANs [4, 7], VAEs [6, 21] and Flows [18]. Diffusion models have risen as effective tools for broad applications, spanning from high-quality images to multi-type content creation. With the growing interest in full information records, immersive virtual reality, artistic expression, and historical preservation, the synthesis of long scrolls is gaining more and more attention, particularly in panoramic image generation tasks [2, 11, 38].

Recent advancements have exhibited the expansibility of pre-trained diffusion models in generating images of arbitrary dimensions, such as super-resolution diffusion [13, 34] and area fusion strategies [17, 35, 39]. The latter often involves cropping from a large space into small patches for individual processing, as well as conducting specific guidance to fuse them together, providing more controllability than the former. However, achieving crop-wise high-resolution generation is non-trivial. To our knowledge, MultiDiffusion [2] represents the state-of-the-art framework among the existing methods, yet it still fails to capture the relationships between neighboring image areas, resulting in unnatural connections or even visible seams in panoramas. Although a finer cropping stride could ease such problems, it comes with a higher time cost.

To tackle these challenges, we propose TwinDiffusion, an optimized framework designed to enhance the capability of panoramic image generation with diffusion models. Drawing from the groundwork laid by the MultiDiffusion, our approach introduces two key innovations to make improvements in both quality and efficiency.

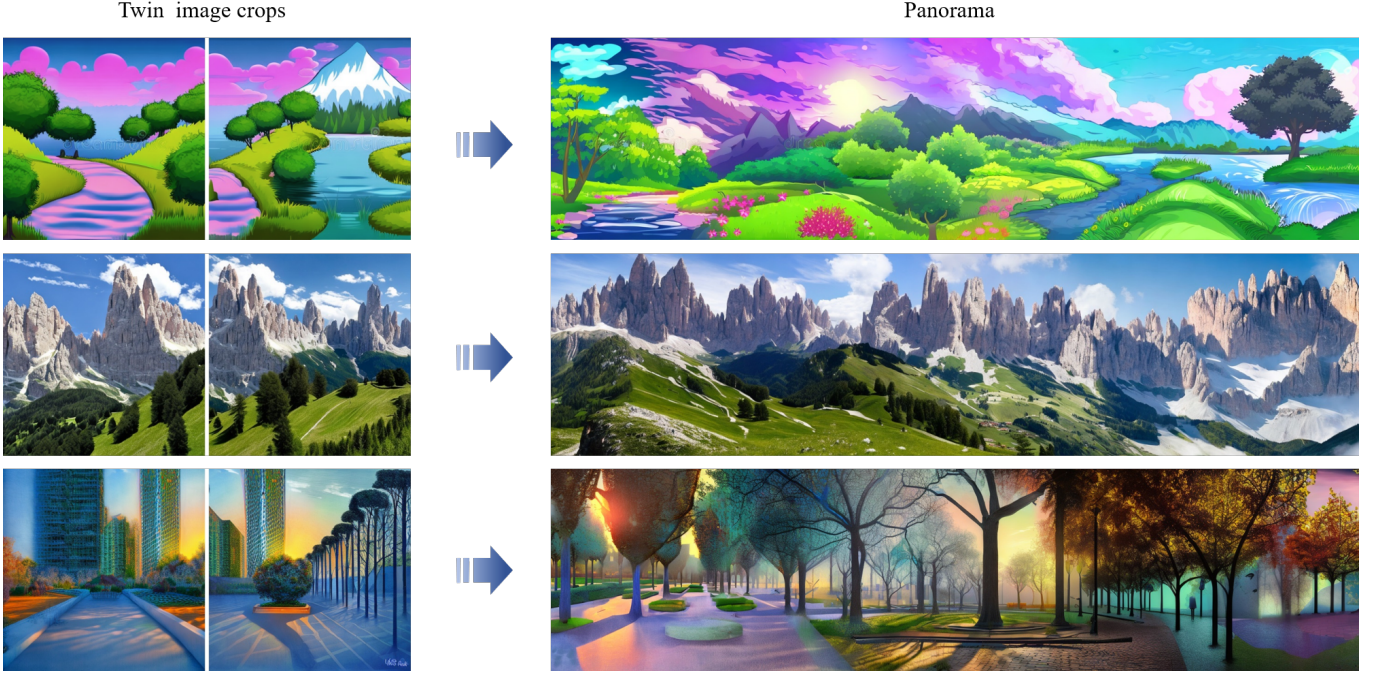
- **(Quality) Crop Fusion:** Our first innovation in TwinDiffusion focuses on refining the coherence of generated panoramas by introducing a training-free optimizing stage. Inspired by the harmonious relationship between twins, this approach is aimed to closely align the adjacent parts of the panoramic image space, leading to smoother transitions and fewer seams in final panoramas as shown in Fig. 1.
- **(Efficiency) Cross Sampling:** Our second innovation handles the efficiency of generating panoramas by adopting an interleaving sampling method. With a group of dynamic strides in the cropping process, we effectively mitigate the loss of image quality caused by larger cropping strides, enabling faster generation while upholding the sampling quality.
- **Performance Trade-off:** Moreover, we analyze the key factors in TwinDiffusion that impact its performance, including the timestep for introducing the crop optimizing stage, the Lagrange multiplier in our core function, the view stride and the cross stride for sampling image patches. This thorough discussion gives insights into the condition of quality-efficiency balance with our method.

Lastly, our comprehensive evaluation of TwinDiffusion compares its performance with baselines in a range of aspects including coherence (measured by LPIPS [40] & DISTS [8]), diversity (FID [14] & IS [27]), compatibility (CLIP [23] & CLIP-aesthetic [28]), efficiency (processing time), etc. Qualitatively, we demonstrate its effectiveness and stability in eliminating seams and generating smoother panoramas. Quantitatively, our method outperforms other baselines across all evaluation metrics, striking a new balance in quality and efficiency for panoramic image generation.

## 2 Related Work

**Diffusion Models** Diffusion models are inspired by non-equilibrium thermodynamics [29]. They define two Markov chains for forward and reverse processes, namely diffusion and denoising. The forward process is to perturb a data distribution  $x_0 \sim q(x)$  into a standard Gaussian distribution  $x_T \sim \mathcal{N}(0, I)$  with  $T$  steps of noise injection. This process is reversed to recreate the sample  $x_0$  that obeys the original data distribution from a Gaussian noise input.

\* Corresponding Author. Email: yctang@zju.edu.cn



**Figure 1.** TwinDiffusion is a crop-wise framework designed for high-resolution panorama generation with diffusion models. Inspired by the strong connection between twins, our approach aims to reconcile adjacent areas of the panoramic image space successively. This alignment produces pairs of locally similar image crops resembling twins (left), leading to improved coherence and smoother transitions in panoramas (right).

Specifically, the reverse process involves training a network to approximate  $q(x_{t-1} | x_t)$ , and then sampling from  $\mathcal{N}(0, I)$  iteratively with the trained model. From DDPM [15] to DDIM [30] to LDM [25], diffusion models have paved the way for text-to-image generation, boosting AI-painting applications like Stable Diffusion [25] and DALLE2 [24]. By capturing the spatial-temporal distribution features, they also show promising prospects in generating multi-modal contents such as videos [3, 5, 10], audio [16, 19], and 3D objects [31, 33, 37].

**High-Resolution Image Generation with Diffusion Models** Extensive studies have been dedicated to leveraging diffusion models for controllable high-resolution image generation tasks. The existing methodologies can be divided into two branches: (i) methods that focus on super-resolution [13, 34] or inpainting [1] techniques utilizing diffusion processes to infer the missing information, and (ii) methods that center around crop-based fusion strategies within diffusion paths [2, 17, 39]. The former often requires training on specific datasets, combining initial noise with low-resolution images as input to the network. Additionally, they involve resizing the input images, which lacks portability and imposes computational demands. On the other hand, the latter offers greater flexibility by manipulating the generation process in different cropped spaces, and reconciling them in a training-free or fine-tuning manner. Among them, the MultiDiffusion framework proves to be feasible and resultful. However, its optimization function only pays attention to the overlapping regions of image crops, ignoring the non-overlapping adjacent areas, which reduces itself into a naive weighted mean method. Hence, we see room for improvement.

**Faster Sampling Method for Diffusion Models** Traditional sampling methods in diffusion require a large number of iterations to generate high-resolution images, which can strain computational resources and slow down operations [12, 32]. Alongside the scheduler optimization [9, 30] and diffusion distillation [26], we refer to the

trajectory stitching and interlaced sampling method [20] for our sampling acceleration request.

### 3 Method

#### 3.1 Preliminary

To start with, we introduce a pre-trained diffusion model denoted by  $\Phi$ , operating in a latent space  $Z = \mathbb{R}^{h \times w \times c}$  and a textual condition  $C$ . Employing the deterministic DDIM sampling [30]:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left( \sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \Phi(z_t, t, C) \quad (1)$$

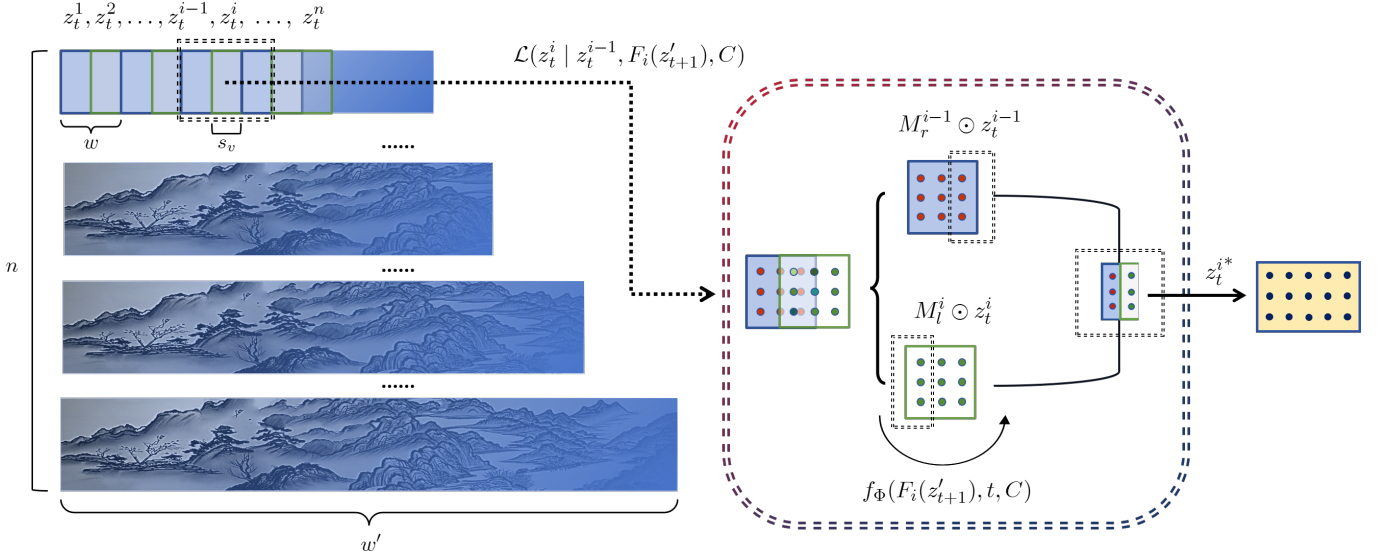
where  $z_t \in Z$  and  $\alpha_t$  is parameterized by the DDIM schedule  $\{\beta_i | i = 1, 2, \dots, T, \beta_i \in (0, 1)\}$ , we get image  $z_0$  from initial Gaussian noise  $z_T$  after  $T$  steps of denoising.

Our intention is to extend  $\Phi$  as a reference model to generate images in a larger space  $Z' = \mathbb{R}^{h' \times w' \times c}$ , where  $h' > h$  and  $w' > w$ . This can be achieved with the MultiDiffusion framework [2], represented by a function called the MultiDiffuser  $\Psi$ . It defines a set of mappings between two model spaces by:

$$z_t^i = F_i(z_t') \quad (2)$$

Specifically,  $F_i$  refers to cropping the  $i$ -th image patch from space  $Z'$  with the stride of  $s_v$ .

During the MultiDiffusion process, firstly, each crop is simultaneously and independently denoised with  $f_\Phi: z_{t-1}^i = f_\Phi(z_t^i, t, C)$  suggested by Eq. 1. Then, based on the Manifold Hypothesis, a least-square optimization for global fusion is formulated to minimize the discrepancy between each crop  $F_i(z_{t-1}')$  and its denoised counterpart  $f_\Phi(F_i(z_t'), t, C)$ , merging different crops into one large image  $z'$ . According to the properties of  $\Psi$ , its optimization problem has an



**Figure 2.** Illustration of our approach applied to panorama generation. The process begins with the mapping function  $F_i$  transforming the image crops into the panoramic space  $Z'$ . This results in a sequence of overlapping crops  $z_t^1, z_t^2, \dots, z_t^n$  arranged spatially, each having an independent denoising path. Our goal is to optimize  $z_t^i$  within the constraints of its adjacent neighbor and itself as well, thus ensuring a unified and progressive fusion of crops. To achieve this alignment, our objective function Eq. 4 is defined into two mutual-restricted parts and reaches the minimizer  $z_t^{i*}$  in each denoising timestep: (i) the matching term: differences at the overlaps of  $z_t^i$  and its neighbor  $z_t^{i-1}$ , (ii) the regularization term: deviations between  $z_t^{i*}$  and its unoptimized self  $f_\Phi(F_i(z_{t+1}'), t, C)$ .

analytical solution. Thus, the minimizer  $z'$  turns out to be a weighted average value:

$$z'_{t-1} = \frac{\sum_i W_i \odot F_i^{-1}(f_\Phi(z_t^i, t, C))}{\sum_i W_i} \quad (3)$$

with  $W_i$  represents the pixel weight matrix of the  $i$ -th crop.

### 3.2 Image Crop Fusion

As Eq. 3 suggests, the optimization process in MultiDiffusion is only taken for where each  $F_i(z_t')$  overlaps, disregarding the adjacent but non-overlapping subareas, which can always disrupt the overall coherence of images. The key idea of our Crop Fusion method is to reconstruct this core function, enabling a more reasonable and high-quality panorama generation.

At each denoising timestep  $t$ , we get a spatial-ordered sequence of overlapping crops  $z_t^1, z_t^2, \dots, z_t^n$  generated by the mapping function Eq. 2 in a panoramic space  $Z'$ . For each crop  $z_t^i$ , our goal is to align its overlapping part of the adjacent crop  $z_t^{i-1}$  as closely as possible, while limiting the deviation from the crop itself. Thus we present:

$$\begin{aligned} z_t^{i*} &= \arg \min_{z_t^i \in Z'} \mathcal{L}(z_t^i | z_t^{i-1}, F_i(z_{t+1}'), C) \\ &= \arg \min_{z_t^i \in Z'} \| M_r^{i-1} \odot z_t^{i-1} - M_l^i \odot z_t^i \|^2 + \\ &\quad \| f_\Phi(F_i(z_{t+1}'), t, C) - z_t^i \|^2 \end{aligned} \quad (4)$$

as our optimization task, where  $z^*$  denotes the optimized crop,  $M_l, M_r \in \{0, 1\}^{h \times w}$  represent the binary masks covering the crop's left and right overlapping regions according to  $s_v$ , and  $\odot$  is the Hadamard product. The second part of Eq. 4 serves as a regularization term, which is used to coordinate the alignment behavior of crops. Therefore, the objective function of TwinDiffusion can be formulated as follows:

$$\begin{aligned} \min \quad & \| M_r^{i-1} \odot z_t^{i-1} - M_l^i \odot z_t^i \|^2 \\ \text{s.t.} \quad & \| f_\Phi(F_i(z_{t+1}'), t, C) - z_t^i \|^2 \leq \aleph \end{aligned} \quad (5)$$

and the Lagrangian function for this problem is given by:

$$L(z_t^i, \lambda) = \| M_r^{i-1} \odot z_t^{i-1} - M_l^i \odot z_t^i \|^2 + \lambda (\| f_\Phi(F_i(z_{t+1}'), t, C) - z_t^i \|^2 - \aleph) \quad (6)$$

where  $\lambda$  is the Lagrange multiplier associated with the constraints of adjacent but non-overlapping regions in the panorama. Using the Karush-Kuhn-Tucker (KKT) conditions, we can reach an optimal solution:

$$z_t^{i*} = \begin{cases} M_l^i \odot z_t^{i*} = (1 + \lambda)^{-1} [M_r^{i-1} \odot z_t^{i-1} + \lambda M_l^i \odot f_\Phi(F_i(z_{t+1}'), t, C)] \\ M_r^i \odot z_t^{i*} = M_r^i \odot f_\Phi(F_i(z_{t+1}'), t, C) \end{cases} \quad (7)$$

which demonstrates that our Crop Fusion is a training-free method with closed-form optimization.

As depicted in Fig. 2, our framework progressively optimizes image crops while considering their coherence in multiple subregions and achieving a unified fusion. This approach fundamentally differs from MultiDiffusion, which performs a single weighted average across the entire panorama space.

### 3.3 Cross Sampling

In Eq. 2,  $F_i$  specifies a sliding window to crop overlapping images with a fixed step size  $s_v$ , referred to as the view stride later. We notice that the quality and generation speed of panoramic images heavily depend on this view stride, which determines the degree of overlap between crops. A finer degree of overlap results in superior panoramas, yet processing numerous image crops during denoising iterations can be time-consuming.

To address this quality-efficiency trade-off, we propose a variant mapping function called Cross Sampling defined by:

$$z_t^i = F_i^{(k)}(z_t'), \quad k = t \bmod r \quad (8)$$



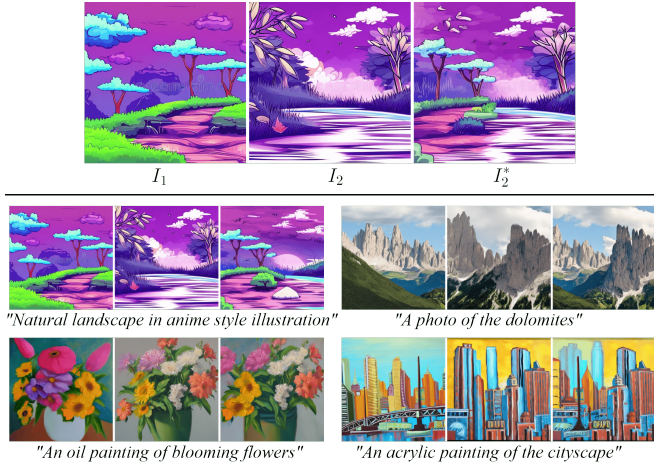
where  $r$  controls the interleaving frequency and  $k$  denotes the sampling mode. Staggering in  $r$  times, our method dynamically forms a set  $\mathcal{Z} = \{z_{i,j} \mid 1 \leq i \leq r, 1 \leq j \leq n\}$  consists of  $r$  groups, each containing  $n$  overlapping crops, with spatial locations incrementing by the cross stride  $s_r$ . Then, in  $T$  rounds of denoising, we alternate between using these crop groups for sampling in  $r$  staggered spaces.

Panorama seams mostly occur where the crops meet. Thus by incorporating the Cross Sampling method, TwinDiffusion takes more flexible control over the fine degree of overlap between neighboring image crops, filling the gaps caused by the enlarged view stride. Our experimental results in Sec. 4.3 have demonstrated that with this straightforward yet effective solution, we can double  $s_v$  to cut the generation time in half or more (with a larger  $s_v$ ), while maintaining panoramic image quality on par with the original MultiDiffusion.

## 4 Experiments

### 4.1 Panorama with Twin Crops

Here, we report two successive implementations of TwinDiffusion: the Single form, generating twin images with high and controllable similarity, and the Multiple form, extending its capabilities to synthesize panoramas composed of multiple, harmoniously interconnected twin crops.



**Figure 3.** Applying our Crop Fusion method to generate twin images. Top: the optimized  $I_2^*$  exhibits a seamless fusion effect that meets our expectations. Bottom: We further test its limits by fixing the regularization term of Eq. 5. The results demonstrate our method’s robustness under extreme conditions.

**Single: Twin Images** As illustrated in Fig. 3 (top), our method generates a pair of images that locally resemble each other like twins.  $I_1$  and  $I_2$  represent the first and second images respectively, generated from initial latent noise satisfying  $M_r^1 \odot z_1 = M_r^2 \odot z_2$ .  $I_2^*$  corresponds to  $I_2$  with our optimization. It successfully retains the content of original  $[I_2]_r$  while closely aligning its left part to  $[I_1]_r$ , achieving a seamless fusion of  $[I_1]_r$  and  $[I_2]_l$ .

We also conduct a stress test on TwinDiffusion’s ability to fuse image crops. Specifically, we replace  $f_\Phi(F_i(z'_{t+1}), t, C)$  in the regularization term of Eq. 5 with a constant reference  $\tilde{z}_t^i$ , representing the raw  $z_t^i$  following its unoptimized denoising trajectory. The results in Fig. 3 (bottom) demonstrate that even under such conditions, our method still achieves a natural and seamless crop fusion. This work serves as an initial validation to ensure higher consistency in the subsequent panorama generation.

**Multiple: Panorama with Twin Crops** Beyond a single pair of images, we generalize this approach to a sequence of images, i.e., crops  $z_t^1, z_t^2, \dots, z_t^n$  within panoramas. Our optimization is applied to each pair of neighboring crops, promoting a high degree of similarity between  $[z_t^i]_r$  and  $[z_t^{i+1}]_l$ , thereby resulting in higher-quality panoramas with consecutive twin crops as depicted in Fig. 1. Refer to Appendix A for more implementation examples.

### 4.2 Comparison

We conduct a comprehensive evaluation of our approach from both qualitative and quantitative perspectives, comparing images generated by TwinDiffusion versus other baselines.

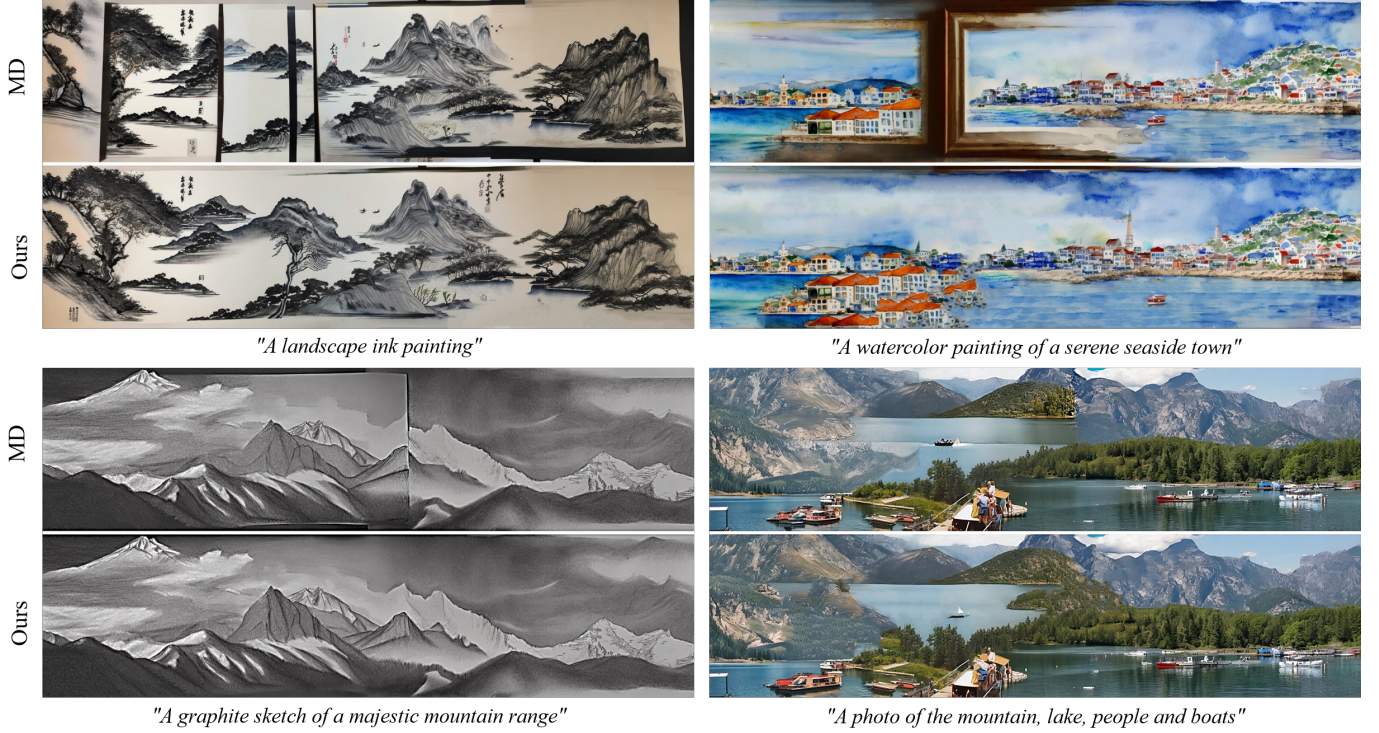
For the reference model  $\Phi$ , we employ two variants: the widely used diffusion model Stable Diffusion v2.0, and its advanced version Stable Diffusion XL v1.0 [22]. They respectively operate in an image space of  $\mathbb{R}^{512 \times 512 \times 3}$  and  $\mathbb{R}^{1024 \times 1024 \times 3}$ . We align the size of crops with the default resolution of  $\Phi$ , creating panoramas in  $512 \times 2048$  and  $1024 \times 4096$  correspondingly. To ensure the reliability of our results, we test 20 different prompts involving various contents and art styles, and generate 200 panoramas per prompt with 5 sets of random seeds. The results presented in Sec. 4.2 are specifically obtained with a reference model of  $\Phi = \text{SD}_{2.0}$ , a crop fusion timestep of  $\tau = T/2$ , an adjacent control factor of  $\lambda = 1$ , a view stride of  $s_v = 16$  and a cross stride of  $s_r = 8$ . More details and comparisons about these contributing factors are thoroughly discussed in Sec. 4.3 and Appendix C.

**Qualitative Comparison** In Fig. 4, we showcase the comparative performance between our method and MultiDiffusion across a series of qualitative examples. Our TwinDiffusion effectively mitigates the problem of visible seams at the overlaps of image crops, achieving a smoother transition where MultiDiffusion tends to struggle. As shown in the first three cases, this improvement is particularly noticeable for art painting, which differs from the natural landscape due to its frame-like incoherence that often arises at the edges of crops. More qualitative results are provided in Appendix B.

**Quantitative Comparison** We utilize a range of quantitative metrics focusing on the following four aspects: (i) coherence at the intersection of crops, (ii) fidelity and diversity of the generated panoramas, (iii) compatibility with the input prompts, as well as (ix) efficiency of the optimization process. Recognizing that resizing the entire panoramic image to meet the small dimensions required by the metrics (e.g.  $299^2$  for FID,  $224^2$  for CLIP) could lead to loss of essential features and distortions, we choose to test with images cropped from panoramas at a  $512^2$  resolution instead.

- **(Coherence) Learned Perceptual Image Patch Similarity (LPIPS) and Deep Image Structure and Texture Similarity (DISTS):** LPIPS and DISTS capture the perceptual differences between two images by computing distances of their feature vectors. Each generated panorama is divided into 8 pairs of adjacent but non-overlapping image crops according to the  $s_v$ . From these cropped views, we randomly take 4,000 pairs to compute LPIPS and DISTS values.
- **(Fidelity & Diversity) Fréchet Inception Distance (FID) and Inception Score (IS):** Leveraging the underlying output of Inception V3 network, FID and IS describe both fidelity and diversity of generated images. Inspired by [2], we measure FID and IS between the distribution of generated and reference image sets, where the former consists of images cropped from panoramas,





**Figure 4.** Qualitative comparisons between MultiDiffusion and ours. Our approach significantly reduces the odd joints and visible seams that commonly occur in MultiDiffusion, resulting in higher-quality panoramic images.

and the latter comprises images generated by reference models. To avoid coherence interfering with diversity, we extract only one random crop from each panorama, and calculate the metrics from each crop to the reference image set.

- **(Compatibility) CLIP and CLIP-aesthetic:** CLIP is used to assess the cosine similarity between generated images and the input prompts, while CLIP-aesthetic score is predicted from a linear estimator on top of CLIP. For a given prompt, we employ the cropped image set mentioned above to compute the scores.
- **(Efficiency) Generation Time:** The time taken to generate a complete panoramic image is evaluated on an A100 GPU.

We make comparisons among the Blended Latent Diffusion [1], MultiDiffusion and our TwinDiffusion, calculating the mean and standard deviation scores of all metrics. Additionally, we measure the performance of  $\Phi$  itself (i.e. the Stable Diffusion), which is measured by internal comparisons within the reference image set.

As reflected in Fig. 8, our approach stands out as the optimal method across all evaluation criteria. Coherence, the most important aspect of panoramic images, is greatly improved by our method, as reported in the first row. Additional progress can also be observed in the CLIP-aesthetic metric, implying that the increase in coherence could facilitate the compatibility and aesthetic appeal of generated results. Meanwhile, we get comparable scores in FID and IS. This keeps in line with our method’s primary focus on the seam issue, which may not have much impact on fidelity and diversity. In terms of efficiency, our method achieves better image quality without any compromise in time cost.

### 4.3 Ablation

TwinDiffusion incorporates several key factors that contribute to its performance, including the timestep  $\tau$  for introducing the Crop Fusion stage, the adjacent control factor  $\lambda$  in our optimization function,

the view stride  $s_v$  for cropping image patches, and the cross stride  $s_r$  in the Cross Sampling method.

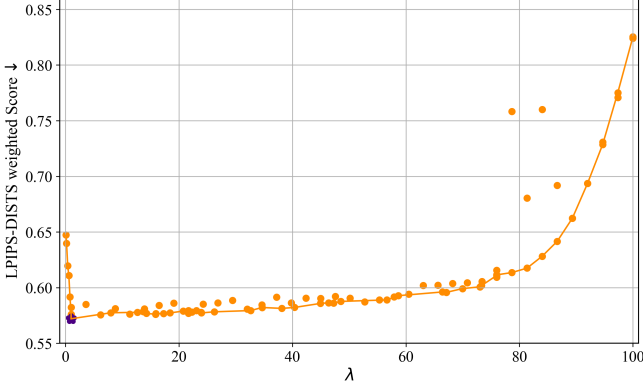


**Figure 5.** The analysis of the proper timestep  $\tau$  for introducing the Crop Fusion stage. As  $\tau$  decreases, a gradual transition from under-optimization to over-optimization can be observed, with the best results attained by  $\tau = T/2$ . The bottom rows of the figure offer two additional examples that further support our findings.

**Optimization Timestep** In the initial attempts at twin-image generation, we apply the Crop Fusion method throughout the entire denoising process. However, the outcomes are unsatisfactory as the generated images exhibit a distinct left-right fragmentation, where an appropriate optimization time window is needed to achieve the desirable results. We know that an earlier guidance plays a greater role in the diffusion trajectory. Thus, we decide to confine the optimization period to the early stages of the sampling process. That

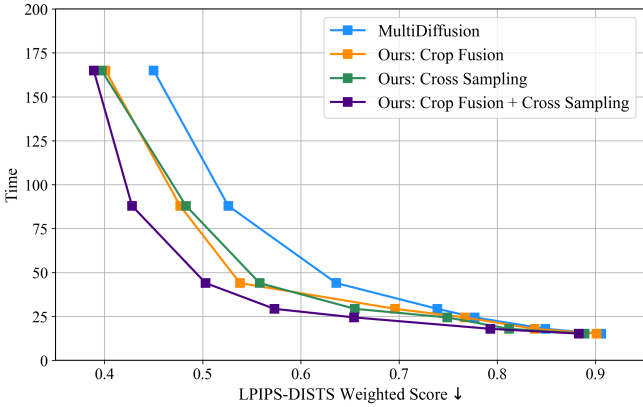
is, the Crop Fusion is carried out from  $t: T \rightarrow \tau$  and then stopped during  $t: \tau \rightarrow 0$ , where  $T$  represents the total timestep of the corresponding diffusion scheduler.

The relationship between  $\tau$  and the stitching effect of twin images is depicted in Fig. 5. We can see a progressive effect on  $I_2^*$  with the decrease of optimization timestep  $\tau$ . (i) When  $\tau > T/2$ , it is under-optimized, as  $[I_2^*]_l$  differs significantly from  $[I_1]_r$ . (ii) When  $\tau = T/2$ , it is well-optimized, generating the most natural and seamless  $I_2^*$ . (iii) When  $\tau < T/2$ , it is over-optimized,  $I_2^*$  stays too close with  $[I_1]_r$  while failing to fuse with  $[I_2^*]_r$ .



**Figure 6.** Further explorations about adjacent control factor  $\lambda$ . It presents a comparative analysis of different  $\lambda$  values to study their impact on the alignment behavior and visual coherence of panoramas. The results demonstrate that our method achieves the best balance around  $\lambda = 1$  and is not sensitive to changes of  $\lambda$  values.

**Adjacent Control** The control factor  $\lambda$  in Eq. 6 plays a crucial role in determining the alignment behavior of image crops. It allows us to adjust the balance between aligning closely with adjacent blocks and maintaining self-alignment, thus significantly influencing the overall quality of panoramic images. Since coherence is an essential attribute in panoramas, we streamline our assessment to focus on LPIPS and DISTS metrics to represent image quality.

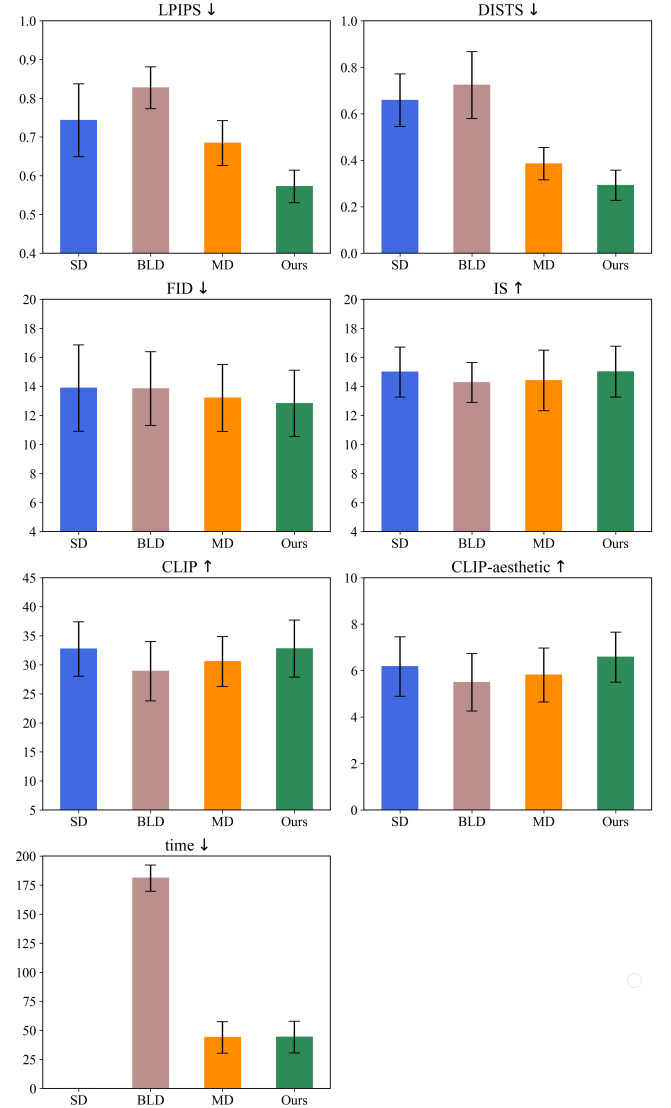


**Figure 7.** Ablation study about our method in improving the quality-speed trade-off. As seen, the effect of the single Crop Fusion and Cross Sampling method is comparable, and they both outperform the MultiDiffusion baseline. With their additive effects, TwinDiffusion exhibits a more robust rate of curve, demonstrating a better speed and quality balance.

In Fig. 6, we compare the results across a wide range of  $\lambda$  values from 0.1 to 100. When  $\lambda$  is between 0.1 and 80, the LPIPS-DISTS weighted score shows a very slow growth trend. It only starts to increase greatly when  $\lambda$  exceeds 80. The lowest point is concentrated

around 1, where our method achieves the optimal balance between the two competing terms mentioned above, leading to the desired visual consistency in panoramic images. This is reasonable and in keeping well with our objective.

**View Stride** As discussed in Sec. 1 and Sec. 3.3, the stride of neighboring views controls the trade-off between quality and efficiency. Smaller  $s_v$  results in better image quality but also takes a lot of time; larger  $s_v$  can accelerate the generation speed of images but the overall quality falls. To test the effectiveness of our approach in improving this problem, we measure the generation time and the aforementioned LPIPS-DISTS weighted score of generated images under  $s_v = 4, 8, 16, 24, 32, 40, 48$ , making a comparison between MultiDiffusion and our TwinDiffusion with and without the two optimization method.



**Figure 8.** Quantitative results comparing our approach with the baselines. Our method demonstrates its superiority in all aspects, particularly in coherence, the best visual representation of panoramas.

As seen in Fig. 7, our method surpasses the baseline in a better quality-efficiency balance. (i) For quality: under the same view stride, both the Crop Fusion and Cross Sampling methods reach a lower LPIPS-DISTS weighted score than the MultiDiffusion, and full



**Table 1.** Here are the findings from our further exploration of the Cross Sampling strategy. Different  $s_r$  values were tested across all metrics. The better effects are more likely to be achieved when  $s_r = s_v/2$ ,  $s_v/3$ , indicating that excessively fine cross strides are unnecessary.

	Coherence		Fidelity & Diversity		Compatibility		Efficiency
	LPIPS↓	DISTS↓	FID↓	IS↑	CLIP↑	CLIP-aesthetic↑	time↓
w/o Cross Sampling	0.69	0.49	<u>14.03</u>	12.31	30.91	6.43	<b>43.98</b>
$s_r = s_v/2$	<b>0.60</b>	<u>0.43</u>	<b>13.16</b>	<b>14.49</b>	<b>32.06</b>	<b>6.76</b>	<u>43.99</u>
$s_r = s_v/3$	<u>0.63</u>	<b>0.42</b>	14.29	<u>14.15</u>	31.38	6.46	45.01
$s_r = s_v/5$	0.71	0.59	15.11	13.98	30.93	6.42	47.52
$s_r = s_v/7$	0.76	0.62	16.00	13.98	30.25	6.48	53.60
$s_r = s_v/s_v$	0.84	0.78	16.21	12.58	30.26	6.05	59.67

TwinDiffusion reaches even lower scores due to the combined effects of the two. (ii) For efficiency: analyzing the blue and green lines, we can observe that our Cross Sampling approach successfully enables generating comparable results within a fraction of the time required by the original MultiDiffusion, specifically achieving a reduction of N-fold when using a  $N \times s_v$  value. (iii) For quality-efficiency trade-off: the slopes indicate that our method takes greater advantage of the efficiency gained from larger  $s_v$  while maintaining high image quality, striking an optimal balance between quality and speed for panoramic image generation.

**Cross Stride** We further investigate the influence of cross stride to ensure an appropriate interleaving frequency in the Cross Sampling method. In particular,  $s_r = 1$  means setting a different sampling mode per pixel, and  $s_r = s_v$  is equivalent to not using our Cross Sampling method. All the results are obtained under the same condition described in Sec. 4.2.

Tab. 1 provides a series of comparisons on all sides, with the best and the second-best results marked in bold and underlined respectively. The scores show a stable and consistent pattern, where the most desirable outcomes are generally achieved when  $s_r$  is set to  $s_v/2$  or  $s_v/3$ . In the extreme case of  $s_r = s_v/s_v = 1$ , the effect regresses to no Cross Sampling or even worse. This observation suggests that a finer interleaving level of sampling does not contribute to the quality scores but adds unnecessary running time.

## 5 Conclusion

In this paper, we have presented TwinDiffusion, an optimized framework for panoramic image generation using state-of-the-art diffusion models. Our work breaks through the existing limitations in quality and efficiency by introducing two key innovations: (i) a lightweight fusion stage to enhance coherence, and (ii) an interleaved sampling method to improve generating speed. By extending this promising framework to wider domains, especially virtual reality and graphic design, we can unlock new possibilities for creating dynamic and immersive visual content.

**Limitations and Social Impact** Although TwinDiffusion works well in most cases, it still faces some limitations. Our approach primarily focuses on optimizing the local similarity of the image areas. However, it could not ensure stability in perceiving the overall layout of the images, which may lead to the generation of visually coherent but spatially illogical panoramas. As for potential negative impact, image generation models may involve personal copyright or generate fake, offensive, discriminatory results. Further research should prioritize the responsible use of the relevant technology to avoid generating content in any harmful way.

## Acknowledgements

We would like to express sincere gratitude to Xiaoyu Zhang, Mingyue Hu, and the entire research group for their inspiration and support. Furthermore, we thank Yunhao Chen for assisting in diagnosing issues and providing crucial feedback and suggestions.

## References

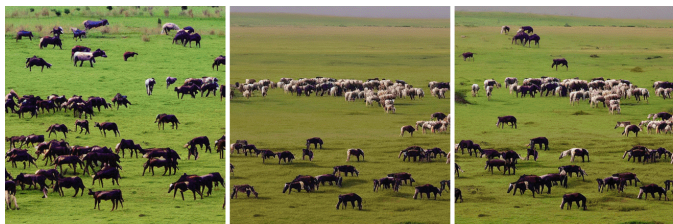
- [1] O. Avrahami, O. Fried, and D. Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023.
- [2] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.
- [3] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.
- [4] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [5] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>.
- [6] D. Chira, I. Haralampiev, O. Winther, A. Dittadi, and V. Liévin. Image super-resolution with deep variational autoencoders. In *European Conference on Computer Vision*, pages 395–411. Springer, 2022.
- [7] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [8] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.
- [9] Z. Duan, C. Wang, C. Chen, J. Huang, and W. Qian. Optimal linear subspace search: Learning to construct fast and high-quality schedulers for diffusion models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 463–472, 2023.
- [10] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.
- [11] M. Feng, J. Liu, M. Cui, and X. Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*, 2023.
- [12] G. Franzese, S. Rossi, L. Yang, A. Finamore, D. Rossi, M. Filippone, and P. Michiardi. How much is enough? a study on diffusion times in score-based generative models. *Entropy*, 25(4):633, 2023.
- [13] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023.
- [14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [15] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.



- [16] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.
- [17] Á. B. Jiménez. Mixture of diffusers for scene composition and high resolution image generation. *arXiv preprint arXiv:2302.02412*, 2023.
- [18] I. Kobyzev, S. J. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.
- [19] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.
- [20] Z. Pan, B. Zhuang, D.-A. Huang, W. Nie, Z. Yu, C. Xiao, J. Cai, and A. Anandkumar. T-stitch: Accelerating sampling in pre-trained diffusion models with trajectory stitching. *arXiv*, 2024.
- [21] K. Pandey, A. Mukherjee, P. Rai, and A. Kumar. Vae meet diffusion models: Efficient and high-fidelity generation. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
- [22] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [24] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- [25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [26] T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.
- [27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [28] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.
- [29] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [30] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [31] J. Tang, T. Wang, B. Zhang, T. Zhang, R. Yi, L. Ma, and D. Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22819–22829, 2023.
- [32] A. Ulhaq, N. Akhtar, and G. Pogrebna. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*, 2022.
- [33] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen, and B. Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4563–4573, June 2023.
- [34] H. Xiao, X. Wang, J. Wang, J.-Y. Cai, J.-H. Deng, J.-K. Yan, and Y.-D. Tang. Single image super-resolution with denoising diffusion gans. *Scientific Reports*, 14(1):4272, 2024.
- [35] J. Xiao, T. Liu, Y. Zhang, B. Zou, J. Lei, and Q. Li. Multi-focus image fusion based on depth extraction with inhomogeneous diffusion equation. *Signal Processing*, 125:171–186, 2016.
- [36] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.
- [37] S. Yao, M. Sun, B. Li, F. Yang, J. Wang, and R. Zhang. Dance with you: The diversity controllable dancer generation via diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8504–8514, 2023.
- [38] C. Zhang, Q. Wu, C. C. Gambardella, X. Huang, D. Phung, W. Ouyang, and J. Cai. Taming stable diffusion for text to 360 {°} panorama image generation. *arXiv preprint arXiv:2404.07949*, 2024.
- [39] Q. Zhang, J. Song, X. Huang, Y. Chen, and M.-Y. Liu. Diffcollage: Parallel generation of large content with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10188–10198. IEEE, 2023.
- [40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

## A More Implementation Examples on Twin-Image Generation

Our experimental results of twin images are provided in Fig. A1, meaning the high stability of our method.



"A scene of animal herds across the grassland"



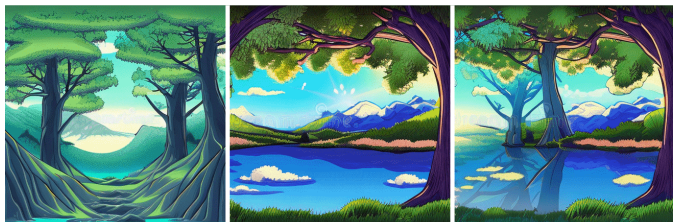
"A mosaic artwork of a peaceful countryside"



"A beach with seagulls, people and palm trees"



"A photo of the mountain, lake, people and boats"



"Natural landscape in anime style illustration"



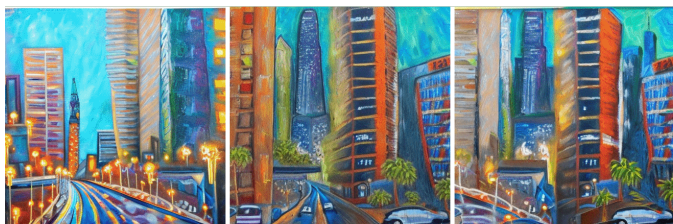
"A surrealistic artwork of urban park at dawn"



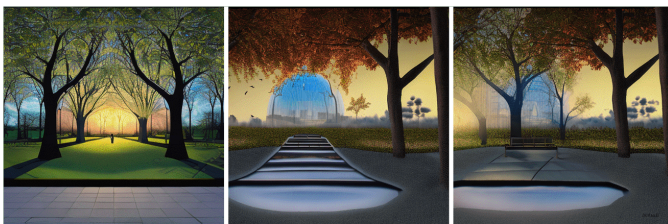
"An oil painting of blooming flowers"



"A picture of birds landing on a telephone pole under blue sky"



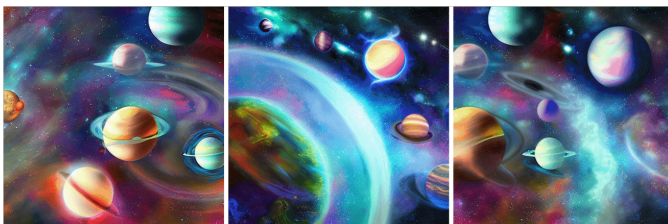
"An acrylic painting of the cityscape"



"A surrealistic artwork of urban park at dawn"



"A landscape ink painting"



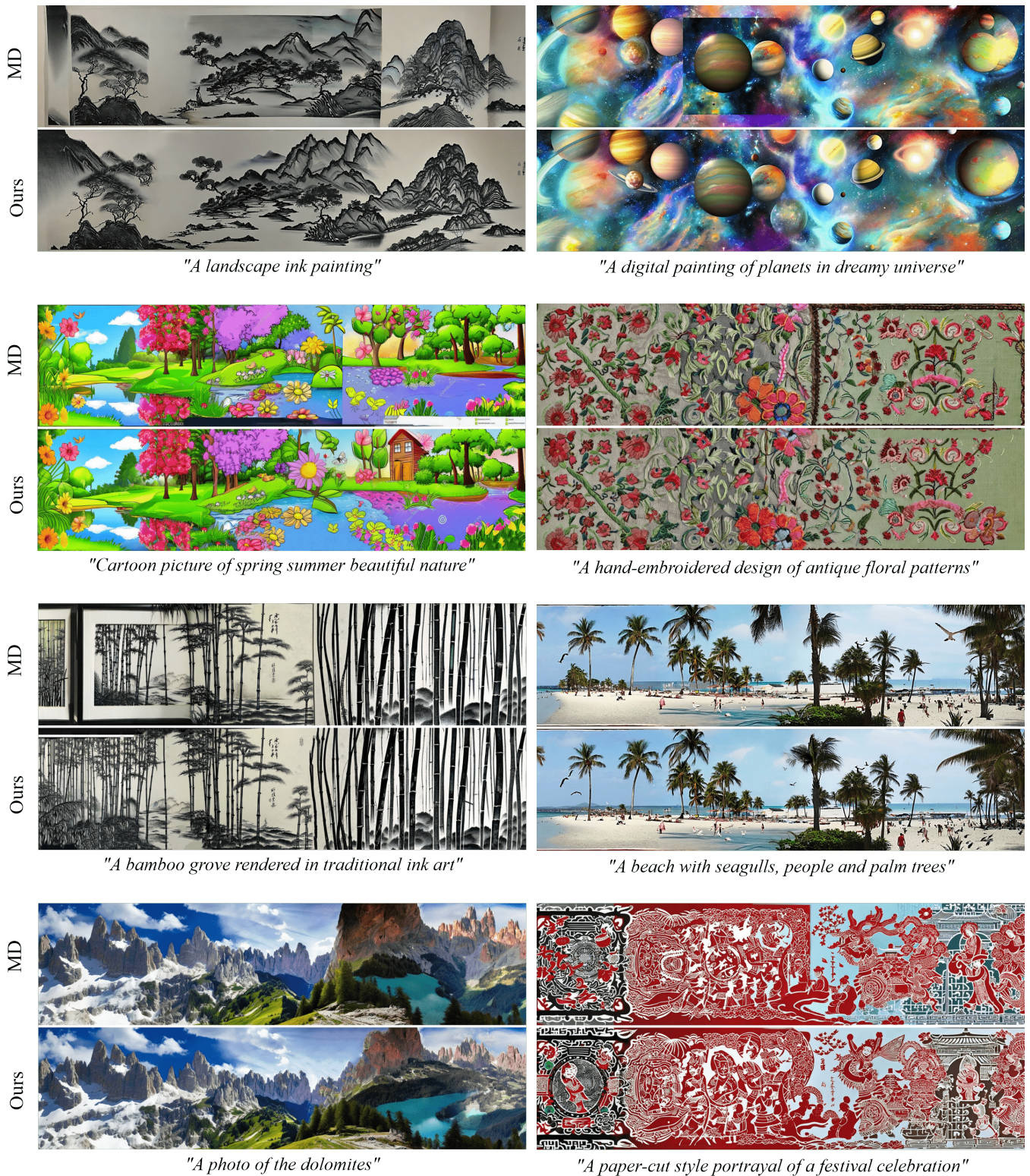
"A digital painting of planets in dreamy universe"

**Figure A1.** More examples of twin images with our TwinDiffusion.



## B More Results of Qualitative Comparison on Panorama Generation

Fig. A2 gives more qualitative comparisons with MultiDiffusion on panoramic image generation, highlighting the areas where improvements have been made.



**Figure A2.** Additional qualitative comparison results on panorama generation with various text prompts.



## C Analysis on Different Reference Models

In addition to Stable Diffusion v2.0, we also conduct a series of experiments on Stable Diffusion XL v1.0 as our reference model  $\Phi$ , aiming to further explore the generalization capability of our approach. Considering that images generated by SDXL have larger dimensions of  $1024^2$ , we also crop them into  $512^2$ , which aligns with the other image sets.

Tab. A1 presents a comprehensive comparison among the methods using different reference models. As observed, even in the context of SDXL, our approach still contributes to refining the visual and semantic coherency of generated panoramas, albeit to a lesser extent than observed with the SD-based reference model. This observation is understandable, considering the inherent attributes of SDXL as an advanced diffusion model that doubles the default resolution in both height and width. Despite the limitations imposed by the reference model, our approach still enhances several metrics in panoramic image generation tasks, demonstrating its ability to improve panorama quality.

**Table A1.** Comparisons among different methods using various reference models. Despite the advanced performance stored in SDXL itself, our approach still demonstrates its ability to improve the quality and coherence of the generated panoramas.

	Coherence		Fidelity & Diversity		Compatibility		Efficiency
	LPIPS↓	DISTS↓	FID↓	IS↑	CLIP↑	CLIP-aesthetic↑	time↓
SD <sub>2.0</sub>	0.74	0.66	13.88	14.98	32.43	6.18	/
MD <sub><math>\Phi</math>=SD<sub>2.0</sub></sub>	0.68	0.39	13.46	13.20	31.55	5.81	43.92
Ours <sub><math>\Phi</math>=SD<sub>2.0</sub></sub>	<b>0.60</b>	<u>0.29</u>	<u>12.83</u>	15.01	32.72	<u>6.56</u>	44.17
SDXL <sub>1.0</sub>	<u>0.61</u>	<u>0.29</u>	<b>11.85</b>	<b>16.01</b>	<b>33.54</b>	<u>6.50</u>	/
MD <sub><math>\Phi</math>=SDXL<sub>1.0</sub></sub>	0.67	<b>0.28</b>	<b>11.85</b>	14.74	32.40	6.37	172.62
Ours <sub><math>\Phi</math>=SDXL<sub>1.0</sub></sub>	<b>0.60</b>	<b>0.28</b>	<b>11.85</b>	<u>15.63</u>	<u>32.78</u>	<b>6.58</b>	174.46