Guided and Variance-Corrected Fusion with One-shot Style Alignment for Large-Content Image Generation

Shoukun Sun[1], **Min Xian**[1], **Tiankai Yao**[2], **Fei Xu**[2], **Luca Capriotti**[2]
[1]Department of Computer Science, University of Idaho
[2]Idaho National Laboratory
sun5322@vandals.uidaho.edu, mxian@uidaho.edu, {tiankai.yao, fei.xu, luca.capriotti}@inl.gov

Abstract

Producing large images using small diffusion models is gaining increasing popularity, as the cost of training large models could be prohibitive. A common approach involves jointly generating a series of overlapped image patches and obtaining large images by merging adjacent patches. However, results from existing methods often exhibit noticeable artifacts, e.g., seams and inconsistent objects and styles. To address the issues, we proposed Guided Fusion (GF), which mitigates the negative impact from distant image regions by applying a weighted average to the overlapping regions. Moreover, we proposed Variance-Corrected Fusion (VCF), which corrects data variance at post-averaging, generating more accurate fusion for the Denoising Diffusion Probabilistic Model. Furthermore, we proposed a one-shot Style Alignment (SA), which generates a coherent style for large images by adjusting the initial input noise without adding extra computational burden. Extensive experiments demonstrated that the proposed fusion methods improved the quality of the generated image significantly. The proposed method can be widely applied as a plug-and-play module to enhance other fusion-based methods for large image generation.

Code — https://github.com/TitorX/GVCFDiffusion

Introduction

Recent years have witnessed remarkable advancements in text-to-image generation models, which can produce realistic and diverse images based on textual prompts. Among them, the Diffusion models, specifically the Stable Diffusion (SD) (Rombach et al. 2022), have emerged as one of the mainstream methods for image generation.

There is a significant demand for producing large images. The pursuit of generating larger images involves 1) producing images with higher resolution that exhibit ultra-fine details and 2) creating images that encompass more content, such as panorama images. To differentiate between these aspects, we refer to them as High-Resolution image generation and Large-Content image generation, respectively. However, training models capable of generating large images requires a substantial investment in hardware and data. For instance, training the SD v2 model to generate 512^2 images took over a month on 256 A100 GPUs. The core U-Net model of it comprises 865 million parameters. The larger SDXL (Podell et al. 2023) model, which can generate 1024^2 images and contains 2.6 billion parameters, demands a longer training period.

Recent progress has been made by using pre-trained smaller models to jointly generate a series of overlapped small patches, which are then combined to form images of arbitrary sizes. A notable work is MultiDiffusion (Bar-Tal et al. 2023), which generates large images by averaging overlapped areas of patches at each denoising step. SyncDiffusion (Lee et al. 2023) achieves more coherent large-content images by ensuring consistent styles across each small patch during the joint denoising process. However, existing methods exhibit three major drawbacks: 1) noticeable seams at overlapped areas, 2) generation of discontinuous objects, and 3) low-quality content.

Each patch derives different values in the overlapped regions at each denoising step. Resolving discrepancies by averaging to achieve uniformity values can interfere with the denoising of individual patches. This interference occurs because diffusion models, during training, assume that the whole denoising process is completed with all intermediate results undisturbed. Persistent changes to the values in certain regions can have unknown impacts on the denoising process, typically resulting in adverse effects.

We propose a method termed Guided Fusion (GF), which assigns a guidance map to each small patch to perform weighted averaging in the overlapped regions. This allows the denoising process to be dominated by patches close to the center (i.e., higher weights). Additionally, we discovered that averaging the overlapped regions while using Stochastic Differential Equation (SDE) samplers, such as Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020), produces highly blurred results. This occurs because the SDE samplers usually introduce a Gaussian-distributed random term during the denoising process, and averaging multiple variables sampled from Gaussian distributions results in a variance lower than expected, leading to blurred images that lack details. To address this, we introduce Variance-Corrected Fusion (VCF) to adjust the variance and generate higher-quality images. Furthermore, we observed that significant differences in the initial noise each patch uses make it more challenging to produce coherent

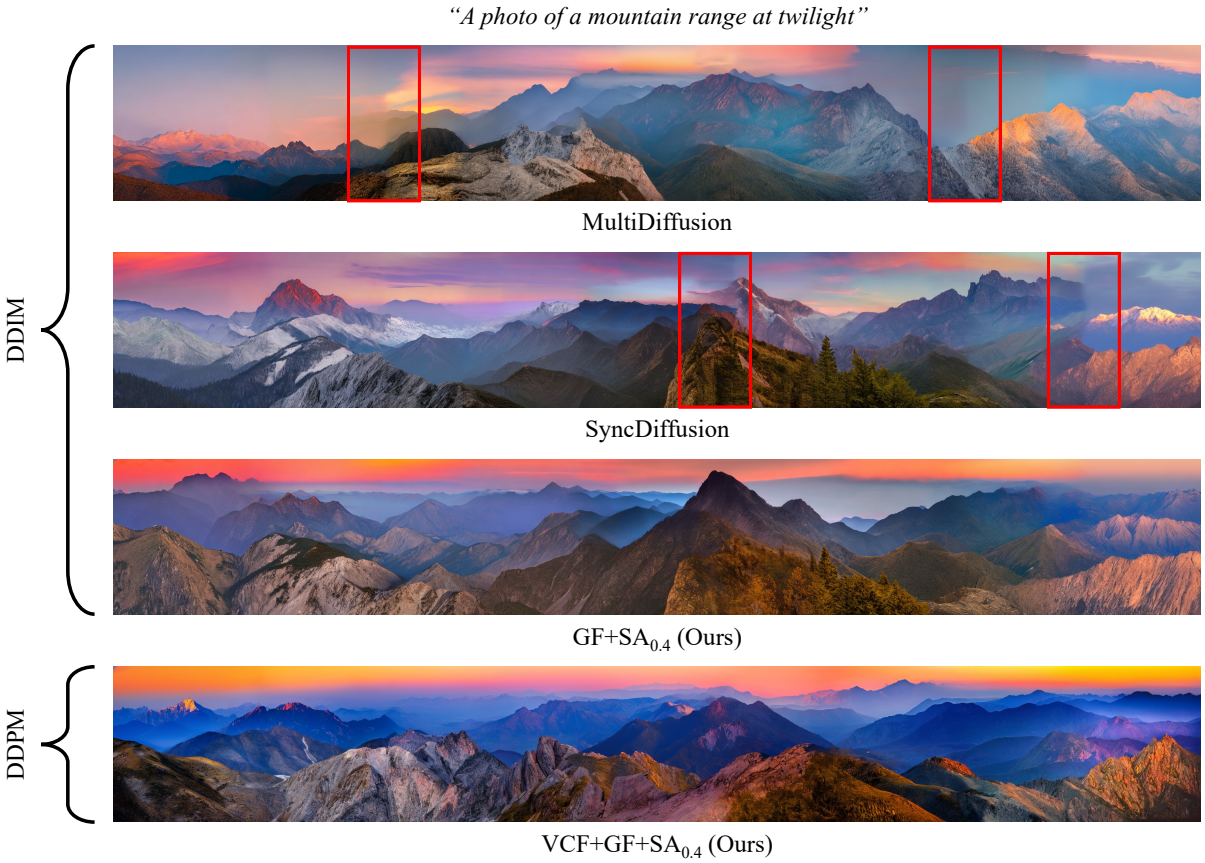

Figure 1: Comparisons of panorama images generated by MultiDiffusion (Bar-Tal et al. 2023), SyncDiffusion (Lee et al. 2023) and our methods: Guided Fusion (GF), Variance-Corrected Fusion (VCF) and Style Alignment (SA). All images are generated with the same initial noise. The red boxes highlight the discontinuous and defective areas on the generated image.

images. Therefore, we propose a one-shot Style Alignment (SA), which aligns the initial noise with semantic interpolation to produce more style-consistent results.

The main contributions of this paper are as follows:

- Guided Fusion was proposed to utilize a guidance map for weighted averaging on overlapped areas, leading to better quality and seamless image generation.
- We proposed the Variance-Corrected Fusion to fix the small variance issue while averaging overlapped regions with SDE samplers. The proposed method prevents generating blurred results with SDE samplers, leading to higher-quality image generation.
- We proposed the one-shot Style Alignment approach that aligns the style of the initial noise only once to generate more coherent content without increasing the computational burden.

Preliminaries

The core of diffusion models (DMs) lies in the concept of a Markov process, specifically, a type of Markov chain where each step adds a controlled amount of Gaussian noise to the data. The forward diffusion process is defined as a sequence of latent variables $\{\mathbf{x}_t\}$ indexed by discrete time steps $t = 0, 1, \ldots, T$, where $\mathbf{x}_0$ represents the original data and $\mathbf{x}_T$ approximates a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The transition from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$ is modeled by a Gaussian distribution, typically formulated as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \quad (1)$$

Here, the schedule of variances β_t is designed to gradually add noise to $\mathbf{x}_t$, which can be learned by reparameterization (Kingma and Welling 2013) or held a sequence of constants as hyperparameters (Rombach et al. 2022; Song, Meng, and Ermon 2020). The choice of the $\{\beta_t\}$ is critical as it controls the rate at which the data is diffused into noise over time.

The reverse diffusion process, or called the denoising process, involves learning a model $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ that approximates the reverse of the forward process. This is done by parameterizing the Gaussian distribution with learnable parameters θ, usually expressed as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)), \quad (2)$$

where $\mu_\theta(\mathbf{x}_t, t)$ and $\mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$ are learned through optimization. The objective is to minimize the difference between the true reverse distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ and the modeled distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$.

A common practice sets the schedule of β_t as an increasing sequence of constants at the forward process. The reverse process sets $\mathbf{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2\mathbf{I}$ and let $\sigma_t^2 = \beta_t$ or $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ (Ho, Jain, and Abbeel 2020), where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\alpha_t := 1 - \beta_t$. Hence, we can formulate:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}). \tag{3}$$

Latent Diffusion Model (LDM) (Rombach et al. 2022) extends diffusion models by operating in a low-dimensional latent space instead of the high-dimensional pixel space. This is achieved by first encoding the data into a latent representation using a suitable encoder and then applying the diffusion process within this more compact space. This reduction in dimensionality leads to more efficient modeling and sampling as the model needs to learn and operate over fewer parameters. The Variational Autoencoders (VAEs) (Kingma and Welling 2013) are often chosen for encoding images to latent space and decoding to pixel space.

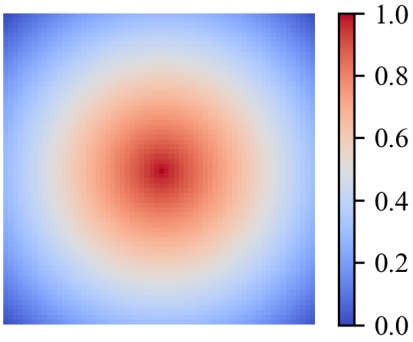

Figure 2: Guided Fusion Map.

Figure 3: Images produced by direct averaging overlapped areas with DDIM and DDPM sampler, and a result from DDPM with Variance-Corrected Fusion (VCF).

Method

The nature of the joint denoising process. We denote a small pretrained diffusion model as a parametric model that has been optimized for a series of Markov chained Gaussian transitions $p_\theta(\mathbf{x}_0) := p_\theta(\mathbf{x}_{0:T})d\mathbf{x}_{1:T}$ at a low-dimensional space $\mathbf{x}_0 \in \mathbb{R}^n$. As the small diffusion model has never been optimized with the high-dimensional dataset, it cannot be directly used to sample larger images. The joint denoising process uses the small model to obtain large images $\mathbf{X}_0 \in \mathbb{R}^m$, where $m > n$, by fusing a series of overlapped patches after each denoising step. Since the distribution in high-dimensional space is unknown, we can only aim to sample a $\mathbf{X}_0$ for which each subview: 1) conforms to a learned distribution in the low-dimensional space so that each generated patch is realistic; 2) shares identical values in the overlapping dimensions so that can be merged to form a large sample.

The drawbacks of averaging latent variables. Use a simple case as an illustration, we denote a large sample with three dimensions as $\mathbf{X} = [x^{(1)}, x^{(2)}, x^{(3)}]$ and use a two-dimensional model to jointly produce overlapped patches $\mathbf{x}^{(1)} = [x^{(1)}, x^{(2)}]$ and $\mathbf{x}^{(2)} = [x^{(2)}, x^{(3)}]$. The MultiDiffusion (Bar-Tal et al. 2023) introduced a joint denoising process that averages values on overlapped dimensions after each denoising step, which can be described as:

$$\begin{aligned}[x_{t-1}^{(1)}, x_{t-1}^{(21)}] &\sim p_\theta(\mathbf{x}_{t-1}^{(1)}|\mathbf{x}_t^{(1)}) \\ [x_{t-1}^{(22)}, x_{t-1}^{(3)}] &\sim p_\theta(\mathbf{x}_{t-1}^{(2)}|\mathbf{x}_t^{(2)})\end{aligned} \tag{4}$$

$$x_{t-1}^{(2)} = \frac{x_{t-1}^{(21)} + x_{t-1}^{(22)}}{2} \tag{5}$$

$$\begin{aligned}\mathbf{x}_{t-1}^{(1)} &:= \widetilde{\mathbf{x}}_{t-1}^{(1)} = [x_{t-1}^{(1)}, x_{t-1}^{(2)}] \\ \mathbf{x}_{t-1}^{(2)} &:= \widetilde{\mathbf{x}}_{t-1}^{(2)} = [x_{t-1}^{(2)}, x_{t-1}^{(3)}].\end{aligned} \tag{6}$$

As shown in Eq. 4, the denoising steps for $\mathbf{x}_t^{(1)}$ and $\mathbf{x}_t^{(2)}$ produce diverged values $x_{t-1}^{(21)}$ and $x_{t-1}^{(22)}$ over the same dimension $x^{(2)}$. Averaging by Eq. 5 solves the divergence to ensure the overlapped dimension shares the same value after each step.

As described in Eq. 3, throughout the denoising process, for $1 < t < T$, $\mathbf{x}_{t-1}$ should be estimated by the conditional probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. However, during the patch averaging, the values of overlapped dimensions have been constantly modified, leading to the next $\mathbf{x}_{t-1}$ being estimated conditioned at an altered $\widetilde{\mathbf{x}}_t$. Such value altering perturbs the denoising transitions, leading to obvious seams and reduced quality.

Mitigate Divergence among Patches with Guided Fusion

The patch fusion using averaging could significantly alter the patch values and shift the input distribution. The diffusion model has not been fine-tuned in a training-free setting for such modified input. Hence, the model may generate degenerated results. To mitigate the input distribution shifting issue, we propose a guidance map as shown in Figure 2,

Stride	Fusion	FID↓	KID↓($\times 10^{-3}$)	GIQA-QS↑	GIQA-DS↑	CLIP↑
128	MD	20.60	9.14	9.311	9.203	**31.65**
	GF	**17.64**	**7.72**	**9.324**	**9.218**	31.59
256	MD	17.32	7.21	9.183	9.117	**31.58**
	GF	**15.99**	**6.68**	**9.280**	**9.188**	31.52
384	MD	15.55	6.49	9.208	9.136	31.50
	GF	**14.88**	**6.28**	**9.236**	**9.159**	**31.51**

Table 1: Quantitative comparisons between MultiDiffusion (MD) (Bar-Tal et al. 2023) and Guided Fusion (GF) with DDIM sampler using various strides. The best results within each stride group are marked in bold.

which linearly decreases its weight from 1 at the center to 0 at the corners to guide the weighted averaging of the overlapping regions. The weighted averaging allows the altered input to attach more to the closer patches, reducing the significant value changes.

Disrupting the denoising process of a patch in different regions may lead to varying degrees of model performance degradation. Intuitively, we consider that the closer the disturbed area is to the center, the greater the impact on the quality of the generated image. Therefore, we propose a guidance map as shown in Figure 2, which linearly decreases its weight from 1 at the center to 0 at the corners to guide the weighted averaging of the overlapping regions. Following the example described by Eq. (5), the weighted average at overlapped dimension can be formulated as:

$$x_{t-1}^{(2)} = \frac{w_1 x_{t-1}^{(21)} + w_2 x_{t-1}^{(22)}}{w_1 + w_2} \tag{7}$$

where the corresponding locations on the guidance map determine the weights w_1 and w_2. To generalize the simple case to *N* overlapped patches, we formulate the weighted average for each dimension from overlapped areas as:

$$x_{t-1} = \frac{\sum_i^N w_i x_{t-1}^{(i)}}{\sum_i^N w_i}. \tag{8}$$

This method is called Guided Fusion (GF). During the joint denoising process, the value of each dimension in the overlapped area is predominantly determined by the geometrically closer patch, thereby reducing the perturbation in the denoising process for that dimension.

Correcting Variance of Fused Patches with SDE Samplers

For Ordinary Differential Equation (ODE) samplers, such as the Denoising Diffusion Implicit Model (DDIM) (Song, Meng, and Ermon 2020), the experimental results demonstrate that although averaging fusion interferes with the denoising process, it can still produce compelling images, as shown in the first row of Figure 3. However, for scenarios requiring the use of Stochastic Differential Equation (SDE) samplers, such as DDPM (Ho, Jain, and Abbeel 2020), averaging can lead to faulty blurred results, as displayed in the second row of Figure 3. We use DDPM as an example to illustrate the reason.

For a single image patch generation using DDPM, the $t-1$ denoised image is computed by:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta \left(\mathbf{x}_t, t \right) \right) + \sigma_t \mathbf{z} \tag{9}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We can consider $\mathbf{x}_t$ as a known variable because the previous step has determined it, hence the:

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_t, \sigma_t^2) \tag{10}$$

where $\mu_t = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta \left(\mathbf{x}_t, t \right) \right)$.

Continuing the example from the Eq. (5) using DDPM sampler, the fused denoised dimension $x_{t-1}^{(2)} = \frac{x_{t-1}^{(21)} + x_{t-1}^{(22)}}{2}$ has:

$$x_{t-1}^{(2)} \sim \mathcal{N}(\frac{\mu_t^{(21)} + \mu_t^{(22)}}{2}, \frac{\sigma_t^2}{2}). \tag{11}$$

We notice that the variance becomes $\sigma_t^2/2$, smaller than the expected σ_t^2 as in Eq. (10). This causes blurred results while applying averaging with DDPM, e.g., the second row of Figure 3. The reduced variance leads to over-homogeneous image content.

We propose the Variance-Corrected Fusion (VCF) by redefining $x_{t-1}^{(2)}$ to correct the variance:

$$\begin{aligned} x_{t-1}^{(2)} =& \sqrt{2} \frac{x_{t-1}^{(21)} + x_{t-1}^{(22)}}{2} + (1-\sqrt{2}) \frac{\mu_t^{(21)} + \mu_t^{(22)}}{2} \\ =& \frac{x_{t-1}^{(21)} + x_{t-1}^{(22)}}{\sqrt{2}} + (1-\sqrt{2}) \frac{\mu_t^{(21)} + \mu_t^{(22)}}{2}, \end{aligned} \tag{12}$$

so that have $x_{t-1}^{(2)} \sim N((\mu_t^{(21)} + \mu_t^{(22)})/2, \sigma_t^2)$.

We generalize the Eq. (12) to averaging *N* overlaps:

$$x_{t-1} = \frac{\sum_i^N x_{t-1}^{(i)}}{\sqrt{N}} + (1-\sqrt{N}) \frac{\sum_i^N \mu_t^{(i)}}{N}, \tag{13}$$

and generalize to Guided Fusion weighted average:

$$\begin{aligned} x_{t-1} =& \frac{\sum_i^N w_i x_{t-1}^{(i)}}{\sqrt{\sum_i^N w_i^2}} \\ &+ (1 - \frac{W}{\sqrt{\sum_i^N w_i^2}}) \frac{\sum_i^N w_i \mu_t^{(i)}}{W}, \end{aligned} \tag{14}$$

Samplers	Methods	FID↓	KID↓ ($\times 10^{-3}$)	GIQA-QS↑	GIQA-DS↑	CLIP↑
DDIM	MD	11.22	3.47	8.952	8.870	31.69
	Sync	11.19	3.34	8.965	<u>8.882</u>	<u>31.70</u>
	Elastic	122.52	53.88	6.562	6.711	22.46
	GF (Ours)	<u>10.71</u>	**3.10**	**8.976**	**8.887**	31.67
	MD + $SA_{0.4}$ (Ours)	10.82	3.27	8.948	8.866	**31.71**
	GF + $SA_{0.4}$ (Ours)	**10.40**	<u>3.12</u>	<u>8.970</u>	8.877	31.68
DDPM (Ours)	VCF	4.78	1.29	8.987	8.970	<u>31.87</u>
	VCF + GF	<u>4.35</u>	<u>1.09</u>	**9.001**	**8.981**	31.85
	VCF + $SA_{0.4}$	4.43	1.19	8.977	8.962	**31.90**
	VCF + GF + $SA_{0.4}$	**4.02**	**1.02**	<u>8.998</u>	<u>8.976</u>	31.86

Table 2: Overall performance. The subscript of SA indicates the value of α. Each sampler group's best and second results are marked in bold and underlined, respectively.

where $W = \sum_i^N w_i$.

The corrected formula can be applied to other SDE samplers that employ Gaussian noise, such as the EDM stochastic sampler (Karras et al. 2022).

One-shot Style Alignment (SA) for Coherent Montages

SyncDiffusion (Lee et al. 2023) inspires us that aligning the style of each small patch reduces the difficulty of generating more coherent content. However, SyncDiffusion requires constantly modifying the intermediate denoised patches to align their style, further disrupting the denoising process.

We noticed that the diffusion model exhibits the semantic interpolation effect (Song, Meng, and Ermon 2020), in which the interpolations between two initial noises can lead to semantically meaningful results.

We propose a one-shot style-control method, Style Alignment (SA), performing interpolation on each non-overlapped patch cropped from the whole initial noise to a reference noise. The SA can be formulated as:

$$\mathbf{x}_T^{(i)} := \text{slerp}(\mathbf{x}_T^{(i)}, \mathbf{z}^{\text{ref}}, \alpha) \tag{15}$$

where the $\text{slerp}(\cdot)$ is the spherical linear interpolation (Shoemake 1985) function; $\mathbf{x}_T^{(i)}$ is the i^{th} non-overlapped crop from the initial noise $\mathbf{X}_T$; $\mathbf{z}^{\text{ref}}$ is a reference noise to be aligned with; $\alpha \in [0, 1]$ is the interpolation ratio where 0 returns the original $\mathbf{x}_T^{(i)}$ and 1 returns $\mathbf{z}^{\text{ref}}$. The reference noise $\mathbf{z}^{\text{ref}}$ can be any standard Gaussian noise. It may originate from a patch of the initial noise $\mathbf{X}_T$ or be obtained through diffusing a specific image.

After SA alignment, all non-overlapped patches rotate towards the reference noise, making them more clustered. Consequently, the distances between them are reduced, and their similarity increases.

Results

Generated Datasets. The text-to-panorama generation task was chosen to assess each method's performance on large-content image generation. For each approach, we sampled a set of 512×3584 sized images, $\times 7$ wider than the original model resolution, with ten prompts and 500 panorama images for each prompt. In total, 5,000 panorama images were generated for each approach. The panorama images were further divided into 7 patches matching the original model size, ultimately producing 35,000 images. The ten used prompts are:

1) A photo of a city skyline at night; 2) A photo of a mountain range at twilight; 3) A photo of a snowy mountain peak with skiers; 4) Cartoon panorama of spring summer beautiful nature; 5) Natural landscape in anime style illustration; 6) A photo of lush forest with a babbling brook; 7) An illustration of a beach in La La Land style; 8) Silhouette wallpaper of a dreamy scene with shooting stars; 9) A beach with palm trees; and 10) A film photo of a beachside street under the sunset.

We conducted both qualitative and quantitative comparative experiments with the results obtained from MultiDiffusion (MD) (Bar-Tal et al. 2023), SyncDiffusion (Sync) (Lee et al. 2023) and ElasticDiffusion (Elastic) (Haji-Ali, Balakrishnan, and Ordonez 2024).

Reference Dataset. Based on the prior works, the ODE samplers, such as DDIM, tend to lead to worse output quality (Song, Meng, and Ermon 2020; Song et al. 2020; Karras et al. 2022). We chose the SDE sampler DDPM to generate the reference dataset, which stands for higher quality. We used Stable Diffusion (Rombach et al. 2022) v2.0 to generate reference images for evaluation. A reference dataset that contains 35,000 of 512×512 images was generated with 3500 images per prompt.

Evaluation Metrics. To assess the image quality, we employed FID (Heusel et al. 2017), KID (Bińkowski et al. 2018) (we use the anti-aliasing implementation (Parmar, Zhang, and Zhu 2022)) and GIQA-QS/GIQA-DS (Gu et al. 2020) to evaluate the fidelity and diversity; CLIP score (Hessel et al. 2021) to evaluate the compatibility with the prompt.

The results presented in Table 1 and Figure 5 were evaluated from the first five prompts. Table 2 results were calculated from ten prompts.

The Effectiveness of Guided Fusion

The stride controls the overlap ratio between patches; a smaller stride indicates a larger overlapping ratio. Additionally, a smaller stride indicates that more patches are needed

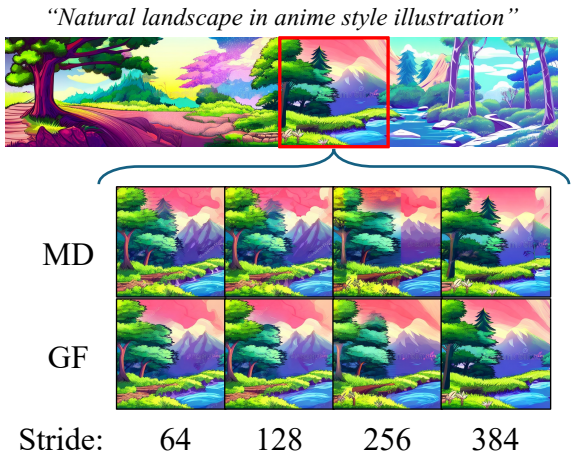

Figure 4: MultiDiffusion (MD) compared with Guided Fusion (GF) with different strides. All images are generated with the same initial noise.

Samplers	Methods	Runtime (s)
DDIM	MD	9.70±0.02
	Sync	102.31±0.52
	Elastic	54.55±0.23
	GF	9.64±0.07
DDPM	VCF	181.16±0.31

Table 3: Average runtime of single image generation. Results were calculated from 500 generations for each method.

in joint denoising to form a large image. Figure 4 shows qualitative results from MultiDiffusion (MD) and Guided Fusion (GF) over 64, 128, 256, and 384 strides with a DDIM sampler. It can be observed that noticeable seams are present in the results of MD with four different strides. Among these, the seams are least apparent with the 64 stride, while they are most pronounced with the 256 stride. After applying GF, the seams are significantly reduced at all strides, resulting in more continuous images.

To thoroughly evaluate the effectiveness of the proposed GF, we compared our method with MD in three stride settings: 128, 256, and 384 with quantitative metrics. As shown in Table 1, the experimental results indicate that GF consistently outperforms MD across different strides. Specifically, GF achieved the best results in several key metrics, including FID, KID, GIQA-QS, and GIQA-DS, while MD demonstrated an advantage in CLIP scores. GF exhibited superior image quality and diversity, highlighting its greater applicability in fusing overlapped patches.

As shown in Table 1, as the stride increases, the FID and KID metrics of the results are gradually improved for both MD and GF. This supports our viewpoint: modifying the values in overlapping regions interferes with the denoising process of each patch and negatively affects the quality of the generated images. Although the seams are less evident with a higher overlap ratio, as the overlap ratio decreases, the FID and KID metrics of MD and GF gradually decrease, indicating that the generated images have better details.

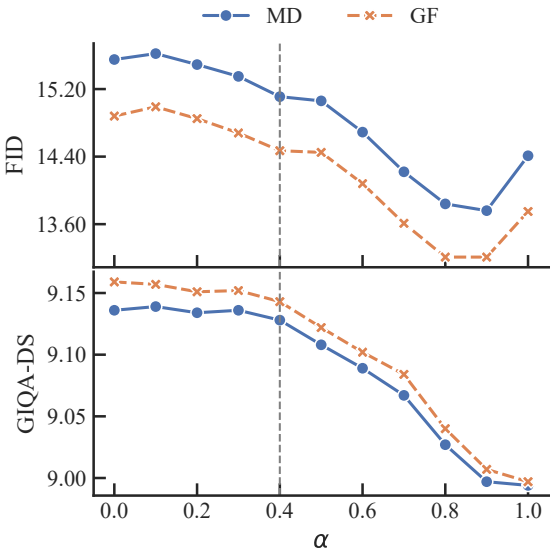

Figure 5: Image quality and diversity assessment using Style Alignment (SA) with different α values. The DDIM sampler is used.

We opted to use a stride of 384 for subsequent experiments because it demonstrated the best image quality and higher computational efficiency. Specifically, when generating a panorama image with a size of 512 × 3,584, employing a stride of 128 requires processing 25 patches, whereas using a stride of 384 requires only nine patches.

High Image Quality Generation using DDPM Sampler with Variance-Corrected Fusion

By examining Table 2, it can be observed that applying DDPM with VCF can produce high-quality and diverse outcomes. The "VCF" rows present substantial improvements to DDIM-based methods. We did not report the result from DDPM applied with MD because it produces blurred images, as shown in the second row of Figure 3. The third row of Figure 3 and Figure 1 shows that DDPM with VCF could generate large images with high contrast and fine details. The "VCF+GF" showing better scores than solely applying VCF indicates that the VCF and GF do not interfere with each other's effectiveness.

The Effectiveness of Style Alignment

For Style Alignment (SA), we use FID and GIQA-DS as the primary metrics to evaluate the quality and diversity of the generated panorama images. We assessed the generated images with α set to 0.0, 0.1, 0.2, ..., and 1.0 for MD and GF with the DDIM sampler. $\alpha = 0.0$ implies the SA is not applied, while $\alpha = 1.0$ initializes the entire large image using repeated reference noise patches. We used a randomly generated standard Gaussian noise as the reference noise to conduct our experiments.

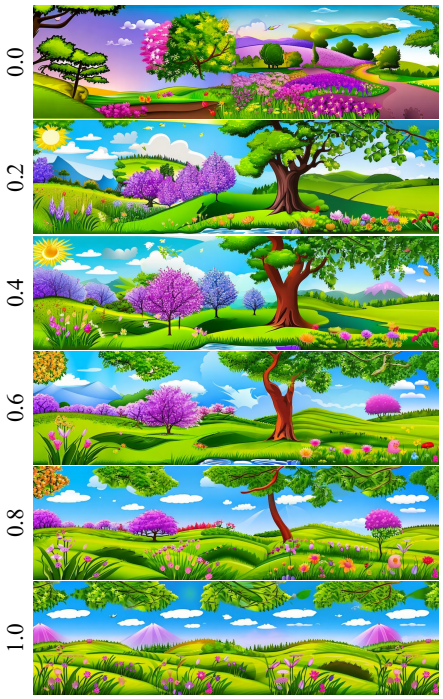

Figure 6: The left half of panorama images generated using Style Alignment (SA) with different α values. DDIM sampler is used.

As shown in Figure 5, with the increase in α, the overall image quality exhibits an upward trend, while diversity shows a downward trend. Figure 6 shows progressive visual results from discontinuous content to the highly repeated pattern generated with increasing values of α. This evidences our assumption: initializing patches with similarity helps to create more coherent content. The trade-off is that as α increases, diversity decreases. We identified the $\alpha = 0.4$ as the optimal value because it balances the quality and diversity. With α larger than 0.4, the diversity drops quickly. The different choices of α provide a control of style consistency that can fit different aesthetic requirements.

It can also be observed from Figure 5 that regardless of the choice of α, applying SA with GF consistently achieves better quality and diversity compared to MD.

As shown in Figure 7, we discovered that when using the same initial noise, the results generated by SyncDiffusion with a 0.1 sync weight and SA with $\alpha = 0.1$ are highly similar to each other but significantly different from MD. In Table 4, we calculated the similarity between the images generated from three methods with DDIM sampler using Structural Similarity Index Measure (SSIM) (Wang et al. 2004), with 2500 panoramic images from each method. The SSIM between SA and SyncDiffusion reached 0.74, indicating that SyncDiffusion and SA produce highly similar outcomes. This implies that SA and SyncDiffusion are potentially equivalent to a specific content. Compared to SyncDiffusion, which uses gradient descent to align patch style at each denoising step, the SA is more computationally efficient as it only performs a one-shot alignment at initial noise. When generating a 3584-width image with the 384 stride, SA takes approximately 8s, while SyncDiffusion requires 102 seconds on a Quadro RTX 6000 card. The computational efficiency makes style control more feasible with the use of SDE samplers, which necessitates more denoising steps. The DDPM sampler requires 1000 denoising steps, which is 20 times longer than a 50-step DDIM sampler.

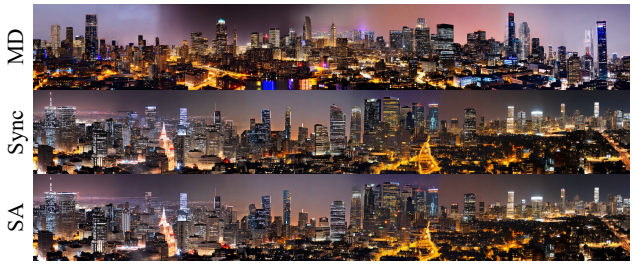

Figure 7: Highly similar results generated by SyncDiffusion (Sync) with a 0.1 sync weight and Style Alignment (SA) with $\alpha = 0.1$. DDIM sampler is used to generate all results.

	MD	MD+SA$_{0.1}$
MD+SA$_{0.1}$	0.30	–
Sync	0.30	0.74

Table 4: SSIM Matrix.

Conclusions

We have revisited joint denoising, which generates large images by creating a series of overlapped patches through small diffusion models. This work addresses the issues presented in the averaging fusion of overlapped regions, e.g., noticeable seams, blurred images, and discontinuous objects. We proposed a novel technique called Guided Fusion (GF), which reduces the disruption to the denoised image by assigning higher weights to the central region of each image patch, allowing the fused values in overlapped regions to be predominantly determined by the geometrically closer patch. Additionally, we presented Variance-Corrected Fusion (VCF), which adjusts the variance of the averaged values to enable its application with SDE samplers, such as DDPM. Furthermore, we introduced the Style Alignment (SA) that eases the fusion process by controlling the similarity of the initial noise, resulting in more coherent images.

Qualitative and quantitative experimental results demonstrate that the proposed methods effectively enhance the quality of the generated images. Our proposed approaches can be broadly applied to other joint denoising-based methods to achieve superior fusion outcomes. For instance, the high-resolution image generation approaches, ScaleCrafter (He et al. 2024) and DemoFusion (Du et al. 2024), both rely on MD to fuse the overlaps. Our approaches offer a potential enhancement for these approaches.

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, 1737–1752. Honolulu, Hawaii, USA: JMLR.org.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*.

Du, R.; Chang, D.; Hospedales, T.; Song, Y.-Z.; and Ma, Z. 2024. DemoFusion: Democratising High-Resolution Image Generation With No $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6159–6168.

Gu, S.; Bao, J.; Chen, D.; and Wen, F. 2020. GIQA: Generated Image Quality Assessment. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, volume 12356, 369–385. Cham: Springer International Publishing. ISBN 978-3-030-58620-1 978-3-030-58621-8. Series Title: Lecture Notes in Computer Science.

Haji-Ali, M.; Balakrishnan, G.; and Ordonez, V. 2024. ElasticDiffusion: Training-free Arbitrary Size Image Generation through Global-Local Content Separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6603–6612.

He, Y.; Yang, S.; Chen, H.; Cun, X.; Xia, M.; Zhang, Y.; Wang, X.; He, R.; Chen, Q.; and Shan, Y. 2024. ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. ArXiv:1312.6114 [cs, stat].

Lee, Y.; Kim, K.; Kim, H.; and Sung, M. 2023. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36: 50648–50660.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11410–11420.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shoemake, K. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques - SIGGRAPH '85*, 245–254. Not Known: ACM Press. ISBN 978-0-89791-166-5.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612. Conference Name: IEEE Transactions on Image Processing.