MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation

Omer Bar-Tal [*1] **Lior Yariv** [*1] **Yaron Lipman** [1 2] **Tali Dekel** [1]

Abstract

Recent advances in text-to-image generation with diffusion models present transformative capabilities in image synthesis. However, user controllability of the generated image, and fast adaptation to new tasks still remains an open challenge, currently mostly addressed by costly and long retraining and fine-tuning or ad-hoc adaptations to specific image generation tasks. In this work, we present *MultiDiffusion*, a unified framework that enables versatile and controllable image generation, using a pre-trained text-to-image diffusion model, without any further training or finetuning. At the center of our approach is a new generation process, based on an optimization task that binds together multiple diffusion generation processes with a shared set of parameters or constraints. We show that MultiDiffusion can be readily applied to generate high quality and diverse images that adhere to user-provided controls, such as desired aspect ratio (e.g., panorama), and spatial guiding signals, ranging from tight segmentation masks to bounding boxes.

1. Introduction

Text-to-image generative models have emerged as a "disruptive technology", demonstrating unprecedented capabilities in synthesizing high-quality and diverse images from text prompts, where diffusion models are currently established as state-of-the-art (Saharia et al., 2022b; Ramesh et al., 2022; Rombach et al., 2022; Croitoru et al., 2022). While this progress holds a great promise in changing the way we can create digital content, deploying text-to-image models to real-world applications remains challenging due to the difficulty to provide users with intuitive control over the generated content. Currently, controllability over diffusion models is achieved in one of two ways: (i) training a model from scratch or finetuning a given diffusion model for the task at hand (e.g., inpainting, layout-to-image training, etc. (Wang et al., 2022a; Ramesh et al., 2022; Rombach et al., 2022; Nichol et al., 2021; Avrahami et al., 2022b; Brooks et al., 2022; Wang et al., 2022b)). With the ever-increasing scale of models and training data, this approach often requires *extensive compute* and *long development period*, even in a finetuning setting. (ii) Reuse a pre-trained model and add some controlled generation capability. Previously, these methods have concentrated on specific tasks and designed a tailored methodology (e.g., replacing objects in an image, manipulating style, or controlling layout (Tumanyan et al., 2022; Hertz et al., 2022; Avrahami et al., 2022a)).

The goal of this work is to design *MultiDiffusion*, a new unified framework that significantly increases the flexibility in adapting a pre-trained (reference) diffusion model to controlled image generation. The basic idea behind the MultiDiffusion is to define a *new generation process* that is composed of several reference diffusion generation processes binded together with a set of shared parameters or constraints. In more detail, the reference diffusion model is applied to different regions in the generated image, predicting a denoising sampling step for each. In turn, the MultiDiffusion takes a global denoising sampling step reconciling all these different steps via least squares optimal solution.

For example, consider the task of generating an image at arbitrary aspect ratio given a reference diffusion model trained on square images (Fig. 2). At each denoising step, the MultiDiffusion fuses the denoising directions, provided by the reference model, from *all* the square crops, and strives to follow them all as closely as possible, constrained by the fact that nearby crops share common pixels. Intuitively, we encourage each crop to be a real sample from the reference model. Note that while each crop might pull to a different denoising direction, our framework yields a unified denoising step, hence produces high-quality and seamless images.

With MultiDiffusion, we are able to harness a reference pre-trained text-to-image to different applications including synthesizing images at desired resolution or aspect ratio, or

[*]Equal contribution [1]Weizmann Institute of Science [2]Meta AI. Correspondence to: Omer Bar-Tal <omer.bar-tal@weizmann.ac.il>, Lior Yariv <lior.yariv@weizmann.ac.il>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Project page is available at https://multidiffusion.github.io.

Figure 1. MultiDiffusion enables flexible text-to-image generation, unifying multiple controls over the generated content, including desired aspect ratio, or simple spatial guiding signals such as rough region-based text-prompts.

synthesizing images using rough region-based text prompts, as seen in Fig. 1. Notably, our framework allows to solve these tasks *simultaneously*, using a common generation process. Comparing to relevant baselines, we found that our approach is able to produce state-of-the-art controlled generation quality even compared to methods that are specifically trained for these tasks. Furthermore, our method works efficiently, without introducing computational overhead.

2. Related Work

Diffusion Models Diffusion models (Sohl-Dickstein et al., 2015; Croitoru et al., 2022; Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021) are a class of generative probabilistic models that aim to approximate a data distribution q, and are easy to sample from. Specifically, these models take a Gaussian noise input $I_T \sim \mathcal{N}(0, I)$, and through a series of gradual denoising steps, transform it into a sample I_0, that should be distributed according to q. The number of denoising steps, and the parameterization of the transformation varies among different works (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020; Lu et al., 2022a;b; Liu et al., 2022). Recently, Diffusion Models have emerged as state-of-the-art generators due to their success in learning complex distributions and generating diverse high quality samples. These models have been successfully used in various domains, including images (Dhariwal & Nichol, 2021; Nichol & Dhariwal, 2021; Saharia et al., 2022b; Ramesh et al., 2022; Rombach et al., 2022), video (Ho et al., 2022; Singer et al., 2022), 3D scenes (Müller et al., 2022), and motion sequences (Yuan et al., 2022; Tevet et al., 2022).

Controllable generation with diffusion models Diffusion models can be trained with guiding input channels (e.g., semantic layout, category label) and successfully perform conditional image generation (Ramesh et al., 2021; Saharia et al., 2022c;a; Wang et al., 2022a; Preechakul et al., 2022; Ho & Salimans, 2022). The most prominent example of conditional diffusion models is recent text-to-image diffusion models, which have demonstrated groundbreaking synthesis capabilities (Nichol et al., 2021; Saharia et al., 2022b; Ramesh et al., 2022; Nichol et al., 2021; Rombach et al., 2022; Sheynin et al., 2022). However, these models provide only little control over the generated content, which is mainly achieved through the input text. Recently, a surge of methods have been proposed to gain wider and better user controllability. Existing methods can be roughly divided into two main approaches: (i) methods that incorporate explicit control by using additional guiding signals to the model (Avrahami et al., 2022b; Rombach et al., 2022; Brooks et al., 2022). However, these works require costly extensive training on curated datasets. (ii) On the other side of the spectrum, numerous methods proposed to implicitly control the generated content by manipulating the generation process of a pre-trained model (Kwon & Ye, 2022;

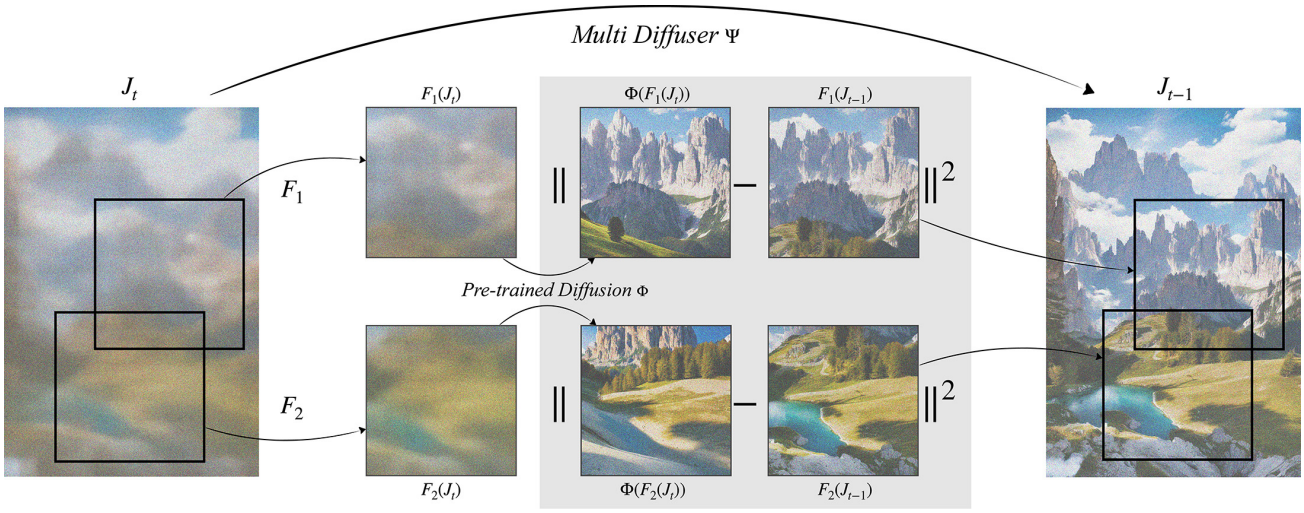

Figure 2. MultiDiffusion: a new generation process, Ψ, is defined over a pre-trained reference model Φ. Starting from a noise image J_T, at each generation step, we solve an optimization task whose objective is that *each* crop $F_i(J_t)$ will follow as closely as possible its denoised version $\Phi(F_i(J_t))$. Note that while each denoising step $\Phi(F_i(J_t))$ may pull to a different direction, our process fuses these inconsistent directions into a *global* denoising step $\Phi(J_t)$, resulting in a high-quality seamless image.

Meng et al., 2021; Tumanyan et al., 2022; Hertz et al., 2022; Avrahami et al., 2022c; Choi et al., 2021; Mokady et al., 2022; Couairon et al., 2022; Kong et al., 2023; Kwon et al., 2022) or by performing lightweight model finetuning (Ruiz et al., 2022; Kawar et al., 2022; Kim et al., 2022; Valevski et al., 2022). Avarahami *et al.* designed image inpainting methods (Avrahami et al., 2022a;c) that do not require finetuning. Recent works (Tumanyan et al., 2022; Hertz et al., 2022) rely on architectural properties and insights about the internal features of the pretrained model, and tailor image editing techniques accordingly. Our work also manipulates the generation process of a pretrained diffusion model, and does not require any training or finetuning. However, in contrast to existing works that target a specific application, without a well defined objective, we propose a more general approach that allows us to unify different user control inputs in a more principled manner.

3. Method

We consider a pre-trained diffusion model, which serves as a reference model:

$$\Phi : \mathcal{I} \times \mathcal{Y} \to \mathcal{I}$$

working in image space $\mathcal{I} = \mathbb{R}^{H \times W \times C}$ and condition space $\mathcal{Y}$, e.g., $y \in \mathcal{Y}$ is a text prompt. Initializing $I_T \sim P_{\mathcal{I}}$, where $P_{\mathcal{I}}$ represents the distribution of Gaussian i.i.d. pixel values, and setting a condition $y \in \mathcal{Y}$, the diffusion model builds a sequence of images,

$$I_T, I_{T-1}, \ldots, I_0 \quad \text{s.t.} \quad I_{t-1} = \Phi(I_t|y) \tag{1}$$

gradually transforming the noisy image I_T into a clean image I_0.

MultiDiffusion. Our goal is to leverage Φ to generate images in a potentially different image space $\mathcal{J} = \mathbb{R}^{H' \times W' \times C}$ and condition space $\mathcal{Z}$, without any training or finetuning. To do so, we define a *MultiDiffusion process*, defined by a function, called *MultiDiffuser*,

$$\Psi : \mathcal{J} \times \mathcal{Z} \to \mathcal{J}$$

The MultiDiffusion, similarly to a diffusion process, starts with some initial noisy input $J_T \sim P_{\mathcal{J}}$, where $P_{\mathcal{J}}$ is a noise distribution over $\mathcal{J}$, and produces a series of images

$$J_T, J_{T-1}, \ldots, J_0 \quad \text{s.t.} \quad J_{t-1} = \Psi(J_t|z) \tag{2}$$

Our key idea is to define Ψ to be *as-consistent-as-possible* with Φ. More specifically, we define a set of mappings between the target and reference image spaces $F_i : \mathcal{J} \to \mathcal{I}$, and a corresponding set of mappings between the condition spaces: $\lambda_i : \mathcal{Z} \to \mathcal{Y}$ where $i \in [n] = \{1, \ldots, n\}$. These mappings are application depended, as will be described later in Sec. 4. Our goal is to make every MultiDiffuser step $J_{t-1} = \Psi(J_t|z)$ follow as closely as possible $\Phi(I_t^i|y_i)$, $i \in [n]$, i.e., the denoising steps of Φ when applied to the images and conditions:

$$I_t^i = F_i(J_t), \quad y_i = \lambda_i(z)$$

Formally, our new process is given by solving the following optimization problem:

$$\Psi(J_t|z) = \arg\min_{J\in\mathcal{J}} \mathcal{L}_{\text{FTD}}(J|J_t, z) \tag{3}$$

$$\mathcal{L}_{\text{FTD}}(J|J_t, z) = \sum_{i=1}^{n} \left\| W_i \otimes \left[F_i(J) - \Phi(I_t^i|y_i) \right] \right\|^2 \tag{4}$$

where $W_i \in \mathbb{R}_{\geq 0}^{H\times W}$ are per pixel weights and $\otimes$ is the Hadamard product. Intuitively, the FTD loss reconciles, in the least-squares sense, the different denoising sampling steps, $\Phi(I_t^i|y_i)$, suggested on different regions, $F_i(J_t)$, of the generated image J_t. Fig. 2 illustrates one step of the MultiDiffuser; Algorithm 2 recaps the MultiDiffusion sampling process.

Closed-form formula. In the applications demonstrated in this paper F_i consist of direct pixel samples (e.g., taking a crop out of image J_t). In this case, Eq. 4 is a quadratic Least-Squares (LS) where each pixel of the minimizer J is a weighted average of all its diffusion sample updates, i.e.,

$$\Psi(J_t|z) = \sum_{i=1}^{n} \frac{F_i^{-1}(W_i)}{\sum_{j=1}^{n} F_j^{-1}(W_j)} \otimes F_i^{-1}(\Phi(I_t^i|y_i)) \tag{5}$$

Properties of MultiDiffusion. The main motivation for the definition of Ψ in Eq. 3 comes from the following observation: If we choose a probability distribution $P_{\mathcal{J}}$ such that

$$F_i(J_T) \sim P_{\mathcal{I}}, \qquad \forall i \in [n] \tag{6}$$

and compute $J_{t-1} = \Psi(J_t|z)$, as defined in Eq. 3, where we reach a zero FTD loss, $\mathcal{L}_{\text{FTD}}(J_{t-1}|J_t, z) = 0$, then:

$$I_{t-1}^i = F_i(J_{t-1}) = \Phi(I_t^i|y_i)$$

That is, I_t^i, for all $i \in [n]$, is a diffusion sequence and thus I_0^t is distributed according to the distribution defined by Φ over the image space $\mathcal{I}$. We summarize

Proposition 3.1. *If $P_{\mathcal{J}}$ is a distribution over $\mathcal{J}$ satisfying Eq. 6, and the FTD cost (Eq. 4) is defined with $W_i \in \mathbb{R}_{>0}^{H\times W}$ and minimized to zero in Eq. 3 for all steps $T, T-1, \ldots, 0$, then the images $I_t^i = F_i(J_t)$ reproduce a Φ diffusion path. In particular $F_i(J_0)$, $i \in [n]$ are distributed identically to samples from the reference diffusion model Φ.*

The implications of this proposition are far reaching: using a single reference diffusion process we can flexibly adapt to different image generation scenarios without the need to retrain the model, while still being consistent with the reference diffusion model. Next, we instantiate this framework outlining several application of the Follow-the-Diffusion-Paths approach.

(a) Generation with per-crop independent diffusion paths.

(b) Generation with fused diffusion paths using MultiDiffusion.

Figure 3. Independent diffusion paths vs. MultiDiffusion. (a) Panoramic image generated by applying the reference model on four crops independently; as expected, there is no coherency between the crops. (b) Starting from the same noise, our generation process steers these initial diffusion paths into a consistent and high quality image.

Algorithm 1 MultiDiffusion sampling.

Input : Φ ▷ pre-trained Diffusion Model
$\{F_i\}_{i=1}^n$ ▷ image space mappings
$\{y_i\}_{i=1}^n$ ▷ text-prompts conditioning
$\{W_i\}_{i=1}^n$ ▷ per-pixel weights
$J_T \sim P_{\mathcal{J}}$ ▷ noise initialization
for $t = T, ..., 1$ **do**
  $I_{t-1}^i \leftarrow \Phi(F_i(J_t), y_i) \quad \forall i \in [n]$ ▷ diffusion updates
  $J_{t-1} \leftarrow \texttt{MultiDiffuser}(\{I_{t-1}^i\}_{i=1}^n)$ ▷ Eq. 5
Output : J_0

4. Applications

4.1. Panorama

As a first instantiation we use our framework to define a diffusion model in an image space $\mathcal{J}$ with $H' \geq H, W' \geq H$ directly from a trained model Φ working in image space $\mathcal{I}$. Let $\mathcal{Z} = \mathcal{Y}$ (namely, generating a panoramic image for a given text-prompt), $F_i(J) \in \mathcal{I}$ is an $H \times W$ crop of image J, and $z = \lambda_i(z)$. We consider n such crops that cover the original images J. Setting $W_i = \mathbf{1}$, we get

$$\Psi(J_t, z) = \arg\min_{J\in\mathcal{J}} \sum_{i=1}^{n} \|F_i(J) - \Phi(F_i(J), z)\|^2 \tag{7}$$

that is a least-squares problem, the solution of which is calculated analytically according to Eq. 5. See the Appendix C.1 for implementation details.

As discussed in Sec. 3, *MultiDiffusion* reconciles multiple diffusion paths provided by the reference model Φ. We illustrate this property in Fig. 3, where we consider a panorama of $H \times 4W$. Fig. 3(a) shows the generation result when in-

dependently applying Φ on four non-overlapping crops. As expected, there is no coherency between the crops since this amounts to four random samples from the model. Starting from the same initial noise, our generation process (Eq. 4.1), allows us to fuse these initially-unrelated diffusion paths, and steer the generation into a high-quality, coherent panorama (b).

4.2. Region-based text-to-image-generation

Given a set of region-masks $\{M_i\}_{i=1}^n \subset \{0,1\}^{H\times W}$ and a corresponding set of text-prompts $\{y_i\}_{i=1}^n \subset \mathcal{Y}^n$, our goal is to generate a high-quality image $I \in \mathcal{I}$ that depicts the desired content in each region. That is, the image segment $I \otimes M_i$ should manifest y_i. Going back to our formulation (Eq. 2), the *MultiDiffusion* process is defined over the condition space $\mathcal{Z} = \mathcal{Y}^n$, i.e., $z = (y_1, \ldots, y_n)$, and the target image space $\mathcal{J} = \mathcal{I}$ is identical to the reference one:

$$\Psi : \mathcal{I} \times \mathcal{Y}^n \to \mathcal{I}$$

Furthermore, the region selection maps are defined as $F_i(I) = I$, the pixel weights are set according to the masks, $W_i = M_i$, and the Ψ step is defined as the solution to the least-squares problem:

$$\Psi(J_t, z) = \arg\min_{J\in\mathcal{I}} \sum_{i=1}^{n} \left\| M_i \otimes \Big[J - \Phi(J_t|y_i) \Big] \right\|^2 \quad (8)$$

The solution to this LS problem is calculated analytically. At each step we apply the pretrained diffusion w.r.t. each of the given prompts, resulting in multiple diffusion directions $\Phi(J_t|y_i)$. We encourage each pixel in J_t to follow the (averaged) directions associated with the regions M_i containing it.

Fidelity to tight masks We further support obtaining high-fidelity to tight masks if provided by the user (see Fig. 5). We noticed that the layout is being determined early on in the diffusion process, and thus we strive to encourage $\Phi(J_t|y_i)$ to focus on the region M_i early on in the process in order to match the desired layout, and to consider the full context in the image next, to achieve an harmonized result. We integrate time dependency in the maps F_i, introducing a bootstrapping phase. That is,

$$F_i(J_t, t) = \begin{cases} J_t, & \text{if } t \le T_{init} \\ M_i \otimes J_t + (1 - M_i) \otimes S_t, & \text{otherwise} \end{cases} \quad (9)$$

Where T_{init} is the bootstrapping stopping step parameter, and S_t is a random image with a constant color, which serves as background (see Appendix C.2 for implementation details).

We demonstrate the efficacy of our bootstrapping approach in Sec. 5.2. We set T_{init} to be 20% of the generation process (i.e., $T_{init} = 800$).

5. Results

We thoroughly evaluate our method when applied to each task as discussed in Sec. 4. In all experiments, we used Stable Diffusion (Rombach et al., 2022), where the diffusion process is defined over a latent space $\mathcal{I} = \mathbb{R}^{64\times64\times4}$, and a decoder is trained to reconstruct natural images in higher resolution $[0,1]^{512\times512\times3}$. Similarly, the MultiDiffusion process, Ψ is defined in the latent space $\mathcal{J} = \mathbb{R}^{H`\times W`\times4}$ and using the decoder we produce the results in the target image space $[0,1]^{8H'\times8W'\times3}$.

5.1. Panorama Generation

To evaluate our method on the task of text-to-panorama generation (Sec. 4.1), we generated a diverse set of 512×4608 panoramas, $\times 9$ wider than the original training resolution. Since there is no direct method for generating images at arbitrary aspect ratio from text, we compare to the following two baselines: (i) Blended Latent Diffusion (BLD) (Avrahami et al., 2022a) (combined with Stable Diffusion (Rombach et al., 2022)), and Stable Inpainting (SI) (Rombach et al., 2022), which has been finetuned on large-scale data for inpainting. For both baselines, the panoramic image is generated gradually, starting from a central image (sampled by Φ given the input text), and extrapolated progressively to the right and left.

Fig. 4 shows sample generation results by our method compared to the above baselines. As seen, both baselines often exhibit visible seams and discontinuities between overlapping crops, as well as degradation in visual quality as moving away from the center pivotal image; this is expected due to the iterative generation process. BLD often generates repetitive content (e.g., skiers example), where SI results in noticeable visual difference between the left and right parts of the image. In contrast, our framework *simultaneously* "samples" the panoramic image by combining the diffusion paths of *all* crops, resulting in seamless and high quality images. Additional comparisons are in the Appendix 11.

	FID ↓	CLIP-score↑	CLIP-aesthetic↑
Stable Diffusion	6.05 ± 3.1	0.27	6.36
SI	45.5 ± 14.5	0.26	5.76
BLD	18.4 ± 7.4	**0.27**	6.02
Ours	$\mathbf{10.3 \pm 4.8}$	**0.27**	**6.36**

Table 1. Panorama generation evaluation. We report FID, CLIP text-image score, and CLIP aesthetic scores for of our method compared to the baselines. See more details in Section. 5.1.

To quantify these observations, we use the Frechet Inception Distance (FID) (Parmar et al., 2022) to measure the distance between the distribution of 512×512 crops from the panoramic images to the distribution of images generated by the reference model Φ. That is, for a given text prompt, we

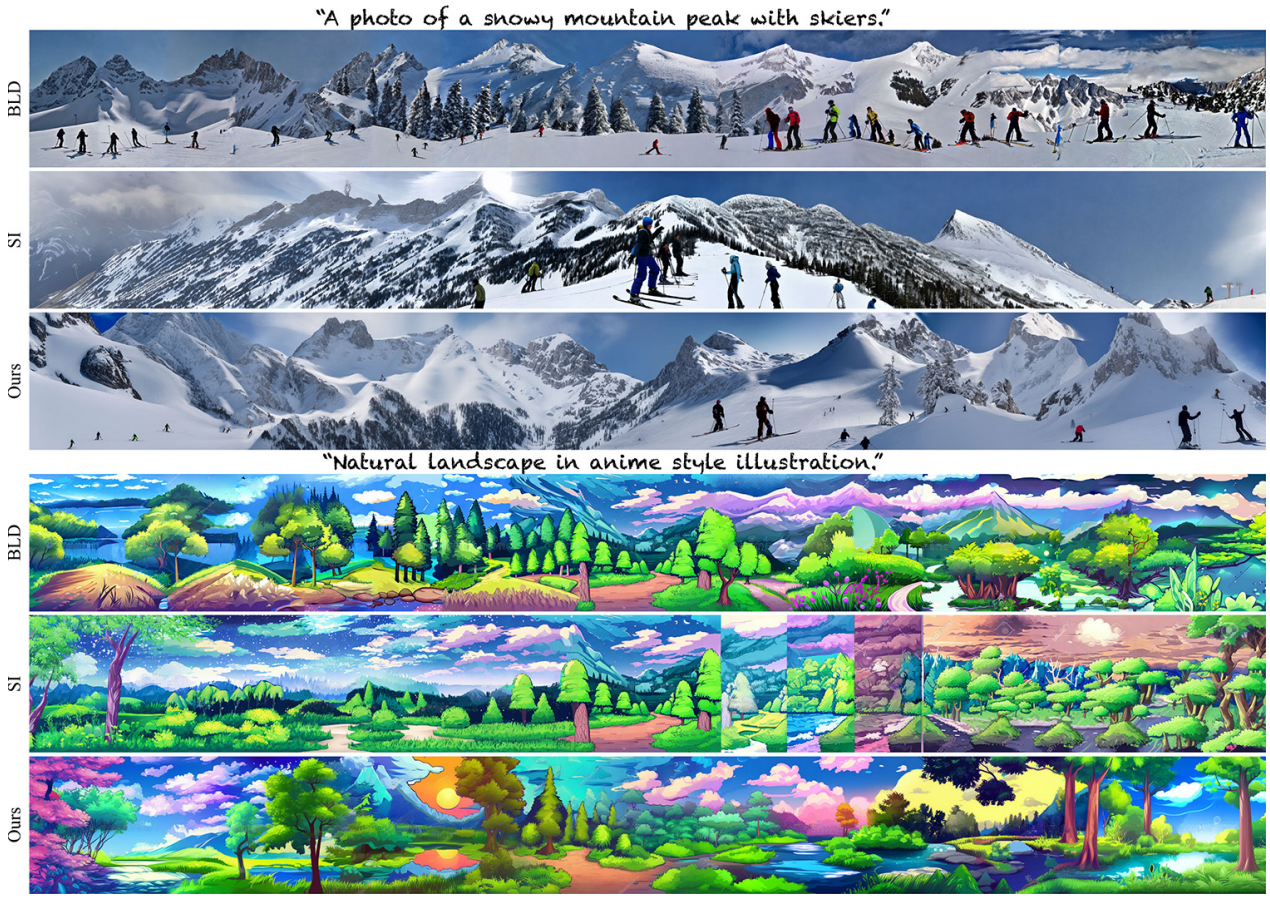

Figure 4. Text-to-Panorama comparison to Blended Latent Diffusion (BLD) (Avrahami et al., 2022a) and Stable Inpainting (SI) (Rombach et al., 2022). Our framework produces seamless and diverse content whereas the baselines either contain repetitive content, visible seams or artifacts.

sample N different 512×512 images from Φ, and consider them as our reference dataset. For the baselines and our method, we generated N panoramic images, and randomly sampled a 512×512 crop from each sample to serves as the generated dataset and computed the FID accordingly.

To further assess the quality of our results, we evaluated two CLIP-based scores: (i) text-image CLIP score (Radford et al., 2021) measured by the cosine similarity between the text prompt and the image embeddings, and (ii) CLIP aesthetic (Schuhmann et al., 2022) measured by a linear estimator on top of CLIP predicting the aesthetic quality of the images.

We used $N = 2000$ samples and repeated this evaluation for 8 different text-conditioning. Table 1 reports the mean and standard-deviation of FID and CLIP scores for our method and the baselines. We additionally report the scores for an independent set of samples images from Φ, which serves as a baseline. As seen, our method outperforms the existing baselines in all metrics.

5.2. Region-based Text-to-Image Generation

Our region-based formulation (Sec. 4.2) allows novice users greater flexibility in their content creation, by lifting the burden of creating accurate tight masks. As can be seen in Fig. 1, Fig. 7 and Fig. 8, our method generates diverse high-quality samples that comply with text description, given only bounding boxes region guidance. As seen in Fig. 7, by starting our generation from a different input noise, we can generate diverse samples, depicting objects in different scales and appearances, all following the same spatial controls. Notably, since we integrate the controls from all regions into a unified generation process, our method can generate complex scene effects (e.g., background blur, shadows or reflections) which are coherently immerse in the scene. More results are included in the Appendix.

We compare our region-based framework with Make-A-Scene (Gafni et al., 2022) and the concurrent work SpaText (Avrahami et al., 2022b). Both baselines perform large-scale training specifically for this task. Note that these models are not publicly available, thus we qualitatively compare to their provided examples.

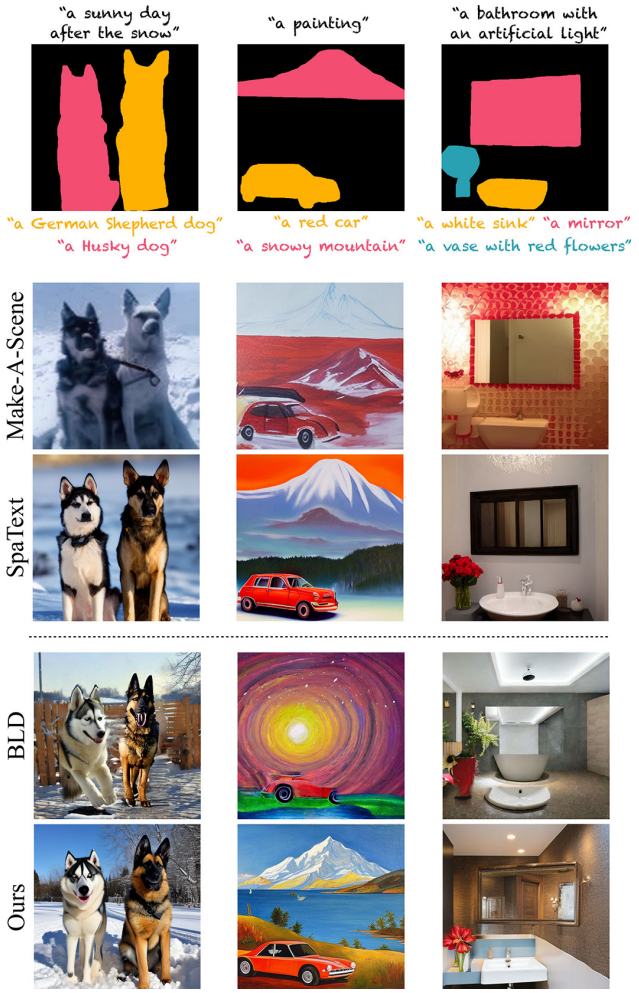

Figure 5. Region-based text-to-image generation. The input segmentation maps with the corresponding region text descriptions are shown above each example. Below: Make-A-Scene (Gafni et al., 2022)and SpaText (Avrahami et al., 2022b) – trained specifically for this task on a large-scale segmentation-text-image dataset; Blended Latent Diffusion (BLD) (Avrahami et al., 2022a), and our results.

Additionally, we consider an adaptation of BLD (Avrahami et al., 2022a) as a baseline. Similarly to Sec. 5.1, this is done by applying their method in an auto-regressive manner by first generating the background, and sequentially generating each of the foreground objects.

As seen in Fig. 5, our framework produces consistent images that adhere to the spatial constraints, and are qualitatively on par with (Avrahami et al., 2022b). The auto-regressive approach based on BLD (Avrahami et al., 2022a) often results in incoherent images and an unnatural scene. (e.g., misplaced sink in "bathroom" example). Additional comparisons to the baselines are in the Appendix.

	IoU ↑
COCO dataset	0.43 ± 0.09
SI	0.16 ± 0.10
BLD	0.17 ± 0.11
Ours *w/o bootstrapping*	0.18 ± 0.10
Ours	$\mathbf{0.26 \pm 0.12}$

Table 2. Region-based generation evaluation of the COCO dataset. We evaluate Intersection over Union (IoU), see Sec. 5.2 for details.

To quantitatively evaluate our performance, we use the COCO dataset (Lin et al., 2014), which contains images with global text caption and instance masks for each object in the image. We apply our method on a subset from the validation set, obtained by filtering examples which consists of 2 to 4 foreground objects, excluding people, and masks that occupy less than 5% of the image. This results in $1K$ diverse samples. Following (Avrahami et al., 2022b), we use the ground truth labels to provide a text prompt for each foreground region, i.e., "a {label}", and use the full image caption as the prompt describing the background.

We evaluate the results with an off-the-shelf segmentation model (Cheng et al., 2022) on the generated images, and measure the Intersection over Union (IoU) w.r.t. to the ground-truth segmentation. Table 2 reports the performance for our method and the baselines described above. As an upper bound, we also report the IoU w.r.t. the original images in the set. Note that our method outperforms the existing baselines SI (Rombach et al., 2022) and BLD (Avrahami et al., 2022a). We additional provide qualitative examples are included in the Appendix.

Finally, we present an ablation of our bootstrapping stage (Eq. 9): qualitatively in Fig. 6, and quantitatively in Table 2. Note that without bootstrapping, our framework still generates the desired object within the mask region, however, the bootstrapping stage makes it tighter to the given mask.

Figure 6. Bootstrapping ablation. Without bootstrapping (middle), our method successfully generates the glass in *some* location inside the mask (left). With our bootstrapping mechanism (right), we achieve high-fidelity to the provided tight mask. See Sec 4.2 for details.

Figure 7. Diverse samples generated by our framework, given rough scene layout guidance (left). All images depict sensible composition, scene effects and relative size of objects.

Figure 8. Rough masks. Sample results of our region-based generation approach (see Sec. 4.2). Our method can work with rough masks, the can intuitively be obtained by novice users.

6. Discussion and Conclusions

Controllable generation is one of the major pending challenges with text-to-image diffusion models. We proposed to tackle this challenge from a fundamentally new direction – defining a new generation process on top of a pre-trained and fixed diffusion model. This approach has several key advantages over previous works: (i) it does not require any further training or finetuning, (ii) it can be applied to various different generation tasks, and (iii) our generation process yields an optimization task which can be solved in closed form for many tasks, hence can be computed efficiently, while ensuring convergence to the global optimum of our objective. As for limitations, our method heavily relies on the generative prior of the reference diffusion model, i.e., the quality of our results depends on the diffusion paths provided by the model. Thus, when a "bad" path is chosen by the reference model (e.g., bad seed, or biased text-prompt), our results will be affected as well. In some cases, we can mitigate it by introducing more constraints into our framework (bootstrapping in Sec. 4.2), or prompt-engineering (Fig. 9). We thoroughly evaluated our framework, demonstrating state-of-the-art results even compared to methods that are tailored-trained for specific tasks.

We believe that our work can trigger further future research in harnessing the power of a pre-trained diffusion model in more principled manner. One way forward, for example, is to generalize the MultiDiffusion with a more general optimization problem,

$$\Psi(J_t|z) = \arg\min_{J\in\mathcal{J}} \mathcal{L}_{\text{FTD}}(J|J_t,z) + \mathcal{L}_0(J,J_t,z) \quad \text{s.t.} \quad J \in \mathcal{C}(J_t,z) \tag{10}$$

where $\mathcal{L}_0$ is a cost function and $\mathcal{C}$ is a set of (hard) constrains that control the MultiDiffusion process by incorporating other priors and/or design constraints. This approach provides a further of freedom in designing MultiDiffusion processes.

Figure 9. Our method heavily relies on the prior of the reference diffusion model. Left: our method when applied to a vertical panorama. The reference diffusion model is biased towards adding a waterfall in each viewing crop, resulting with an unnatural scene. Right: we can try to overcome this by adding a region-based constraint.

7. Acknowledgments

LY is supported by a grant from Israel CHE Program for Data Science Research Centers and the Minerva Stiftung. OB is supported by the Israeli Science Foundation (grant 2303/20). The research was supported also in part by a research grant from the Carolito Stiftung (WAIC). We thank Michal Geyer and Dolev Ofri-Amar for proofreading the paper.

References

Avrahami, O., Fried, O., and Lischinski, D. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022a.

Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., and Yin, X. Spatext: Spatio-textual representation for controllable image generation. *arXiv preprint arXiv:2211.14305*, 2022b.

Avrahami, O., Lischinski, D., and Fried, O. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022c.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. November 2022.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV)*, 2022.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

Kim, G., Kwon, T., and Ye, J. C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022.

Kong, C., Jeon, D., Kwon, O., and Kwak, N. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Kwon, G. and Ye, J. C. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022a.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022b.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

Múller, N., , Siddiqui, Y., Porzi, L., Rota Buló, S., Kontschieder, P., and Nießner, M. Diffrf: Rendering-guided 3d radiance field diffusion. arxiv, 2022.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Parmar, G., Zhang, R., and Zhu, J.-Y. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

Preechakul, K., Chatthee, N., Wizadwongsa, S., and Suwajanakorn, S. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and 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.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022a.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022b.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022c.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022.

Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., and Taigman, Y. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., and Taigman, Y. Make-a-video: Text-to-video generation without text-video data, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A. H., and Cohen-Or, D. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022.

Valevski, D., Kalman, M., Matias, Y., and Leviathan, Y. Unitune: Text-driven image editing by fine tuning an

image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022.

Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., and Wen, F. Pretraining is all you need for image-to-image translation. In *arXiv*, 2022a.

Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., and Li, H. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022b.

Yuan, Y., Song, J., Iqbal, U., Vahdat, A., and Kautz, J. Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*, 2022.

A. Proof to Proposition 3

Let $P_{\mathcal{J}}$ be a distribution over $\mathcal{J}$ satisfying Eq. 6, and assume that the FTD cost (Eq. 4) is minimized to zero in Eq. 3 for all steps $T, T-1, \ldots, 1$. Notice that Eq. 4 is a sum of squares, hence

$$W_i \otimes \Big[F_i(J_{t-1}) - \Phi(I_t^i|y_i)\Big] = 0 \quad \forall t = T, \ldots, 1$$

Under the assumption that $W_i \in \mathbb{R}_{>0}^{H \times W}$, for each $t \in [T]$, J_{t-1} satisfies

$$F_i(J_{t-1}) = \Phi(I_t^i|y_i)$$

Implying that $F_i(J_0)$ is an image sample from the trajectory $\{\Phi(I_t^i|y_i)\}_{t=1}^T$ with a starting condition $I_T^i \sim P_{\mathcal{I}}$, in particular $F_i(J_0)$ is distributed according to the reference diffusion model Φ.

B. Additional Results

In the following section we provide additional results and comparisons for the applications shown in the main paper.

B.1. Panorama Generation

We provide additional results and qualitative comparisons for the task of text-to-panorama (Sec. 5.1). Fig. 11 depicts additional comparisons of our method vs Stable Inpainting (SI) (Rombach et al., 2022) and Blended Latent Diffusion (BLD) (Avrahami et al., 2022a). We also show vertical panorama result in Fig. 13 left.

B.2. Region-based Text-to-Image Generation

We provide additional qualitative results and comparisons for the task of region-based generation (Sec.4.2) in Fig. 13 and Fig. 10.

B.3. Region-based Text-to-Image Generation on COCO

We include sample results and comparison on the subset from the validation set of COCO in Fig. 14. See more details about this experiment in Sec. 5.2.

C. Additional Implementation Details.

C.1. Panorama (Sec. 4.1)

In the case of panorama generation, our maps F_i are defined as fixed-size crops from the full panorama. Specifically, for a panorama with spatial resolution $H' \times W'$, we consider overlapping crops of size $H \times W$ where $H = W = 64$ defined in the Stable Diffusion latent space (which translates to size 512×512 in RGB space). Our maps $F_i, ..., F_n$ provide crops with a sliding window of size $\texttt{step} = 8$ in the latent space (64 pixels in RGB space). In particular, $n = \frac{H'-64}{\texttt{step}} \cdot \frac{W'-64}{\texttt{step}}$. We summarize,

Algorithm 2 MultiDiffusion sampling - Panorama.

Input : Φ ▷ pre-trained Diffusion Model
H', W' ▷ resolution of the desired panorama
$\{F_i\}_{i=1}^n$ ▷ mappings defining crops from the panorama
y ▷ conditioned text-prompt
$J_T \sim \mathcal{N}(0, I)$ $J_T \in \mathcal{R}^{H' \times W' \times C}$ ▷ noise initialization
for $t = T, ..., 1$ **do**
  $I_t^i \leftarrow F_i(J_t) \ \ \forall i \in [n]$ ▷ take crops from the panorama
  $I_{t-1}^i \leftarrow \Phi(I_t^i, y) \ \ \forall i \in [n]$ ▷ per-crop diffusion updates
  $J_{t-1} \leftarrow \texttt{MultiDiffuser}(\{I_{t-1}^i\}_{i=1}^n)$ ▷ Eq. 5
$\texttt{Panorama} \leftarrow \mathcal{D}(J_0)$ ▷ Decode the panorama to RGB space
Output : $\texttt{Panorama}$

Note that we can compute the per-crop diffusion updates in parallel (i.e., in a batch), resulting in total of $\frac{T \cdot n}{b}$ calls to the reference diffusion Φ, where b denotes the batch size.

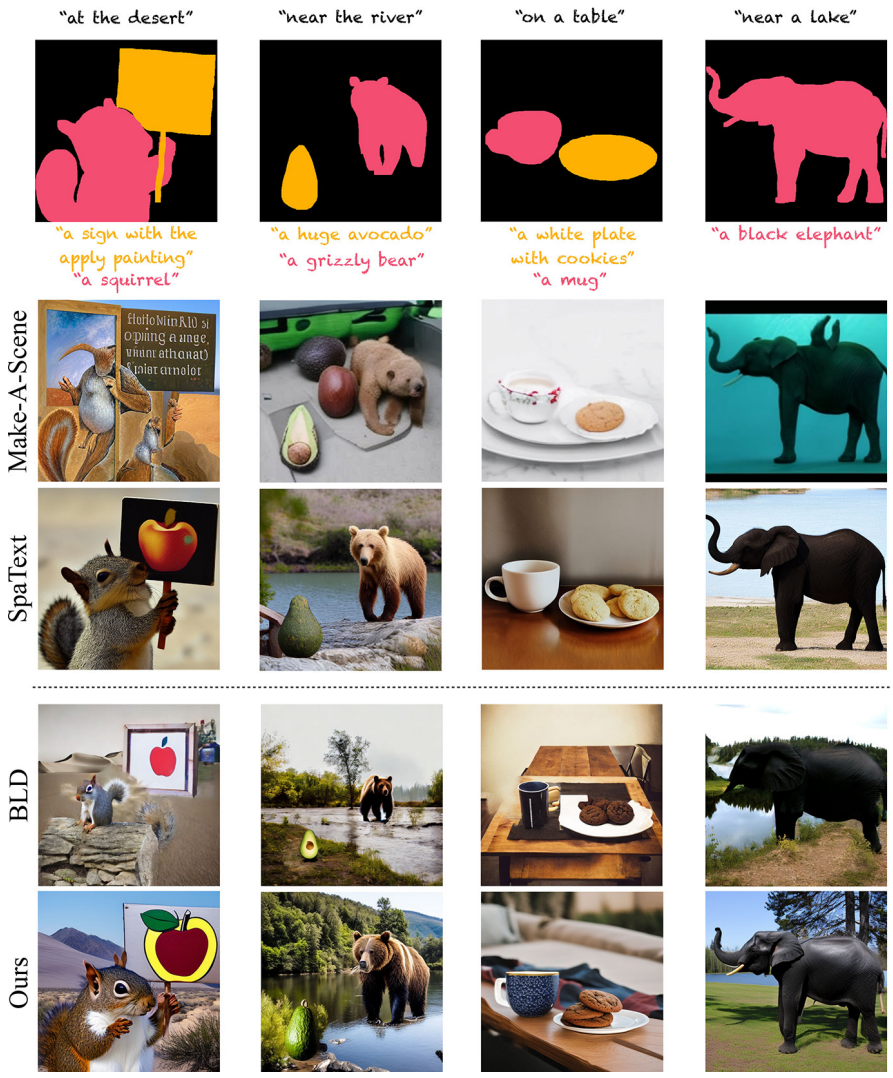

Figure 10. Additional qualitatively comparison to Make-A-Scene (Gafni et al., 2022), Blended-Latent-Diffusion (Avrahami et al., 2022a), Spa-Text (Avrahami et al., 2022b), and ours. See Sec. 4.2, and Sec. 5.2

C.2. Bootstrapping (Sec. 4.2)

In case the user desires to maintain high fidelity to tight masks (see Fig. 4), we introduce a bootstrapping phase to our maps F_i (see Eq. 9). Specifically, we pre-compute each S_t as follows: we randomize an image $I \in [0,1]^{512x512x3}$ with a constant RGB value, and encode it to Stable Diffusion latent space $S = \mathcal{E}(I)$, where $\mathcal{E}$ is the pre-trained encoder provided by the Stable Diffusion framework. Finally, we obtain S_t by noising S to the noise level of time-step t. That is, $S_t \sim \mathcal{N}$ where $(\mu_t \cdot S, \sigma_t^2)$, μ_t and σ_t are the diffusion noise schedulers (Ho et al., 2020).

D. Societal Impact

Our primary goal in this work is to enable novice users to generate visual content in a more intuitive and flexible way. However, it is prone to societal biases the underlying text-based generative model inherits from its training data (Birhane et al., 2021). There is a risk of misuse for creating fake or harmful content with our technology, and we believe that it is crucial to conduct further research on identifying synthetic content and to develop tools for detecting and addressing biases and malicious use cases.

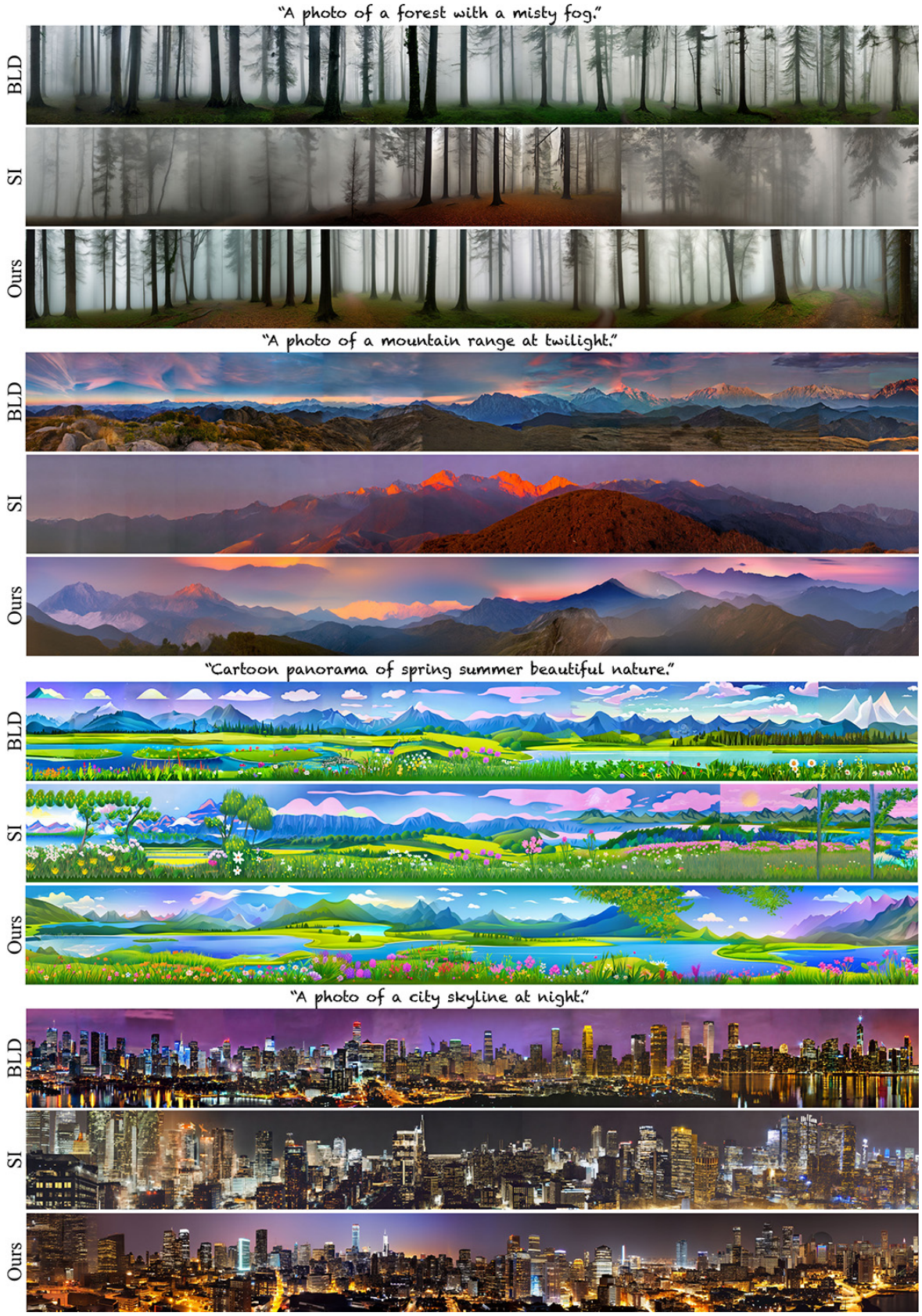

Figure 11. Text-to-Panorama additional results and comparisons to Sec. 5.1.

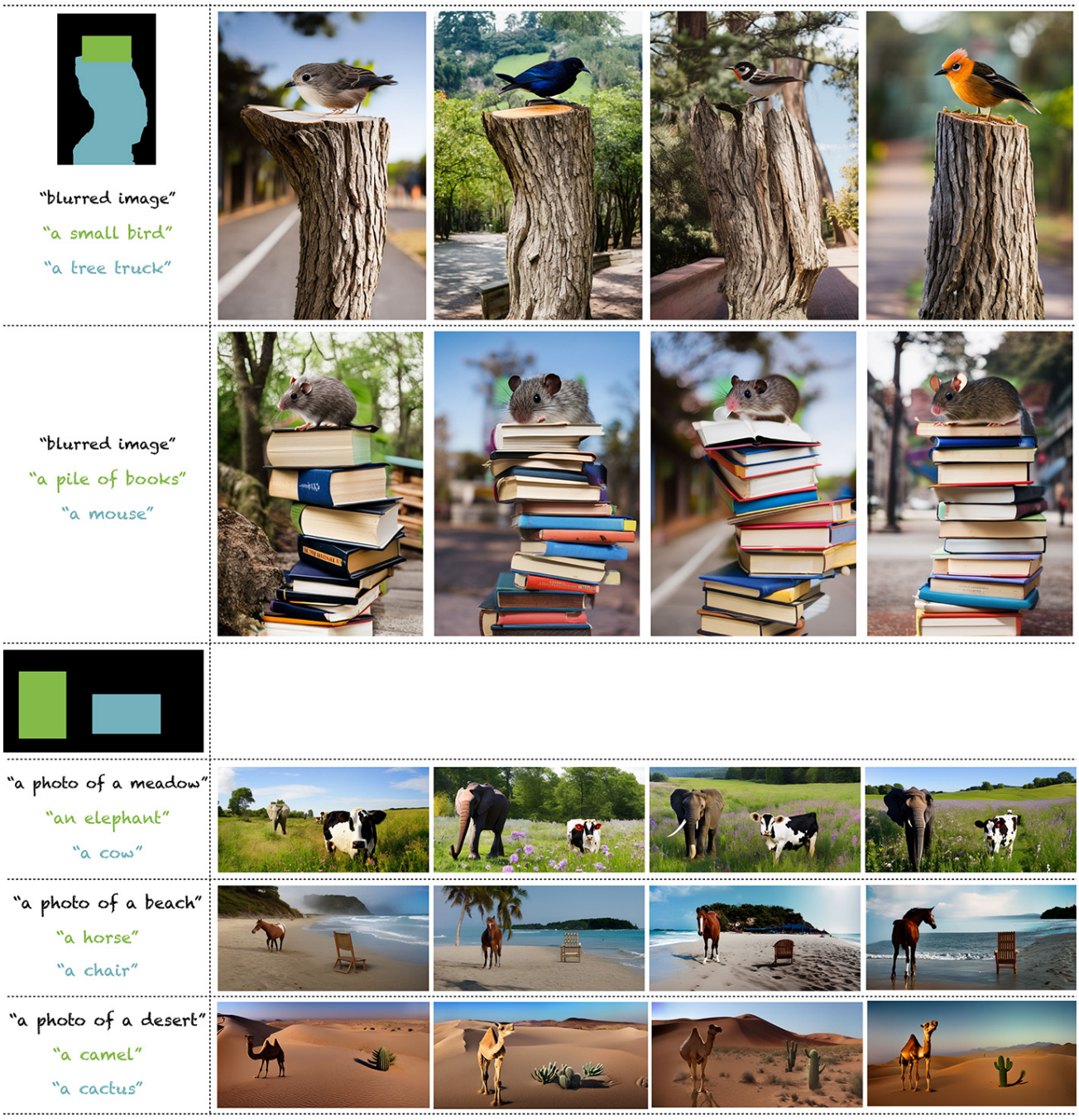

Figure 12. Additional results of our method on generation from rough scene layouts (Sec. 5.2). For each spatial layout, we and for each text prompt, we show different samples from our method.

Figure 13. Left: vertical panoramic image (1024×512) generated by our method. Right: additional generation results using a combination of rough and tight regions; for each layout, we present diverse generated samples.

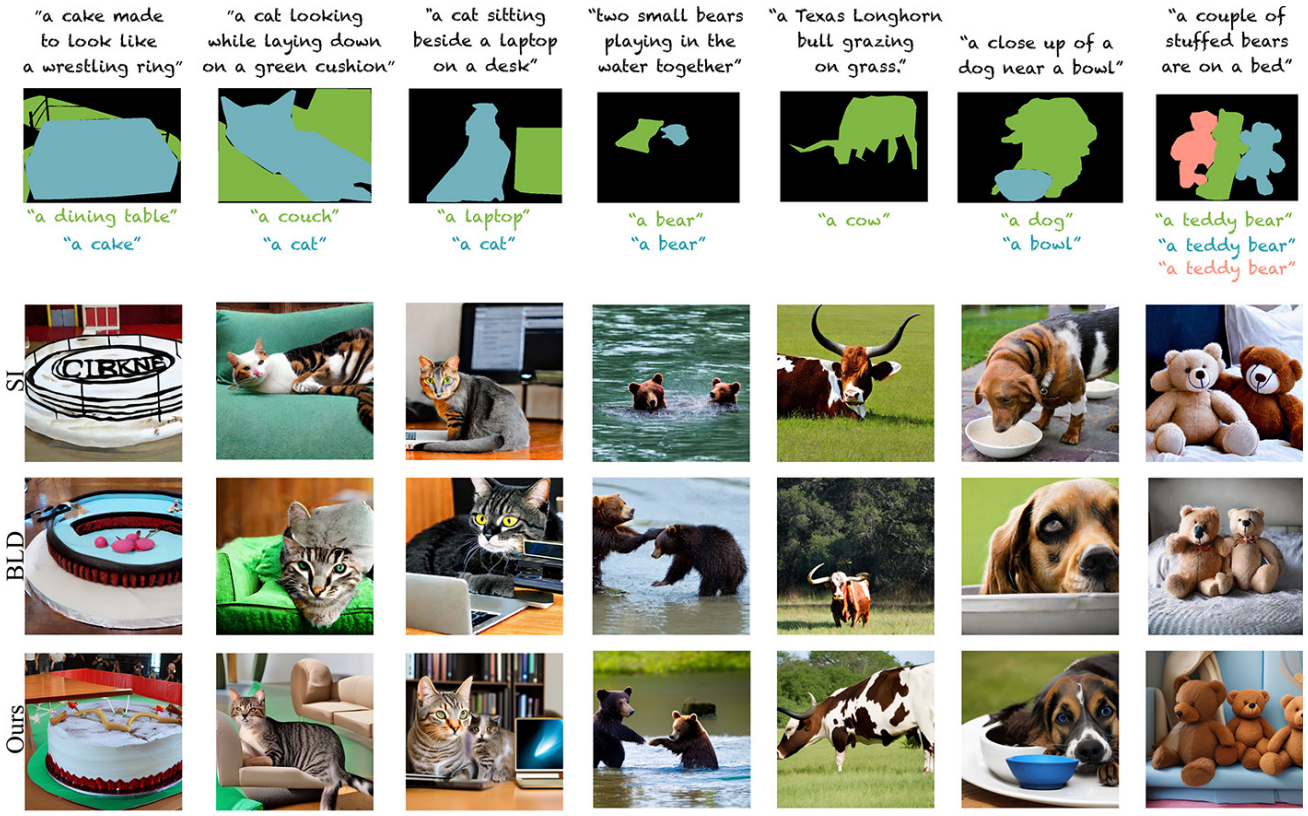

Figure 14. Sample results from COCO validation set by BLD, SI and our method. See more details in Sec. 5.2.