# SyncDiffusion: Coherent Montage via Synchronized Joint Diffusions

**Yuseung Lee** **Kunho Kim** **Hyunjin Kim** **Minhyuk Sung**
KAIST
{phillip0701,kaist984,rlaguswls98,mhsung}@kaist.ac.kr

## Abstract

While ControlNet [8] has enabled pretrained image diffusion models (i.e. Stable Diffusion [4]) to take additional condition inputs, it still has the limitation of only generating images of certain sizes. To this end, we demonstrate an application of SyncDiffusion [2] for conditional image generation, which takes advantage of *synchronized joint diffusions* to allow conditions of arbitrary size as input and in turn generate globally coherent images.

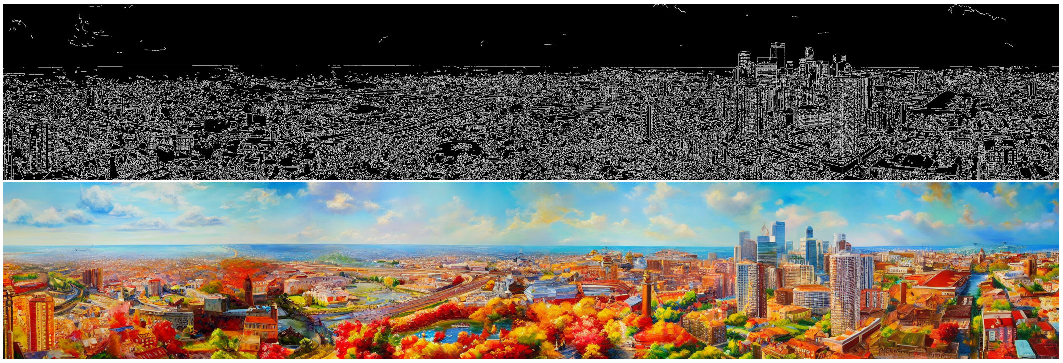


Figure 1: A Canny edge map of $512 \times 3072$ size (top), and a panorama image generated from it along with the prompt *"a beautiful city on a sunny day in oil painting* by our ControlSync.

## 1 Introduction & Related Work

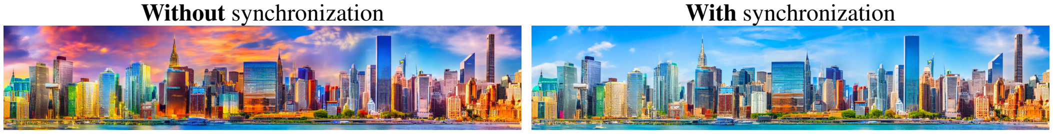
Conditional text-to-image diffusion models [8, 3] have provided controllabilty in image generation by enabling the use of diverse types of additional conditions such as Canny edges and scribbles along with a text prompt. ControlNet [8] is a representative work and has been extended in numerous downstream applications [10, 7, 6]. However, as ControlNet builds upon Stable Diffusion [4] using the same denoising UNet with frozen parameters, it inherits the limitation of Stable Diffusion that the generation resolution is limited to a certain size (namely $512 \times 512$). In this paper, we address this issue and aim to extend the conditional image generation task to *arbitrary* resolutions.

## 2 ControlSync: Extending Conditional Control to Arbitrary Resolutions

We introduce ControlSync, an extension of SyncDiffusion [2] for the task of conditional image generation of *arbitrary* resolutions. Let $c \in \mathbb{R}^{H \times W \times 3}$ denote an input condition (e.g. Canny edge, Hough line) from which the user wants to generate a corresponding RGB image. We focus on the case when the size of the condition differs by a large margin from the training data for the denoising UNet of ControlNet [8] (namely $512 \times 512$). If we have a pretrained ControlNet model $\epsilon_\theta$ trained on the same type of condition as $c$, we can first define a **joint conditional diffusion** following MultiDiffusion [1] (See Alg. 1 line 6-10 in Supplementary). Here, the condition $c$ is cropped with a fixed stride into $N$ windows with overlapping regions. Let function $\psi_{c \to i} : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^{h \times w \times 3}$ map the condition $c$ to the $i$-th cropped window. Then the *one-step denoising* of the $i$-th window with latent $\mathbf{x}_t$ can be defined as:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, c^{(i)}) := q(\mathbf{x}_{t-1} | \mathbf{x}_t, \phi_\theta(\mathbf{x}_t, t, c^{(i)})), \quad c^{(i)} := \psi_{c \to i}(c) \tag{1}$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).


**Without** synchronization

**With** synchronization

Figure 2: Images generated from the same random seed without (left) and with (right) the *synchronization* of joint conditional diffusions. While the left image displays a city in both daytime and dusk, applying synchronization leads to a more coherent result. Prompt is *"a beautiful city on a sunny day"*.

where

$$\phi_\theta(\mathbf{x}_t, t, c^{(i)}) := \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t, c^{(i)})) \tag{2}$$

denotes the *predicted clean latent* at timestep $t$ of DDIM [5].

While such joint conditional diffusion can generate seamless images, it still fails to achieve the global coherence between distant windows, often leading to unrealistic images (see Fig. 2 left image). Following SyncDiffusion [2], ControlSync addresses this coherence issue by *synchronizing* the joint conditional diffusion processes through gradient descent, based on the perceptual loss [9] between the *foreseen denoised* image from different windows. Let $\mathcal{L}(\cdot, \cdot)$ denote the LPIPS [9] value between two images, then we define the **synchronization loss** $\mathcal{L}_{SYNC}$ between the $i$-th and $j$-th window as follows:

$$\mathcal{L}_{SYNC}\left(\mathbf{x}_t^{(i)}, \mathbf{x}_t^{(j)}, c^{(i)}, c^{(j)}, t\right) := \mathcal{L}\left(\mathcal{D}(\phi_\theta(\mathbf{x}_t^{(i)}, t, c^{(i)})), \mathcal{D}(\phi_\theta(\mathbf{x}_t^{(j)}, t, c^{(j)}))\right), \tag{3}$$

where $\mathcal{D}$ is the pretrained latent decoder. After obtaining $\mathcal{L}_{SYNC}$, ControlSync performs a gradient descent through $\mathbf{x}_t^{(i)}$ to update the current latent:

$$\hat{\mathbf{x}}_t^{(i)} = \mathbf{x}_t^{(i)} - w\nabla_{\mathbf{x}_t^{(i)}}\mathcal{L}_{SYNC}\left(\mathbf{x}_t^{(i)}, \mathbf{x}_t^{(0)}, c^{(i)}, c^{(0)}, t\right) \tag{4}$$

We set $j = 0$ and fixed it as the anchor window. The final *one-step denoising* pipeline is shown in Alg. 1 in Supplementary. Plugging this synchronization step into the joint conditional diffusion results in a globally coherent image while following the input condition $c$ (see Fig. 2 right image).

## 3 Experiments & Results

ControlSync is implemented in PyTorch based on the Hugging Face implementation of ControlNet [8]. We set the gradient descent weight $w$ in Eq. 4 as 20 and set weight decay as 0.95. We used DDIM [5] sampling of 50 steps and applied the gradient descent for the first 10 sampling steps. As shown in Fig. 1, ControlSync can take an arbitrary sized condition input and generate a *globally coherent* image, while properly reflecting the semantic properties of the given condition. The effect of our synchronization step is clearly shown in Fig. 2, and more generated results are provided in the Supplementary.

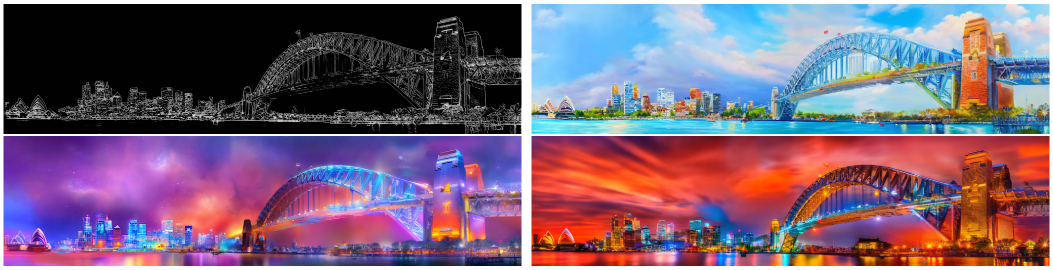


Figure 3: Images generated by ControlSync from a line art condition input (top left) with different text prompts.

## 4 Conclusion

In this paper, we propose ControlSync, an extension of SyncDiffusion [2] for extending the conditional image generation task [8] to arbitrary resolutions by leveraging the idea of *synchronizing* joint diffusion processes. We plan seek further applications in image-related tasks such as image-to-image translation and image editing.

**Potential Negative Social Impacts** As image generation models can potentially generate deepfakes, copyrighted materials, and biased contents, further research is required to advance detection of manipulated content and establish a solid barrier for protecting intellectual property rights.

## References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

[2] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *arXiv preprint arXiv:2306.05178*, 2023.

[3] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn 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.

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[6] Jinheng Xie, Kai Ye, Yudong Li, Yuexiang Li, Kevin Qinghong Lin, Yefeng Zheng, Linlin Shen, and Mike Zheng Shou. Visorgpt: Learning visual prior via generative pre-training. *arXiv preprint arXiv:2305.13777*, 2023.

[7] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023.

[8] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[9] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver 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.

[10] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023.