Infinite-Canvas: Higher-Resolution Video Outpainting with Extensive Content Generation

Qihua Chen[1,3*] , **Yue Ma**[1,2*], **Hongfa Wang**[1,4*], **Junkun Yuan**[1*†],
Wenzhe Zhao[1], **Qi Tian**[1], **Hongmei Wang**[1], **Shaobo Min**[1], **Qifeng Chen**[2†], **Wei Liu**[1]
[1]Tencent, Shenzhen, China
[2]The Hong Kong University of Science and Technology, HongKong
[3]University of Science and Technology of China, Hefei 230027, China
[4]Tsinghua University, Beijing, China
cqh@mail.ustc.edu.cn, ymacn@cse.ust.hk, {hongfawang, junkunyuan, carsonzhao, noaltian, mayhmwang, bobmin}@tencent.com, cqf@ust.hk, wliu@ee.columbia.edu

Abstract

This paper explores higher-resolution video outpainting with extensive content generation. We point out common issues faced by existing methods when attempting to largely outpaint videos: the generation of low-quality content and limitations imposed by GPU memory. To address these challenges, we propose a diffusion-based method called *Infinite-Canvas*. It builds upon two core designs. First, instead of employing the common practice of "single-shot" outpainting, we distribute the task across spatial windows and seamlessly merge them. It allows us to outpaint videos of any size and resolution without being constrained by GPU memory. Second, the source video and its relative positional relation are injected into the generation process of each window. It makes the generated spatial layout within each window harmonize with the source video. Coupling with these two designs enables us to generate higher-resolution outpainting videos with rich content while keeping spatial and temporal consistency. Infinite-Canvas excels in large-scale video outpainting, e.g., from 512×512 to 1152×2048 ($9\times$), while producing high-quality and aesthetically pleasing results. It achieves the best quantitative results across various resolution and scale setups.

Code — https://github.com/mayuelala/FollowYourCanvas

1 Introduction

Video outpainting aims to expand spatial contents of a video beyond its original boundaries to fill a designated canvas region. This task has numerous applications, such as enhancing viewing experience by adjusting aspect ratio of videos to match different users' smartphones (Wang et al. 2024a).

Recently, diffusion models (Ho, Jain, and Abbeel 2020) have emerged as the dominant approach for visual generation, demonstrating exceptional visual synthesis ability by producing appealing results (Rombach et al. 2022). Meanwhile, several diffusion-based video outpainting methods, such as M3DDM (Fan et al. 2023) and MOTIA (Wang et al. 2024a), have been proposed. They utilize the source video as a condition and generate the canvas region through step-by-step denoising, showing great performance. However, their results are limited in terms of *resolution*, such as 256×256 (Fan et al. 2023) and 512×1024 (Wang et al. 2024a), or *content expansion ratio*, for example, from 256×85 to 256×256 ($3\times$) (Fan et al. 2023) and from 512×512 to 512×1024 ($2\times$) (Wang et al. 2024a). This raises an intriguing question: *"Is it possible to outpaint a video to higher resolution with a higher content expansion ratio?"*

This question drives us to evaluate the capability of existing methods in tackling this difficult task. However, we find that they fall short due to limitations in GPU memory. To further explore their potential, we reduce the resolution of the source video through resizing and then resizing it back after outpainting (see details in Section 4). The results are depicted in Fig 1. We observe that both M3DDM (Fan et al. 2023) and MOTIA (Wang et al. 2024a) produce low-quality results, e.g., blurry content and temporal inconsistencies. This motivates us to delve deeper into understanding the reasons behind this. We speculate that there are two possible factors contributing to this: (i) the reduced resolution after resizing negatively affects the performance, and (ii) the content expansion ratio is too high to achieve satisfactory results. We conduct experiments with respect to the variations of these factors, see Fig 3. The results demonstrate that both low resolution and a high content expansion ratio significantly reduce generation quality. In other words, achieving high-quality results requires performing outpainting in the *original/high resolution* with a *low content expansion ratio*. Based on the analysis above, we propose a diffusion-based method called Infinite-Canvas for higher-resolution video outpainting with extensive content generation. We identify that the GPU memory limitations arises from the "single-shot" outpainting practice (Fan et al. 2023; Wang et al. 2024a): directly taking the entire video as the input. In contrast, our Infinite-Canvas is designed to distribute the task across spatial windows. It kills two birds with one stone. First, it enables us to outpaint any videos to higher resolution with a high content expansion ratio, without be-

[*] Equal Contribution.
[†] Corresponding author.

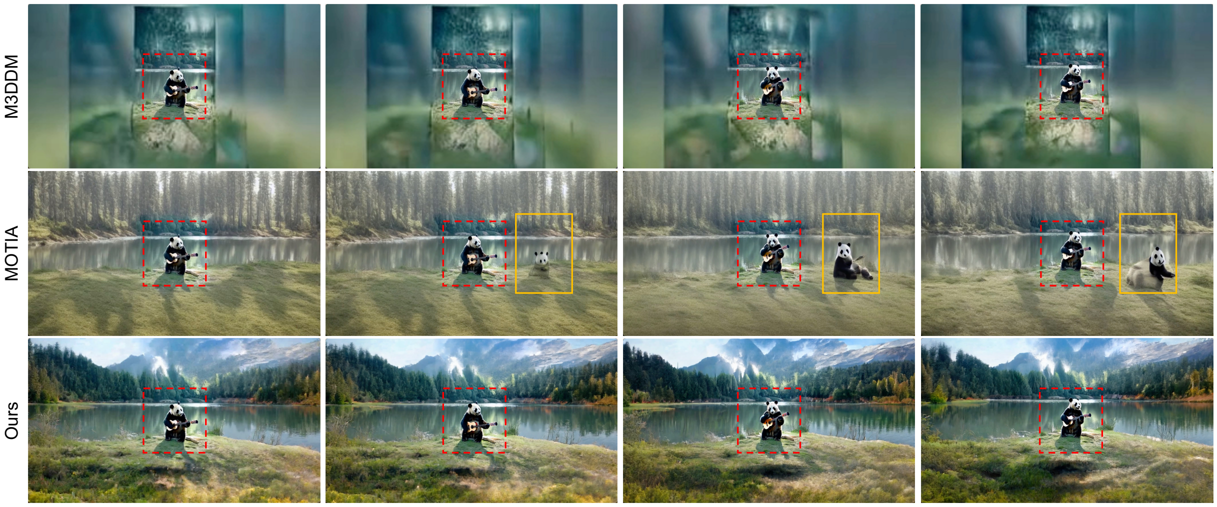


Figure 1: Results of higher-resolution outpainting with a high content expansion ratio. The source video (the red box) is outpainted from 512 × 512 to 1152 × 2048 (9×). Existing methods often suffer from blurry content and temporal inconsistencies (yellow boxes). In comparison, our Infinite-Canvas method generates well-structured scenes with aesthetically pleasing results.

ing constrained by GPU memory. Second, it simplifies the challenging task by breaking it down into smaller and easier sub-tasks: outpainting each window in the original/high resolution with a low content expansion ratio. Specifically, during the training phase, we randomly sample an anchor window and a target window from the source video, mimicking the "source video" and "outpainting region" for inference respectively. It helps model learn how to flexibly outpaint with different relative positions and overlaps between the source video and outpainting region. During the inference phase, we outpaint a video by denoising windows that covering the entire video. To accelerate the generation process, we perform window outpainting in parallel on multiple GPUs. After each step of denoising, we seamlessly merge the windows using Gaussian weights (Bar-Tal et al. 2023) to ensure a smooth transition between them. Due to the fact that videos of any resolution can be covered by a certain number of fixed size windows, while each window is limited within the GPU memory range, our method could be applied to situations where the canvas size is very large ("infinite").

Despite the advantages offered by the spatial window strategy, we observe conflicts between the layout generated within each window and the overall layout of the source video (see Fig 4). This issue arises due to the fact that the model input for each window is only a portion of the source video. Consequently, while the outpainting results within each window are reasonable, they fail to align with the overall layout, particularly when the overlap is low. To address this challenge, our Infinite-Canvas method incorporates the source video and its relative positional relation into the generation process of each window. This ensures that the generated layout harmonizes with the source video. Specifically, we introduce a **L**ayout **E**ncoder (LE) module, which takes the source video as input and provides overall layout information to the model through cross-attention. Meanwhile, we incorporate a **R**elative **R**egion **E**mbedding (RRE) into the output of the LE module, which offers information about the relative positional relation. The RRE is calculated based on the offset of the source video to the target window (outpainting region), as well as the size of them. The LE and RRE guide each window to generate outpainting results that conform to the global layout based on its relative position, effectively improving the spatial-temporal consistency.

Coupling with the strategies of spatial window and layout alignment, our Infinite-Canvas excels in large-scale video outpainting. For example, it outpaints videos from 512 × 512 to 1152 × 2048 (9×), while delivering high-quality and aesthetically pleasing results (Fig 2). When compared to existing methods, Infinite-Canvas produces better results by maintaining spatial-temporal consistency (Fig 1). Infinite-Canvas also achieves the best quantitative results across various resolution and scale setups. For example, it improves FVD from 928.6 to 735.3 (+193.3) when outpainting from 512 × 512 to 2048 × 1152 (9×) on the DAVIS 2017 dataset.

Our main contributions are summarized as follow:

- We emphasize the importance of high resolution and a low content expansion ratio for video outpainting.
- Based on the observation, we distribute the task across spatial windows, which not only overcomes GPU memory limitations but also enhances outpainting quality.
- To ensure alignment between the generated layout and the source video, we incorporate the source video and its relative positional relation into the generation process.
- Our Infinite-Canvas demonstrates great outpainting capabilities through both qualitative and quantitative results.

2 Related Work

Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) are a class of generative models that progressively convert noise into structured data through

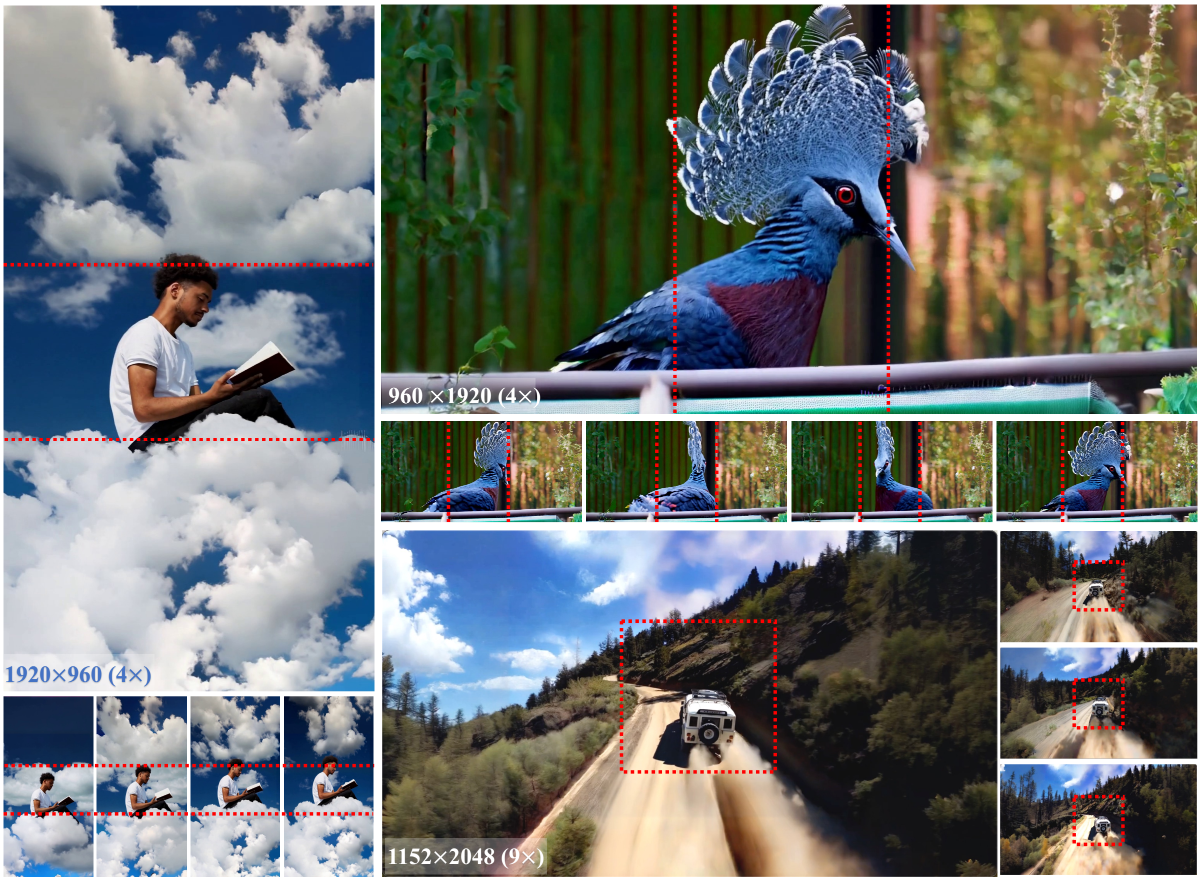


Figure 2: Results of Infinite-Canvas. The videos (from OpenAI's Sora demo cases) within the red dotted boxes are largely outpainted from 4× to 9×. Given a video of any size and resolution, Infinite-Canvas can generate outpainting results in higher resolution with extensive content, while maintaining consistency of spatial layout, temporal changes, and overall aesthetics.

a learned denoising process. It has garnered significant attention in visual generation (Ramesh et al. 2022; Zhang, Rao, and Agrawala 2023; Podell et al. 2023; Feng et al. 2024; Li et al. 2024; Kong et al. 2024; Wang et al. 2024c; Sun et al. 2022, 2024). By applying diffusion models in the latent space, LDM (Rombach et al. 2022) has demonstrated the ability to generate high-quality images by utilizing limited computational resources. Meanwhile, many works (Guo et al. 2024; Ma et al. 2023a; Zhu et al. 2024b; Blattmann et al. 2023; Ma et al. 2022; Ho et al. 2022) generate impressive videos by inserting temporal layers into the model structure. This has promoted the rapid development of video generation in editing (Ceylan, Huang, and Mitra 2023; Liu et al. 2024; Qi et al. 2023), controllable generation (Ma et al. 2024a; Xue et al. 2024; Ma et al. 2023b, 2024b), outpainting (Fan et al. 2023; Wang et al. 2024a), etc.

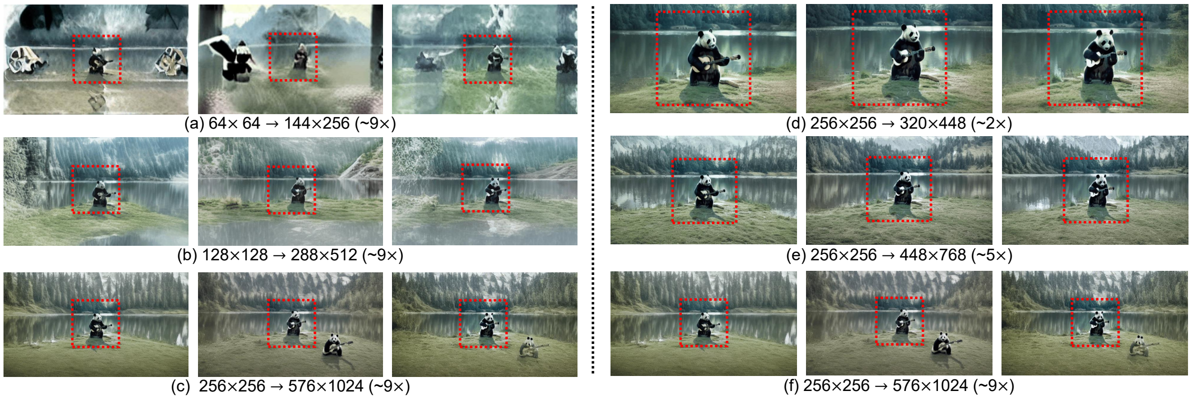
Video outpainting seeks to extend the spatial contents of a video beyond its initial boundaries, allowing it to fill a specific canvas region. Although image outpainting (Zhang et al. 2024; Yu et al. 2024; Cheng et al. 2022; Wang et al. 2024b; Zhu et al. 2024a) has been extensively studied, video outpainting (Dehan et al. 2022) still needs to be fully researched. Recently, some diffusion-based approaches have been introduced. M3DDM (Fan et al. 2023) presents global frame-guided training with a coarse-to-fine inference pipeline to tackle the artifact accumulation issue. Meanwhile, MOTIA (Wang et al. 2024a) proposes a test sample-specific fine-tuning strategy to learn the patterns of each sample. Despite their great results, they are limited in terms of resolution such as 256×256 and 512×1024, or content expansion ratio such as 2× and 3×. We makes the first attempt to study video outpainting with high resolution, *e.g.*, 1152×2048, and a high content expansion ratio, *e.g.*, 9×.

3 Method

We present Infinite-Canvas, a diffusion-based method, which enables higher-resolution video outpainting with extensive content generation. Our approach is built upon two key designs. First, we employ spatial windows to divide the outpainting task into smaller and easier sub-tasks. Second, we introduce a layout encoder module as well as a relative region embedding to align the generated spatial layout.

3.1 Outpainting by Spatial Windows

To address the GPU memory limitations, we distribute the outpainting task across spatial windows. It allows us to out-



(a) 64× 64 → 144×256 (~9×)

(b) 128×128 → 288×512 (~9×)

(c) 256×256 → 576×1024 (~9×)

(d) 256×256 → 320×448 (~2×)

(e) 256×256 → 448×768 (~5×)

(f) 256×256 → 576×1024 (~9×)

Figure 3: Results of MOTIA with different resolution (a-c) and content expansion ratio (d-f) setups. Increasing source video resolution improves the generation quality, while reducing content expansion ratio improves spatial-temporal consistency.

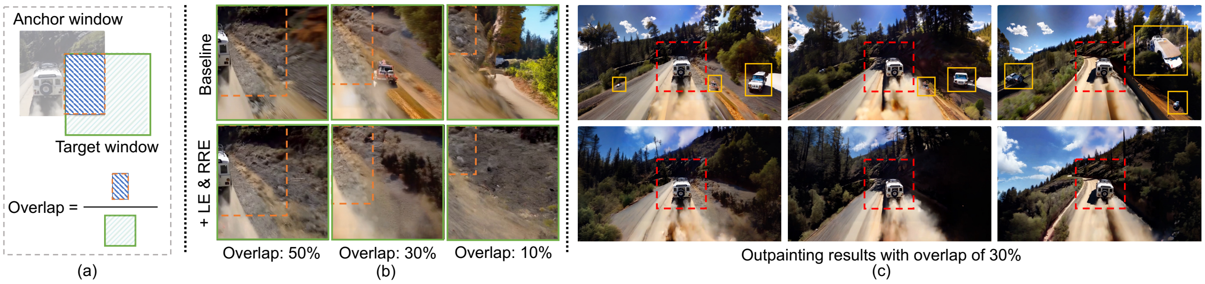


Figure 4: Ablation of layout encoder (LE) & relative region embedding (RRE). Under different overlap (a), results within target windows (b) and the final results (c) are presented. The orange dashed line represents the model input for target windows. While the results appear reasonable within windows, they fail to align with the overall layout (see yellow boxes). By incorporating RRE and LE, the model unifies window layout with that of the anchor window, improving spatial-temporal consistency.

paint any videos to higher resolution with a high content expansion ratio without being constrained by GPU memory. Moreover, it simplifies the task by breaking it down into smaller and easier sub-tasks: outpainting each window in its original/high resolution with a low content expansion ratio.

Training phase. Fig 5 illustrates the training phase of Infinite-Canvas. Given each training video sample, we randomly crop an anchor window and a target window. They serve as the "source video" and the "region to perform outpainting" respectively, mimicking the source video and the outpainting windows during inference, respectively. The conventional training practice of the latent diffusion model adds noise to the latent representation of the data (the target window) to build the model input and makes the model predict the noise. Here, we concatenate it with conditions: the latent representation of a masked target window and the binary mask. They offer information of the original video and its position. Since the channel of the mask and the latent representations output by the VAE encoder are 1 and 4 respectively, the final model input has 9 channels. We modify the first convolution layer of the denoising UNet to adjust to the channel changes, similar to previous work (Fan et al. 2023). However, instead of employing a fixed region for outpainting (Fan et al. 2023; Wang et al. 2024a), we use a random sample of the anchor window and the target window. It helps the model learn to flexibly outpaint with different relative positions and overlaps between the source video and the outpainting region, enabling the sliding window-based inference phase described next. Note that the size of the anchor window, the target window, and their overlap are all variables. See details in experiments.

Inference phase. Fig 6 illustrates the inference phase of Infinite-Canvas. Given a source video to be outpainted, our Infinite-Canvas first determines the number (denoted as N) of spatial windows and their positions, which should cover the source video and fill the target region to be outpainted (find more details in experiments). During each denoising step t, Infinite-Canvas performs outpainting within each window k on noisy data $\mathbf{x}_t^k$, where $k \in \{1, ..., N\}$. Here, the source video and the window correspond to the anchor window and the target window of the training phase respectively. The denoised outputs in the N windows, i.e., $\{\mathbf{x}_{t-1}^k\}_{k=1}^N$, are then merged via Gaussion weights (Bar-Tal et al. 2023) to get a smooth outcome $\mathbf{x}_{t-1}$. The process is repeated until the final outpainting result $\mathbf{x}_0$ is obtained. Importantly, the inference process of each window is independent of the others, allowing us to perform outpainting within each window in parallel on separate GPUs, thereby acceler-

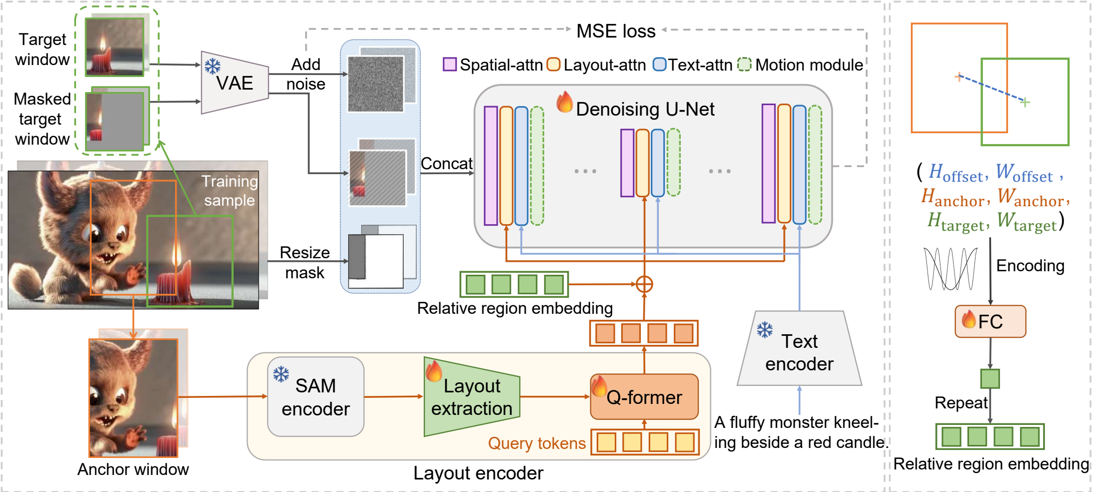


Figure 5: The training phase of Infinite-Canvas. An anchor window and a target window are randomly sampled, mimicking the "source video" and "region to perform outpaint" for inference respectively. The anchor window is injected into the model through a layout encoder, as well as a relative region embedding calculated by the positional relation between the anchor window and the target window, helping the model align the generated layout of the target window with the anchor window.

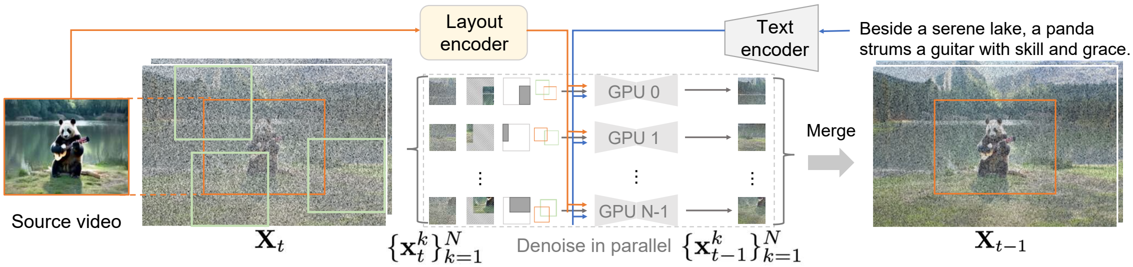

Layout encoder
Text encoder
Beside a serene lake, a panda strums a guitar with skill and grace.
Source video
$\mathbf{X}_t$
$\{\mathbf{x}_t^k\}_{k=1}^N$
GPU 0
GPU 1
GPU N-1
Denoise in parallel
$\{\mathbf{x}_{t-1}^k\}_{k=1}^N$
Merge
$\mathbf{X}_{t-1}$

Figure 6: The inference phase of Infinite-Canvas. The given source video is covered by N spatial windows. During each denoising step t, outpainting is performed within each window in parallel on separate GPUs to accelerate inference. The windows are then merged through Gaussian weights to get the outcome at step $t-1$. Note that these windows may cover layer upon layer, allowing Infinite-Canvas to outpaint any videos to a higher resolution without being limited by the GPU memory.

ating the inference. We analyze its efficiency in experiments.

3.2 Layout Alignment

Despite the advantages offered by the spatial window strategy, we observe conflicts between the layout generated within each window and the overall layout of the source video, as shown in Fig 4. The outpainting results within each window of the "baseline", which only applies the spatial window strategy, are reasonable. However, they do not align with the global layout because each window is provided with a view of only a part of the source video. To enable spatial and temporal consistency, we introduce a layout encoder and relative region embedding. They deliver the layout information of the source video and its relative position relation to each window respectively, effectively helping the model generate more stable and consistent outpainting videos (see the results of "+LE & RRE" method in Fig 4).

Layout Encoder (LE). Similar to the text encoder that injects the text prompts into the model, we introduce LE to incorporate layout information from the source video, see Fig 5. Specifically, LE consists of a SAM encoder (Kirillov et al. 2023), a layout extraction module, and a Q-former (Li et al. 2023). Instead of employing the CLIP visual encoder (Radford et al. 2021) like many previous works (Ye et al. 2023; Xue et al. 2024), we find SAM encoder (ViT-B/16 structure) is more effective to extract visual features by providing finer visual details (see comparisons in experiments). Then, the layout features are extracted by the layout extraction module, including a pseudo-3D convolution layer, two temporal attention layers, and a temporal pooling layer. Inspired by Li et al. 2023, we employ a Q-former (Querying Transformer) to extract and refine visual representations of the layout information by learnable query tokens. We train the layout extraction module and the Q-former while fixing the SAM encoder. The relative region embedding is added to the output of the LE to provide a positional relation between the anchor window and the target window, introduced next.

Resolution	Method	FVD↓	LPIPS↓	AQ↑	IQ↑	PSNR↑	SSIM↑
1280×720 (720P, $\sim 3.5\times$)	MOTIA (Wang et al. 2024a)	473.7	0.418	0.494	0.634	**15.38**	0.582
	Dehan (Dehan et al. 2022)	736.0	0.604	0.435	0.542	13.95	0.605
	M3DDM (Fan et al. 2023)	631.3	0.524	0.446	0.556	15.28	0.605
	Infinite-Canvas (Ours)	**440.0**	**0.390**	**0.509**	**0.658**	**15.38**	**0.606**
1440×810 (1.5K, $\sim 4.5\times$)	MOTIA (Wang et al. 2024a)	575.9	0.457	0.484	0.648	14.52	0.539
	Dehan (Dehan et al. 2022)	857.2	0.650	0.415	0.543	13.38	0.553
	M3DDM (Fan et al. 2023)	767.4	0.579	0.447	0.519	14.43	0.542
	Infinite-Canvas (Ours)	**486.1**	**0.440**	**0.505**	**0.650**	**14.90**	**0.559**
2048×1152 (2K, $9\times$)	MOTIA (Wang et al. 2024a)	928.6	0.587	0.419	0.629	12.45	0.524
	Dehan (Dehan et al. 2022)	1302.1	0.707	0.394	0.607	11.40	0.501
	M3DDM (Fan et al. 2023)	1181.4	0.691	0.411	0.473	12.43	0.530
	Infinite-Canvas (Ours)	**735.3**	**0.573**	**0.472**	**0.657**	**12.72**	**0.535**

Table 1: Quantitative comparisons for higher resolution video outpainting with high content expansion ratios. The resolution of the source video is 512×512. MOTIA is noted by gray because it is based on test sample-specific fine-tuning.

method	PSNR↑	SSIM↑	LPIPS↓	FVD↓
MOTIA (Wang et al. 2024a)	20.36	**0.758**	0.159	286.3
Dehan (He et al. 2022)	17.96	0.627	0.233	363.1
SDM (He et al. 2022)	20.02	0.708	0.216	334.6
M3DDM (Fan et al. 2023)	20.26	0.708	0.203	300.0
Infinite-Canvas (Ours)	**20.80**	0.726	**0.160**	**242.8**

Table 2: Quantitative comparisons for low resolution video outpainting (to 256×256). MOTIA is noted by gray because it is based on test sample-specific fine-tuning.

Relative Region Embedding (RRE). RRE provides the positional relation between the anchor window and the target window (see Fig 5). We denote the height, width, and center point coordinates of the anchor window as H_{anchor}, W_{anchor}, and $(X_{\text{anchor}}, Y_{\text{anchor}})$ respectively. The target window is defined in the same way. RRE employs sinusoidal position encoding (Zhang et al. 2024) to embed the size and relative position relation between the anchor and target windows, i.e., $\{H_{\text{anchor}}, W_{\text{anchor}}, H_{\text{target}}, W_{\text{target}}, H_{\text{offset}}, W_{\text{offset}}\}$, where $H_{\text{offset}} = Y_{\text{target}} - Y_{\text{anchor}}$, $W_{\text{offset}} = X_{\text{target}} - X_{\text{anchor}}$. The embeddings are then fed to a fully-connected (FC) layer. The output of the FC layer is repeated to match the output of the LE. We incorporate the LE and RRE using a cross-attention layer inserted in each spatial-attention block of the model.

4 Experiments

4.1 Setup

Dataset. Fan et al. 2023 use a private dataset with $\sim$5M video samples. Here, we employ a random subset ($\sim$1M video samples) of the public Panda-70M dataset (Chen et al. 2024) for training, improving reproducibility of our work.

Implementation details. Our implementation and model initialization are based on the popular video generation framework of AnimateDiff-V2 (Guo et al. 2024). Due to the limitation of paper length, *we leave more details about the training recipe, the design of the anchor and target windows, and the inference pipeline in the appendix and code.*

Evaluation metrics. We first employ metrics of PSNR, SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018), and FVD (Unterthiner et al. 2018) by following Wang et al. 2024a. To evaluate high-resolution video generation, we further utilize aesthetic quality (AQ) and imaging quality (IQ) (Huang et al. 2024), assessing the layout/color harmony and visual distortion (e.g., noise and blur) respectively.

Baselines. We compare our Infinite-Canvas with the following baseline methods. (1) Dehan et al. 2022 use the approach of flow estimation and background prediction. (2) M3DDM (Fan et al. 2023) employs global-frame features to achieve global and long-range information transfer. 3) MOTIA (Wang et al. 2024a) trains a LoRA (Hu et al. 2021) to learn patterns of test samples. We reproduce these baseline methods using their official codes for high-resolution video outpainting and directly cite their results in low-resolution.

4.2 Comparisons to Baseline Methods

Quantitative results. We compare methods in both high and low-resolution settings. (1) *High-resolution with large content expansion ratios.* Table 1 shows the results. Our Infinite-Canvas consistently achieves the best performance for all metrics and outpainting settings. Meanwhile, as the resolution and content expansion ratio increase, the performance improvement of many metrics becomes more significant. For example, Infinite-Canvas improves FVD from 473.7 to 440.0 (+33.7) in 720P ($\sim$3.5×), improves from 575.9 to 486.1 (+89.8) in 1.5K, and improves from 928.6 to 735.3 (+193.3) in 2K. Our Infinite-Canvas effectively improves performance in the challenging task of high-resolution outpainting with high content expansion ratios. (2) *Conventional settings in low-resolution.* Following Fan et al. 2023 and Wang et al. 2024a, we also compare results in low-resolution, which outpaint videos to 256×256 in the horizontal direction using mask ratio of 0.25 ($\sim 1.3\times$) and 0.66 ($\sim 3\times$) and calculate the average performance. Table 2 shows the results. Our Infinite-Canvas still achieves excellent performance under this conventional setting. Note that

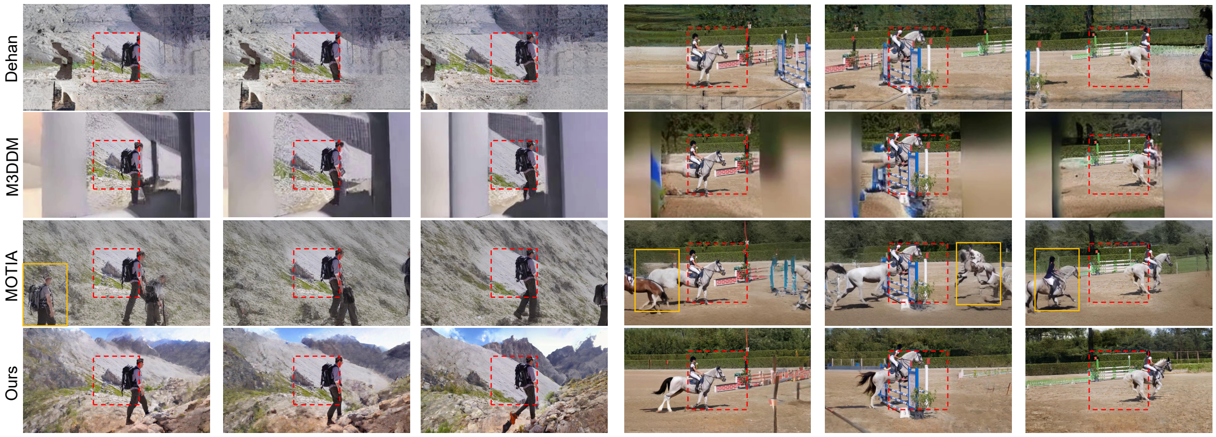


Figure 7: Qualitative results. The source video (the red dotted box) is outpainted from 512 × 512 to 2048 × 1152 (left) or 1440 × 810 (right). Baseline methods suffer from blurry content, and spatial and temporal inconsistencies (yellow boxes).

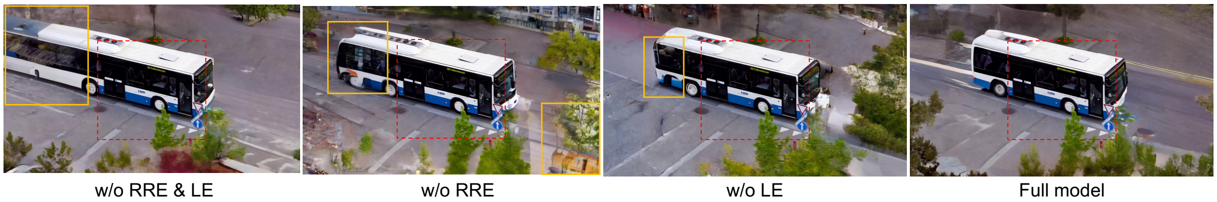


Figure 8: Visual results of ablation study. Layout encoder (LE) and relative region embedding (RRE) effectively guide the generation by providing information of the source video and its positional relation to the outpainting window respectively.

MOTIA (Wang et al. 2024a) fine-tunes the model for each test sample which may not be efficient, while our Infinite-Canvas method performs zero-shot inference after training.

Qualitative results. In Fig. 7, we showcase the qualitative results. It is evident that M3DDM fails to generate meaningful content in the majority of outpainting regions. On the other hand, MOTIA faces difficulties in maintaining spatial and temporal consistencies, which can be attributed to the challenging task of handling high resolution and content expansion ratios. In contrast, our Infinite-Canvas successfully generates well-structured visual content. It is because the design of spatial windows that outpaint within each window in its original/high resolution with a low content expansion ratio. Moreover, the layout alignment plays a crucial role in guiding the overall layout of the outpainting results.

4.3 Ablation Study

We conduct the ablation study by outpainting the source video from 512×512 to 1440×810, as shown in Table 3. We find relative region embedding (RRE), layout encoder (LE), and layout extraction module are all important to achieve the best results. Compared to the popular CLIP encoder, we observe that the SAM encoder helps the model to further improve outpainting results. Visual results are shown in Fig 8.

Method	PSNR↑	SSIM↑	LPIPS↓	FVD ↓
w/o LE & RRE	13.44	0.527	0.464	774.1
w/o LE	14.02	0.542	0.450	512.2
w/o RRE	13.63	0.532	0.458	670.3
w/o layout extraction	13.77	0.535	0.456	550.2
w/ CLIP image encoder	14.56	0.553	0.441	506.8
Infinite-Canvas (ours)	**14.90**	**0.559**	**0.440**	**486.1**

Table 3: Ablation study.

Resolution	1 GPU	2 GPUs	4 GPUs	8 GPUs
1280 × 720	25.2	14.8	7.8	4.3
1440 × 810	58.3	33.5	18.2	11.5
2048 × 1152	85.8	51.9	28.9	16.2

Table 4: Run time (minutes). Parallel inference for outpainting a video of 512 × 512 resolution with 64 frames.

5 Conclusion

Largely expanding an image/video is the core of the outpainting task. In this study, we take the first step towards exploring higher-resolution video outpainting with high content expansion ratios. We achieve this by introducing the spatial window strategy combined with the design of layout alignment. Our method allows for large-scale video outpainting, e.g., from 512 × 512 to 1152 × 2048 (9×).

Limitations. Infinite-Canvas may require long inference time by performing the spatial window strategy. See Table2. We recommend users to use multiple GPUs in parallel to accelerate inference and reduce inference time consumption.

Acknowledgments

This project was supported by the National Key R&D Program of China under grant number 2022ZD0161501.

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *arXiv preprint arXiv:2302.08113*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23206–23217.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.-w.; Jeon, B. E.; Fang, Y.; Lee, H.-Y.; Ren, J.; Yang, M.-H.; and Tulyakov, S. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. *arXiv preprint arXiv:2402.19479*.

Cheng, Y.-C.; Lin, C. H.; Lee, H.-Y.; Ren, J.; Tulyakov, S.; and Yang, M.-H. 2022. Inout: Diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11431–11440.

Dehan, L.; Van Ranst, W.; Vandewalle, P.; and Goedemé, T. 2022. Complete and temporally consistent video outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 687–695.

Fan, F.; Guo, C.; Gong, L.; Wang, B.; Ge, T.; Jiang, Y.; Luo, C.; and Zhan, J. 2023. Hierarchical masked 3d diffusion model for video outpainting. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7890–7900.

Feng, K.; Ma, Y.; Wang, B.; Qi, C.; Chen, H.; Chen, Q.; and Wang, Z. 2024. Dit4edit: Diffusion transformer for image editing. *arXiv preprint arXiv:2411.03286*.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *International Conference on Learning Representations*.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*.

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

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

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; et al. 2024. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. *arXiv preprint arXiv:2405.08748*.

Liu, S.; Zhang, Y.; Li, W.; Lin, Z.; and Jia, J. 2024. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8599–8608.

Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023a. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023b. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. *arXiv preprint arXiv:2304.01186*.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Qi, C.; Cai, C.; Li, X.; Li, Z.; Shum, H.-Y.; Liu, W.; et al. 2024a. Follow-your-click: Open-domain regional image animation via short prompts. *arXiv preprint arXiv:2403.08268*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024b. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*.

Ma, Y.; Wang, Y.; Wu, Y.; Lyu, Z.; Chen, S.; Li, X.; and Qiao, Y. 2022. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4132–4141.

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. arXiv:2307.01952.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15932–15942.

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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.

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

Sun, S.; Chen, Y.; Zhu, Y.; Guo, G.; and Li, G. 2022. Skflow: Learning optical flow with super kernels. *Advances in Neural Information Processing Systems*, 35: 11313–11326.

Sun, S.; Liu, J.; Li, T. H.; Li, H.; Liu, G.; and Gao, W. 2024. StreamFlow: Streamlined Multi-Frame Optical Flow Estimation for Video Sequences. *Advances in Neural Information Processing Systems*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Wang, F.-Y.; Wu, X.; Huang, Z.; Shi, X.; Shen, D.; Song, G.; Liu, Y.; and Li, H. 2024a. Be-Your-Outpainter: Mastering Video Outpainting through Input-Specific Adaptation. *arXiv preprint arXiv:2403.13745*.

Wang, J.; Ma, Y.; Guo, J.; Xiao, Y.; Huang, G.; and Li, X. 2024b. COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing. *arXiv preprint arXiv:2406.08850*.

Wang, J.; Pu, J.; Qi, Z.; Guo, J.; Ma, Y.; Huang, N.; Chen, Y.; Li, X.; and Shan, Y. 2024c. Taming Rectified Flow for Inversion and Editing. *arXiv preprint arXiv:2411.04746*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xue, J.; Wang, H.; Tian, Q.; Ma, Y.; Wang, A.; Zhao, Z.; Min, S.; Zhao, W.; Zhang, K.; Shum, H.-Y.; et al. 2024. Follow-Your-Pose v2: Multiple-Condition Guided Character Image Animation for Stable Pose Control. *arXiv preprint arXiv:2406.03035*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, H.; Li, R.; Xie, S.; and Qiu, J. 2024. Shadow-Enlightened Image Outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7850–7860.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, S.; Huang, J.; Zhou, Q.; Wang, Z.; Wang, F.; Luo, J.; and Yan, J. 2024. Continuous-Multiple Image Outpainting in One-Step via Positional Query and A Diffusion-based Approach. *arXiv preprint arXiv:2401.15652*.

Zhu, C.; Li, K.; Ma, Y.; He, C.; and Xiu, L. 2024a. MultiBooth: Towards Generating All Your Concepts in an Image from Text. *arXiv preprint arXiv:2404.14239*.

Zhu, C.; Li, K.; Ma, Y.; Tang, L.; Fang, C.; Chen, C.; Chen, Q.; and Li, X. 2024b. Instantswap: Fast customized concept swapping across sharp shape differences. *arXiv preprint arXiv:2412.01197*.