

# DiT4Edit: Diffusion Transformer for Image Editing

Kunyu Feng<sup>1\*</sup>, Yue Ma<sup>2\*</sup>, Bingyuan Wang<sup>3\*</sup>, Chenyang Qi<sup>2</sup>, Haozhe Chen<sup>1</sup>,  
Qifeng Chen<sup>2†</sup>, Zeyu Wang<sup>3,2†</sup>

<sup>1</sup>Peking University, China

<sup>2</sup>The Hong Kong University of Science and Technology, China

<sup>3</sup>The Hong Kong University of Science and Technology (Guangzhou), China

{feng\_ky21, chenhaozhe}@stu.pku.edu.cn, ymacn@cse.ust.hk, bwang667@connect.hkust-gz.edu.cn  
chenyangqi67@gmail.com, {cqf, zeyuwang}@ust.hk

## Abstract

Despite recent advances in UNet-based image editing, methods for shape-aware object editing in high-resolution images are still lacking. Compared to UNet, Diffusion Transformers (DiT) demonstrate superior capabilities to effectively capture the long-range dependencies among patches, leading to higher-quality image generation. In this paper, we propose DiT4Edit, the first Diffusion Transformer-based image editing framework. Specifically, DiT4Edit uses the DPM-Solver inversion algorithm to obtain the inverted latents, reducing the number of steps compared to the DDIM inversion algorithm commonly used in UNet-based frameworks. Additionally, we design unified attention control and patch merging, tailored for transformer computation streams. This integration allows our framework to generate higher-quality edited images faster. Our design leverages the advantages of DiT, enabling it to surpass UNet structures in image editing, especially in high-resolution and arbitrary-size images. Extensive experiments demonstrate the strong performance of DiT4Edit in various editing scenarios, highlighting the potential of diffusion transformers for image editing.

**Code** — <https://github.com/fkyyyy/DiT4Edit>

## Introduction

Recent advances in diffusion models have witnessed impressive progress in text-driven visual generation. The development of these text-to-image (T2I) models, e.g., Stable Diffusion (SD) (Rombach et al. 2022), DALL-E 3 (Betker et al. 2023), and PixArt (Chen et al. 2023), has led to significant impacts on numerous downstream applications (Ma et al. 2024a,b; Wang et al. 2024; Zhu et al. 2024; Wang et al. 2023), with image editing as one of the most challenging tasks. Given a synthetic or real input image, image editing algorithms aim to add, remove, or replace entire objects or object attributes according to the user’s intent.

A primary challenge in text-driven image editing is maintaining the consistency between the source and target images. Earlier approaches (Choi et al. 2021) (Kawar et al.

2023) (Zhang et al. 2023) often relied on fine-tuning diffusion models to address this issue. However, these methods typically require considerable time and computational resources, which limits their practical applicability. Recent approaches often utilize DDIM (Song, Meng, and Ermon 2020) inversion to obtain latent maps and then control the attention mechanism in diffusion models for real image editing (Mokady et al. 2023) (Cao et al. 2023). However, the consistency of the edited images depends heavily on the invertibility of the DDIM inversion process. Although some efforts have focused on optimizing this inversion (Ju et al. 2024) (Dong et al. 2023) for better results, the editing framework still relies on too many timesteps (e.g., 50 steps).

In addition, current research on image editing tasks mainly uses the UNet-based diffusion model structure (Rombach et al. 2022), making the final editing results heavily bounded by the generative capacity of UNet. Although the attention mechanism in UNet is also derived from the transformer, DiT (Peebles and Xie 2023) based on pure transformers offers a global attention calculation between patches, allowing them to capture broader and more detailed features compared to the UNet with convolution blocks, leading to higher-quality images. In addition, evidence from DiT demonstrates that transformer-based diffusion models offer better scalability and outperform UNet-based models in large-scale experiments.

To address these challenges, we explore image editing tasks using the diffusion transformer architecture and provide a valuable empirical baseline for future research. First, we aim to leverage solvers that require fewer inversion steps to reduce our inference time while maintaining the image quality of the results. Specifically, we employ an inversion algorithm based on a high-order DPM-Solver (Lu et al. 2023) to obtain better latent maps with fewer timesteps. We then implement a unified attention control scheme for text-guided image editing while preserving background details. Third, to mitigate the increased computational complexity of transformers compared to UNet, we use patch merging to accelerate computation. By integrating these key components, we introduce DiT4Edit, the first diffusion transformer-based editing framework to our knowledge. Experiments demonstrate that our framework achieves superior editing results with fewer inference steps.

\*These authors contributed equally.

†Corresponding authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

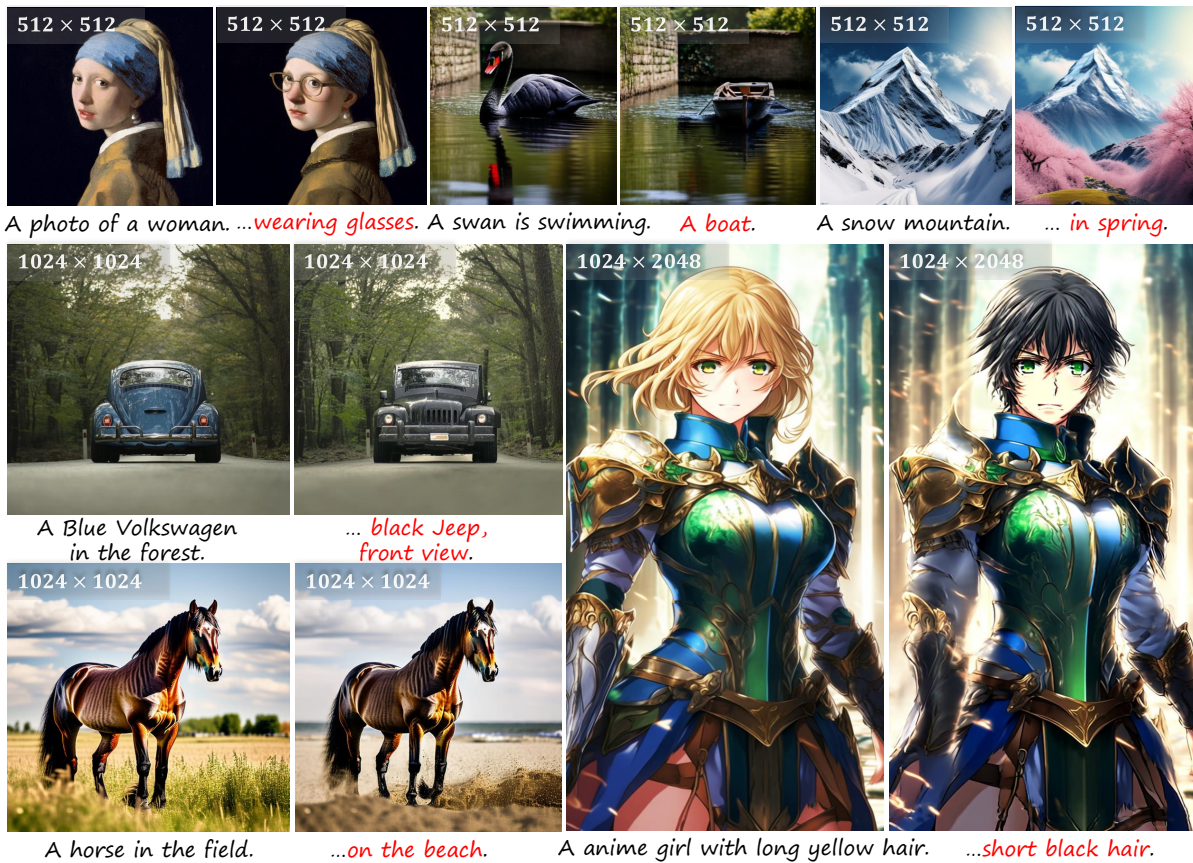


Figure 1: **Visual results of DiT4Edit.** Our method is the first DiT-based image editing framework, which is capable of handling images of various sizes: from small ( $512 \times 512$ ) to large ( $1024 \times 1024$ ), and even arbitrary dimensions (up to  $1024 \times 2048$ ).

In summary, our contributions are as follows:

- Based on the advantages of transformer-based diffusion models in image editing, we introduce DiT4Edit, the first tuning-free image editing framework using Diffusion Transformers (DiT).
- To adapt to the computing mechanism of transformer-based denoising, we first propose a unified attention control mechanism to achieve image editing. Then, we introduce the DPM-Solver inversion and patches merging strategy to reduce inference time.
- Extensive qualitative and quantitative results demonstrate the superior performance of DiT4Edit in object editing, style editing, and shape-aware editing for various image sizes, including  $512 \times 512$ ,  $1024 \times 1024$ ,  $1024 \times 2048$ .

## Related Work

### Text-to-Image Generation

Prior to the advent of diffusion models, researchers predominantly relied on a generative adversarial network (GAN) for image synthesis (Goodfellow et al. 2014). Zhang et al. (Zhang, Xie, and Yang 2018) developed a single-stream generator capable of producing high-resolution images,

while Liang et al. (Liang, Pei, and Lu 2020) enhanced the performance of text-to-image synthesis by incorporating Memory-Attended Text Encoder and Object-Aware Image Encoder. Subsequently, the Denoising Diffusion Probabilistic Models (DDPMs) introduced by Ho et al. (Ho, Jain, and Abbeel 2020) marked significant leaps forward, achieving breakthroughs in image quality, controllability, and diversity. The designs and applications of diffusion-based methods can be classified by tasks such as controllable generation, stylization, and quality improvement (Wang, Chen, and Wang 2024). Dhariwal et al. (Dhariwal and Nichol 2021) introduced a classifier guidance method to improve the generation quality of diffusion models, while Yang et al. (Yang et al. 2024) used the CLIP latents to produce realistic images closely aligned with human expectations in text-to-image generation tasks. In recent text-to-image generation tasks, ControlNet (Zhang, Rao, and Agrawala 2023) allows for the integration of user-specified conditional information into the image generation process. Meanwhile, ScaleCrafter (He et al. 2023) addresses the issue of limited perception in convolutional layers during the diffusion model generation process, enabling the production of higher resolution and higher quality images. Moreover, these T2I models have also driven the development of a series of video generation and editing applications (Chen et al. 2024; Ma et al. 2022a, 2023, 2022b,

2024c; Song et al. 2023; Wang, Zhang, and Wang 2023).

## Interactive Image Editing

Image editing encompasses scenarios like iterative generation, collaborative creation, and image inpainting. Compared to other generative models, diffusion models offer enhanced controllability during the image generation process, allowing for precise manipulation of image attributes (Choi et al. 2021; Zhang, Rao, and Agrawala 2023). These advantages enable diffusion models to achieve outstanding performance in image editing tasks. Hertz et al. (Hertz et al. 2022) introduced a framework for image editing through textual prompts, which transforms the original image into the target image by modifying, adding, and adjusting the weights of the cross-attention map. Methods like InstructPix2Pix (Brooks, Holynski, and Efros 2023) and Custom Diffusion (Kumari et al. 2023) employ user-guided approaches to achieve image editing. These techniques allow for modifications by inputting various types of guiding prompts, allowing diffusion models to swiftly adapt to new concepts. Parmar et al. (Parmar et al. 2023) employed pix2pix-zero to address the challenge of preserving the original structure while incorporating user-specified changes during image editing. Although there have been significant advancements in image editing using diffusion models, these existing attempts at image editing are still bounded by the pretrained generative power of a UNet. Compared with UNet-based diffusion models, DiT is more scalable and has more succinct architectures, while DiT’s application in image editing is still under-explored.

## Methodology

Our proposed framework aims to achieve high-quality image editing for various sizes based on a diffusion transformer. Our method is the first editing strategy based on a pre-trained text-to-image transformer-based diffusion model. With our approach, users can achieve better editing results compared to existing UNet-based methods by providing a target prompt. In this section, we first introduce the latent diffusion models and DPM inversion. Then we illustrate the superiority of transformer-based denoising in image editing tasks. Finally, we discuss the implementation details of our editing framework.

### Preliminaries: Latent Diffusion Models

The Latent Diffusion Model (LDM) (Rombach et al. 2022) proposes an image generation method with a denoising process within a latent space  $\mathcal{Z}$ . In particular, it uses a pre-trained image encoder  $\mathcal{E}$  to encode the input image  $x$  into low-resolution latents  $z = \mathcal{E}(x)$ . During training, the model optimizes a denoising UNet  $\epsilon_\theta$  by removing artificial noise, conditioned on both text prompt embedding  $y$  and current image sample  $z_t$ , which is a noisy sample of  $z_0$  at step  $t \in [0, T]$ :

$$\min_{\theta} E_{z_0, \epsilon \sim \mathcal{N}(0, I), t} \|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2, \quad (1)$$

After training, it is capable of converting random noise  $\epsilon$  to an image sample  $z$  by the learned denoising process.

**DPM-Solver Sampling.** During the inversion stage in Diffusion probabilistic models (DPMs), a clean image  $x_0$  is gradually added with Gaussian noise and turned into a noisy sample  $x_t$ :

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}), \quad (2)$$

where  $\frac{\alpha_t}{\sigma_t^2}$  is the signal-to-noise ratio (SNR), which is a strictly decreasing function of  $t$  (Lu et al. 2023). By solving the diffusion ODE, the DPM sampling can be faster than other methods:

$$\frac{d\mathbf{x}_t}{dt} = \left(f(t) + \frac{g^2(t)}{2\sigma_t^2}\right)\mathbf{x}_t - \frac{\alpha_t g^2(t)}{2\sigma_t^2} \mathbf{x}_\theta(\mathbf{x}_t, t), \quad (3)$$

where  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \tilde{\alpha}^2, \mathbf{I})$ , and  $f(t) = \frac{d \log \alpha_t}{dt}$ ,  $g(t) = \frac{d\alpha_t^2}{dt} - 2 \frac{d \log \alpha_t}{dt} \alpha_t^2$  (Lu et al. 2023). It is shown in the previous works (Lu et al. 2022) (Zhang and Chen 2022) that ODE solver using the exponential integrator exhibits faster convergence compared to traditional solvers during solving the Eq. 3. By setting the value of  $\mathbf{x}_s$ , the solution  $\mathbf{x}_t$  of Eq. 3 can be calculated by:

$$\mathbf{x}_t = \frac{\alpha_t}{\alpha_s} \mathbf{x}_s - \alpha_t \int_{\lambda_s}^{\lambda_t} e^{-\lambda} \mathbf{x}_\theta(\hat{\mathbf{x}}_\lambda, \lambda) d\lambda, \quad (4)$$

where the  $\lambda_t = \log(\alpha_t/\sigma_t)$  is a decreasing function of  $t$  with the inversion function  $t_\lambda(\cdot)$ , and recent research demonstrates that the DPM-Solver can sample the realistic images in 10–20 steps.

### Diffusion Model Architecture

**PIXART- $\alpha$ .** Compared to the UNet structure, Diffusion Transformers (DiT) (Peebles and Xie 2023) exhibits superior scaling properties, generating images of higher quality and demonstrating better performance.

PIXART- $\alpha$  is a Transformer-based text-to-image (T2I) diffusion model that consists of three main components: Cross-Attention layer, AdaLN-single, and Re-parameterization. (Chen et al. 2023) Researchers have trained this T2I diffusion model with three sophisticated designs: decomposition training strategies, efficient T2I transformer and high-informative data. Many experimental results demonstrate that PIXART- $\alpha$  performs better in image quality, artistry, and semantic control. Compared to advanced T2I SOTA models, PIXART- $\alpha$  has faster training speed, lower inference cost, and superior comprehensive performance. In this paper, we use PIXART- $\alpha$  as the baseline for our proposed image editing method.

**The reason for using a transformer as the denoising model.** Compared to the UNet structure, the transformer incorporates a global attention mechanism, allowing the model to focus on a broader range within the image. This enhanced scalability enables transformers to generate high-quality images at large sizes (e.g., greater than  $512 \times 512$ ), and even at arbitrary sizes. The editing results of our DiT-based editing framework for large-sized images are demonstrated in Figures 1 and 2. These represent editing tasks not previously addressed by UNet-based frameworks. Therefore, we adopted a transformer-based denoising model for our editing framework, leveraging the transformer’s capabilities to tackle these more complex editing challenges.

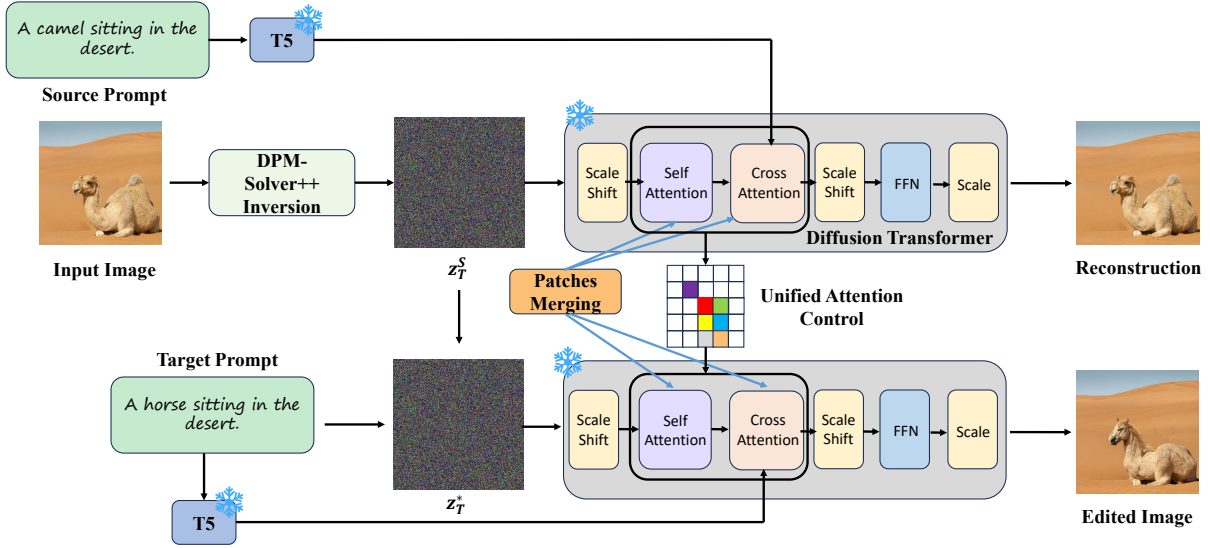


Figure 2: Overview of the DiT4Edit framework. During the image editing process, our inversion algorithm generates high-quality latent maps, and the final edited image is achieved through unified attention control.

A standing horse.

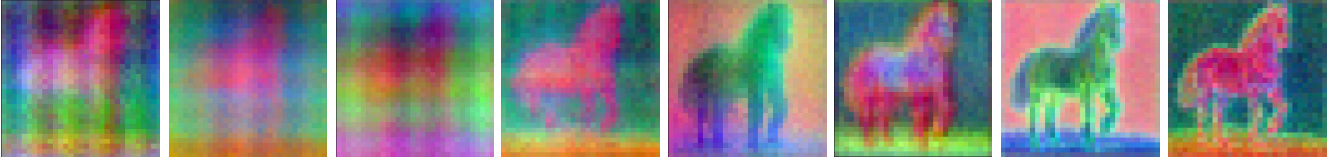


Figure 3: Visualization of the *Query* features in the self-attention layers of the PixArt- $\alpha$ . Features in the deeper layers (right side) are observed to capture the semantic layout more effectively than those in the shallow layers (left side).

### Diffusion Transformer-based Image Editing

In this section, we introduce the components of our proposed DiT4Edit. As shown in Figure 2, based on a pretrained diffusion transformer, the pipeline of our image editing framework is as follows.

**DPM-Solver inversion.** As we discussed earlier, using a high-order DPM-Solver (e.g., DPM-Solver++), can effectively improve the sampling speed. To approximate the integral term  $\int_{\lambda_s}^{\lambda_t} e^{-\lambda} \mathbf{x}_\theta d\lambda$  in equation Eq. 4, given the  $x_{t_{i-1}}$  at time  $t_{i-1}$ , using the Taylor expansion at  $\lambda_{t_{i-1}}$ , and the DPM-Solver++ can obtain an exact solution value at time  $t_i$ :

$$\mathbf{x}_{t_i} = \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} \mathbf{x}_{t_{i-1}} + \sigma_{t_i} \underbrace{\sum_{n=0}^{k-1} \mathbf{x}_\theta^{(n)}(\mathbf{x}_{\lambda_{t_{i-1}}}, \lambda_{t_{i-1}})}_{\text{estimated}} + \underbrace{\int_{\lambda_{t_{i-1}}}^{\lambda_{t_i}} \frac{e^{\lambda(\lambda - \lambda_{t_{i-1}})^n}}{n!} d\lambda}_{\text{analytically computed}} + \underbrace{\mathcal{O}(h_i^{k+1})}_{\text{omitted}}, \quad (5)$$

Especially when  $k = 1$ , the Eq. 5 is equivalent to DDIM sampler (Song, Meng, and Ermon 2020) as follows:

$$\mathbf{x}_{t_i} = \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} \mathbf{x}_{t_{i-1}} - \alpha_{t_i} (e^{-h_i} - 1) \mathbf{x}_\theta(\mathbf{x}_{t_{i-1}}, t_{i-1}), \quad (6)$$

In practical applications, it is common to set  $k = 2$ , enabling a rapid inference and minimizing discretization errors. This DPM-Solver named DPM-Solver++ (2M) (Lu et al. 2023):

$$\mathbf{x}_{t_i} = \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} \mathbf{x}_{t_{i-1}} - \alpha_{t_i} (e^{-h_i} - 1) \cdot \left[ \left(1 + \frac{1}{2r_i}\right) \mathbf{x}_\theta(\mathbf{x}_{t_{i-1}}, t_{i-1}) - \frac{1}{2r_i} \mathbf{x}_\theta(\mathbf{x}_{t_{i-2}}, t_{i-2}) \right], \quad (7)$$

where 2M means this solver is a second-order multistep solver.

However, during the inversion stage for the high-order samplers such as DPM-Solver++(2M), to obtain the inversion result  $\mathbf{x}_{t_i}$  in current timestep  $t_i$ , we need to approximate the values in prior timesteps like  $\{t_{i-2}, t_{i-3}, \dots\}$  for the estimated and the analytically computed terms in Eq. 5:

$$\sigma_{t_i} \sum_{n=0}^{k-1} \mathbf{x}_\theta^{(n)}(\mathbf{x}_{\lambda_{t_{i-1}}}, \lambda_{t_{i-1}}) \int_{\lambda_{t_{i-1}}}^{\lambda_{t_i}} \frac{e^{\lambda(\lambda - \lambda_{t_{i-1}})^n}}{n!} d\lambda, \quad (8)$$

A recent work (Hong et al. 2024) introduced a strategy via the backward Euler method to get the high-order term ap-

proximation in Eq. 8 as follows:

$$\mathbf{d}'_i = \mathbf{z}_\theta(\hat{\mathbf{z}}_{t_{i-1}}, t_{i-1}) + \frac{\mathbf{z}_\theta(\hat{\mathbf{y}}_{t_{i-1}}, t_{i-1}) - \mathbf{z}_\theta(\hat{\mathbf{y}}_{t_{i-2}}, t_{i-2})}{2r_i}, \quad (9)$$

where  $\mathbf{z}_\theta$  is the denoising model,  $\{\hat{\mathbf{y}}_{t_{i-1}}, \hat{\mathbf{y}}_{t_{i-2}}, \dots\}$  is a set of value calculated by  $\hat{x}_{t_i}$  through the DDIM inversion to estimate the  $(\hat{x}_{t_{i-1}}, \hat{x}_{t_{i-2}})$  in Eq. 8, and  $r_i = \frac{\lambda_{t_{i-1}} - \lambda_{t_{i-2}}}{\lambda_{t_i} - \lambda_{t_{i-1}}}$ . Then we can get a inversion latent  $\hat{\mathbf{z}}_{t_{i-1}}$  in current timestep by:

$$\hat{\mathbf{z}}_{t_{i-1}} = \hat{\mathbf{z}}_{t_{i-1}} - \rho(\mathbf{z}'_{t_i} - \hat{\mathbf{z}}_{t_i}), \quad (10)$$

where  $\mathbf{z}'_{t_i} = \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} \hat{\mathbf{z}}_{t_{i-1}} - \alpha_{t_i}(e^{-h_i} - 1)\mathbf{d}'_i$ . In DiT4Edit, we utilize the DPM-Solver++ inversion strategy to obtain an inversion latent from input image  $x_0$  for the editing task. Additionally, this technique was not used in previously UNet-based image editing methods. Furthermore, we observe that we can still obtain a good inversion latent map without using DDIM inversion to calculate the values of  $\hat{\mathbf{y}}$ .

**The unified control of attention mechanism.** In the previous work Prompt to Prompt (P2P) (Hertz et al. 2022), researchers demonstrate that the cross attention layers contain rich semantic information from prompt texts. This finding can edit images by replacing the cross attention maps between the source image and target image during the diffusion process. Specifically, the two commonly used text-guided cross attention strategies are cross attention replacement and cross-attention refinement. These two methods ensure the seamless flow of information from the target prompt to the source prompt, thereby guiding the latent map toward the desired direction.

Unlike cross attention, the self-attention mechanism in the diffusion transformer is utilized to guide the formation of image layout, a feature that cannot be accomplished by the cross-attention mechanism. As shown in Figure 3, the object and layout information from the prompt are not fully captured in the query vectors of the transformer’s shallow layers but are well-represented in the deeper layers. Moreover, with an increasing number of transformer layers, the query vectors’ ability to capture object details becomes clearer and more specific. This suggests that the transformer’s global attention mechanism is more effective at capturing long-range object information, making DiT particularly advantageous for large-scale deformation and editing of extensive images. This observation suggests that non-rigid editing of images can be achieved by controlling self attention mechanism. In MasaCtrl (Cao et al. 2023), researchers introduced the mutual self attention control mechanism. To be specific, in the early steps of diffusion, the feature in the editing steps  $Q_{tar}$ ,  $K_{tar}$ , and  $V_{tar}$  will be used in self attention calculation to generate an image layout closer to the target prompt, while in the later stages, the feature in the reconstruction steps —  $K_{src}$  and  $V_{src}$  will be used to guide the generation of the target image layout closer to the original image.

However, MasaCtrl may still encounter some failure cases, which can be caused by its use of  $Q_{tar}$  throughout the entire editing process, as mentioned in a recent work (Xu

et al. 2024). To address this issue, we determine when to adopt  $Q_{src}$  by setting a threshold  $S$  for the number of steps:

$$\text{Mutual Edit} = \begin{cases} \text{Attention}\{Q_{src}, K_{src}, V_{src}\}, & \text{if } t > S \\ \text{Attention}\{Q_{tar}, K_{src}, V_{src}\}, & \text{otherwise} \end{cases} \quad (11)$$

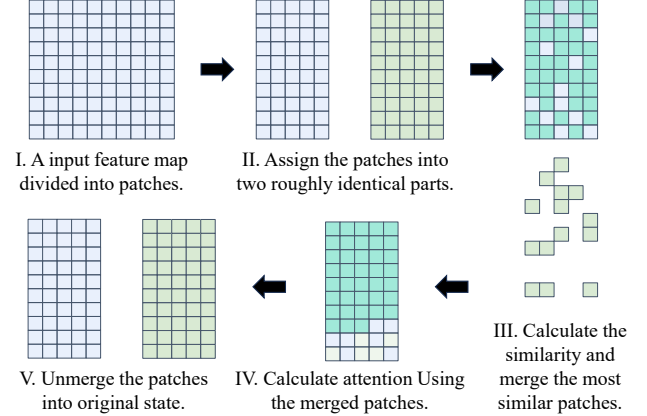


Figure 4: The process of patch merging.

**Patches merging.** To enhance the inference speed, inspired by token merging (Bolya et al. 2023), we embed patches merging into the denoising model. This approach is motivated by the observation that the number of patches involved in attention calculations within the transformer architecture is significantly greater than that in UNet. The calculation flow is shown in Figure 4. For a feature map, we first compute the similarity between each patch and merge the most similar ones to reduce the number of patches processed by the attention mechanism. After attention calculation, we unmerge the patches to maintain the original input size for the next layer in the model. By incorporating patches merging into our framework, we aim to streamline the process and improve overall efficiency, without altering the fundamental operations of each layer.

## Experiments

### Implementation Details

For editing tasks involving images with a scale of  $512 \times 512$  and larger sizes up to  $1024 \times 2048$ , we use pre-trained models PixArt- $\alpha$  - XL -  $512 \times 512$  version for the smaller scale and PixArt- $\alpha$  - XL -  $1024 \times 1024$  - MS version for the larger scale (Chen et al. 2023). We conduct editing on both real and generated images. For the real image input, we use DPM-Solver inversion to get the latent noise map. We configured the DPM-Solver with 30 steps, the classifier-free guidance of 4.5, and a patch merging ratio of  $0.3 \sim 0.7$ . All experiments were run on an NVIDIA Tesla A100 GPU.

### Qualitative Comparison

We evaluate the qualitative performance differences between our proposed DiT4Edit editing framework and six prior baselines, all implemented using official open-source code.

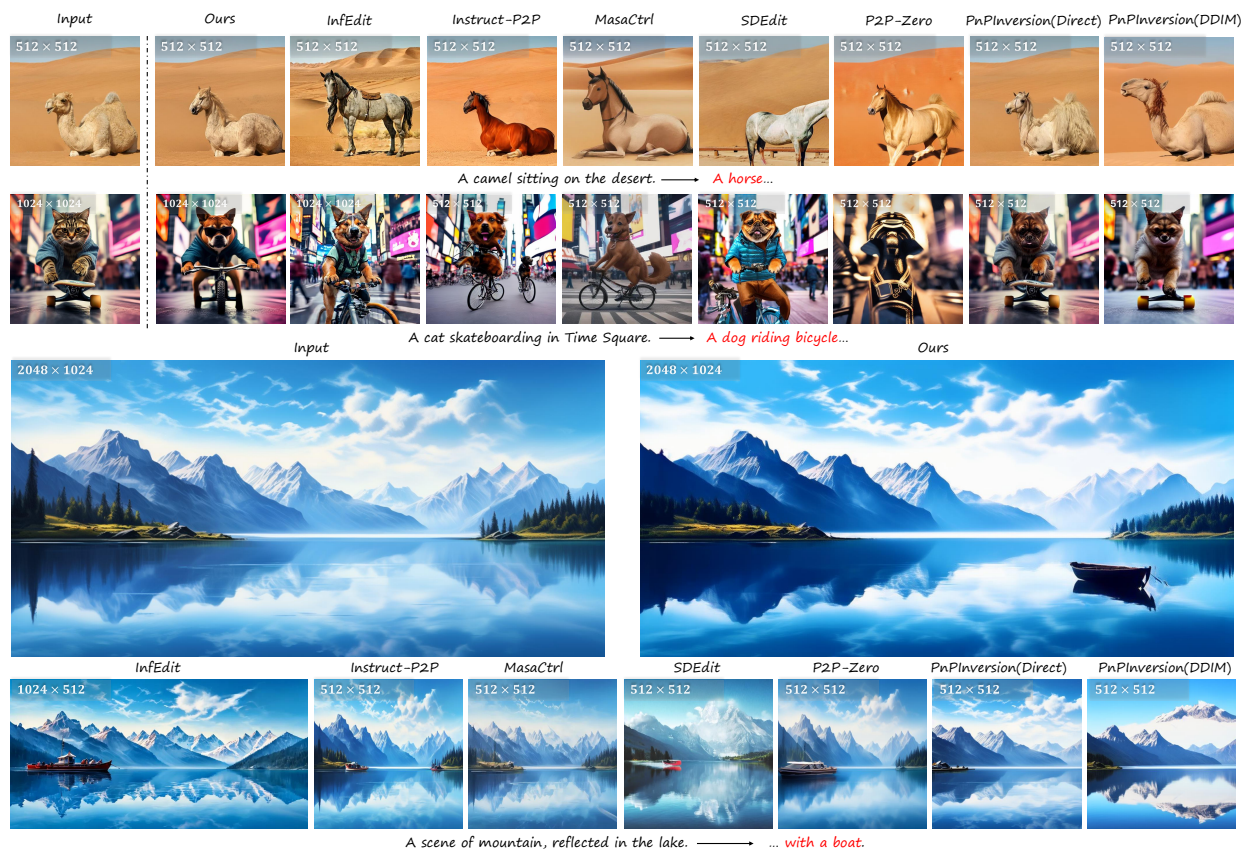


Figure 5: Comparison of  $512 \times 512$ ,  $1024 \times 1024$ , and high-resolution, non-typical aspect-ratio images against the baseline. DiT4Edit achieves satisfactory consistency in editing results.

Metrics		FID			PSNR			CLIP			Inference Time (s)
Models	Structure	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	(a)
SDEdit (Meng et al. 2021)	UNet-based	93.25	88.56	143.87	21.54	20.76	16.35	22.41	21.39	20.73	15.62
IP2P (Brooks, Holynski, and Efros 2023)		88.63	98.73	103.52	20.36	19.73	17.13	20.64	19.36	21.53	12.43
Pix2Pix-Zero (Parmar et al. 2023)		92.31	101.32	158.29	20.6	16.27	23.58	22.12	17.85	20.39	31.45
MasaCtrl (Cao et al. 2023)		110.75	176.15	236.49	17.35	20.51	16.92	23.51	21.96	15.23	19.76
InEdit (Xu et al. 2024)		86.28	87.42	98.75	21.54	22.36	21.74	24.46	23.74	22.16	<b>5.08</b>
PnPInversion (Ju et al. 2024)		84.73	85.33	110.73	21.71	23.42	24.59	21.71	20.76	19.35	30.48
Ours	DiT-based	<b>72.36</b>	<b>62.45</b>	<b>75.43</b>	<b>22.85</b>	<b>29.75</b>	<b>27.46</b>	<b>25.39</b>	<b>26.97</b>	<b>25.66</b>	5.15

Table 1: Quantitative comparison results. We compare our model with six prior works, all implemented using official open-source code. Different settings: (a)  $512 \times 512$ , (b)  $1024 \times 1024$ , (c)  $1024 \times 2048$ .

As shown in Figure 5, we compare our method on  $512 \times 512$  and  $1024 \times 1024$  images. The first row of Figure 5 demonstrates that our framework can generate edited images that remain consistent with the original content when editing real  $512 \times 512$  images, whereas existing methods often alter the background or target details of the original image. Furthermore, the second and third rows of Figure 5 illustrate our experiments with large-scale images and arbitrarily sized images—tasks that previous UNet-based methods struggled to address. The results indicate that our proposed framework effectively handles style and object shape modifications in larger images. In contrast, some state-of-the-art UNet-based methods, despite being capable of performing

editing tasks, frequently result in significant alterations and damage to the background and object locations in the source image. Additionally, due to the limitations of the UNet structure, these methods typically generate target images only at a size of  $512 \times 512$ . These findings emphasize the substantial potential of transformer-based diffusion models in large-scale image editing. We also perform the user study for comprehensive comparisons. The details of the user study can be found in the supplementary material.

### Quantitative Comparison

For quantitative evaluation, we used three indicators: Fréchet Inception Distance (FID) (Heusel et al. 2017), Peak

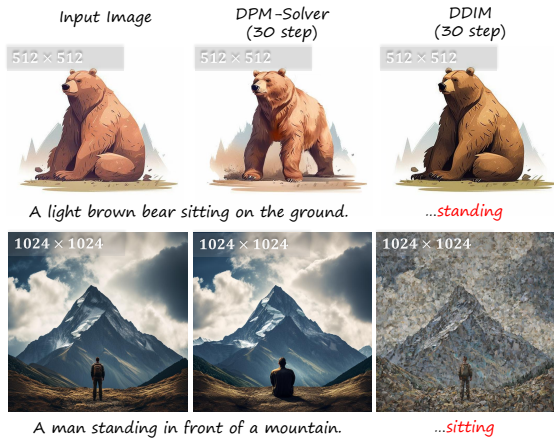


Figure 6: Ablation study on DPM-Solver inversion. Experiments have demonstrated that the DPM solver achieves superior editing results compared to DDIM.

Signal-to-Noise Ratio (PSNR), and CLIP to evaluate the performance differences between our model and SOTA in image generation quality, background preservation, and text alignment. We compared images at three sizes:  $512 \times 512$ ,  $1024 \times 1024$ , and  $1024 \times 2048$ , with results detailed in Table 1. We perform the performances with Pix2Pix-Zero, PnP-Inversion, SDEdit, IP2P, MasaCtrl, and InfEdit. It should be noted that, since no DiT-based editing framework previously existed, all our comparison baselines are based on the UNet architecture. The experimental results show that our proposed DiT4Edit editing strategy outperforms SOTA methods in image generation quality, background preservation, and text alignment. Due to the global attention capabilities of the integrated transformer structure, the DiT4Edit framework exhibits strong robustness across editing tasks of various sizes. The generated images not only show higher quality but also offer better control over the background and details, resulting in greater consistency with the original image. Particularly for editing large or arbitrarily sized images, DiT4Edit demonstrates significant advantages over other methods, showcasing the powerful scaling ability of the transformer architecture. Meanwhile, our editing framework has a shorter inference time, comparable to the inversion-free editing method (InfEdit).

Image Size	Patches Merging	Speed (s)
$512 \times 512$	✓	6.01
	✗	9.13
$1024 \times 1024$	✓	34.27
	✗	41.52
$1024 \times 2048$	✓	99.39
	✗	122.41

Table 2: Ablation study on patches merging. The result demonstrates that this technique accelerates model inference speed, without affecting the quality of the image generation.



Figure 7: Ablation study on patch merging. The results indicate that this module does not impact the editing quality.

### Ablation Study

We perform a series of ablation studies to demonstrate the effectiveness of **DPM-Solver inversion** and **patches merging**. The results of our ablation experiments on **patches merging** are presented in Figure 4 and Table 2. Implementing patch merging led to a notable reduction in the editing time for large-sized images while maintaining editing quality comparable to that achieved without patch merging. This indicates that patch merging can significantly enhance the overall performance of image editing frameworks. Furthermore, the ablation experiment results for **DPM-Solver and DDIM** are illustrated in Figure 6. When comparing the two methods with the same number of inference steps ( $T = 30$ ), DPM Solver consistently outperformed DDIM in terms of image editing quality. This demonstrates that our use of the DPM-Solver inversion strategy allows for the generation of superior latent maps, resulting in better editing outcomes within fewer steps.

### Discussion and Conclusion

**Conclusion.** We introduce DiT4Edit, the first image-editing framework based on a diffusion transformer. Unlike previous UNet-based frameworks, DiT4Edit offers superior editing quality and supports images of various sizes. Leveraging DPM Solver inversion, a unified attention control mechanism, and patch merging, DiT4Edit outperforms the UNet structure in editing tasks for images sized  $512 \times 512$  and  $1024 \times 1024$ . Notably, DiT4Edit can handle images of arbitrary sizes, such as  $1024 \times 2048$ , showcasing the transformer’s advantages in global attention and scalability. Our research can set a baseline for DiT-based image editing and help further explore the potential of transformer structures in generative AI.

**Limitation.** In our experiment, we observed that the T5-tokenizer occasionally encounters issues with word segmentation, which can lead to failures in the final editing process. Additionally, our model might experience color inconsistencies compared to the original image.

## Acknowledgments

This research was supported by the Research Grant Council of the Hong Kong Special Administrative Region under grant number 16203122, the Innovation and Technology Fund of HKSAR under grant number GHX/054/21GD and Guangzhou Education Bureau Project #2024312075.

## References

- Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving Image Generation with Better Captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3): 8.
- Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT But Faster. In *The Eleventh International Conference on Learning Representations*.
- Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning To Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.
- Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.
- Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. PixArt- $\alpha$ : Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*.
- Chen, Q.; Ma, Y.; Wang, H.; Yuan, J.; Zhao, W.; Tian, Q.; Wang, H.; Min, S.; Chen, Q.; and Liu, W. 2024. Follow-Your-Canvas: Higher-Resolution Video Outpainting with Extensive Content Generation. *arXiv preprint arXiv:2409.01055*.
- Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2108.02938*.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.
- Dong, W.; Xue, S.; Duan, X.; and Han, S. 2023. Prompt Tuning Inversion for Text-driven Image Editing Using Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7430–7440.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27.
- He, Y.; Yang, S.; Chen, H.; Cun, X.; Xia, M.; Zhang, Y.; Wang, X.; He, R.; Chen, Q.; and Shan, Y. 2023. ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models. In *The Twelfth International Conference on Learning Representations*.
- Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626*.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.
- Hong, S.; Lee, K.; Jeon, S. Y.; Bae, H.; and Chun, S. Y. 2024. On Exact Inversion of DPM-Solvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7069–7078.
- Ju, X.; Zeng, A.; Bian, Y.; Liu, S.; and Xu, Q. 2024. PnP Inversion: Boosting Diffusion-based Editing with 3 Lines of Code. In *The Twelfth International Conference on Learning Representations*.
- Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-Based Real Image Editing With Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.
- Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.
- Liang, J.; Pei, W.; and Lu, F. 2020. CPGAN: Content-Parsing Generative Adversarial Networks for Text-to-Image Synthesis. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 491–508. Springer.
- Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.
- Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2023. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *arXiv:2211.01095*.
- Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047*.
- Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024a. Follow Your Pose: Pose-Guided Text-to-Video Generation Using Pose-Free Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.
- Ma, Y.; He, Y.; Wang, H.; Wang, A.; Qi, C.; Cai, C.; Li, X.; Li, Z.; Shum, H.-Y.; Liu, W.; et al. 2024b. 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. 2024c. 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. 2022a. Visual Knowledge Graph for Human Action Reasoning in Videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4132–4141.
- Ma, Y.; Yang, T.; Shan, Y.; and Li, X. 2022b. SimVTP: Simple Video Text Pre-training with Masked Autoencoders. *arXiv preprint arXiv:2212.03490*.
- Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *arXiv preprint arXiv:2108.01073*.
- Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. NULL-Text Inversion for Editing Real Images Using Guided Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.
- Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot Image-to-Image Translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.
- Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.
- 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 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*.
- Song, J.; Wang, B.; Wang, Z.; and Yip, D. K.-M. 2023. From expanded cinema to extended reality: How AI can expand and extend cinematic experiences. In *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*, 1–5.
- Wang, B.; Chen, Q.; and Wang, Z. 2024. Diffusion-Based Visual Art Creation: A Survey and New Perspectives. *arXiv preprint arXiv:2408.12128*.
- Wang, B.; Meng, H.; Cai, Z.; Li, L.; Ma, Y.; Chen, Q.; and Wang, Z. 2023. MagicScroll: Nontypical Aspect-Ratio Image Generation for Visual Storytelling via Multi-Layered Semantic-Aware Denoising. *arXiv preprint arXiv:2312.10899*.
- Wang, B.; Zhang, K.; and Wang, Z. 2023. Naturality: A Natural Reflection of Chinese Calligraphy. In *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*, 1–8.
- Wang, J.; Ma, Y.; Guo, J.; Xiao, Y.; Huang, G.; and Li, X. 2024. COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing. *arXiv preprint arXiv:2406.08850*.
- Xu, S.; Huang, Y.; Pan, J.; Ma, Z.; and Chai, J. 2024. Inversion-Free Image Editing with Language-Guided Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9452–9461.
- Yang, L.; Liu, J.; Hong, S.; Zhang, Z.; Huang, Z.; Cai, Z.; Zhang, W.; and Cui, B. 2024. Improving Diffusion-Based Image Synthesis with Context Prediction. *Advances in Neural Information Processing Systems*, 36.
- 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, Q.; and Chen, Y. 2022. Fast Sampling of Diffusion Models with Exponential Integrator. *arXiv preprint arXiv:2204.13902*.
- Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023. SINE: SINGle Image Editing With Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6027–6037.
- Zhang, Z.; Xie, Y.; and Yang, L. 2018. Photographic Text-to-Image Synthesis With a Hierarchically-Nested Adversarial Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6199–6208.
- Zhu, C.; Li, K.; Ma, Y.; Tang, L.; Fang, C.; Chen, C.; Chen, Q.; and Li, X. 2024. InstantSwap: Fast Customized Concept Swapping across Sharp Shape Differences. *arXiv preprint arXiv:2412.01197*.