AdapEdit: Spatio-Temporal Guided Adaptive Editing Algorithm for Text-Based Continuity-Sensitive Image Editing

Zhiyuan Ma[1,2*], **Guoli Jia**[3*], **Bowen Zhou**[1†]

[1]Department of Electronic Engineering, Tsinghua University
[2]Shanghai Artificial Intelligence Laboratory
[3]College of Computer Science, Nankai University
{mzyth, zhoubowen}@tsinghua.edu.cn, exped1230@gmail.com

Abstract

With the great success of text-conditioned diffusion models in creative text-to-image generation, various text-driven image editing approaches have attracted the attention of many researchers. However, previous works mainly focus on discreteness-sensitive instructions such as adding, removing or replacing specific objects, background elements or global styles (*i.e., "hard editing"*), while generally ignoring subject-binding but semantically fine-changing continuity-sensitive instructions such as actions, poses or adjectives, and so on (*i.e., "soft editing"*), which hampers generative AI from generating user-customized visual contents. To mitigate this predicament, we propose a spatio-temporal guided **ada**ptive image **edit**ing algorithm AdapEdit, which realizes adaptive image editing by introducing a soft-attention strategy to dynamically vary the guiding degree from the editing conditions to visual pixels from both temporal and spatial perspectives. Note our approach has a significant advantage in preserving model priors and does not require model training, fine-tuning, extra data, or optimization. We present our results over a wide variety of raw images and editing instructions, demonstrating competitive performance and showing it significantly outperforms the previous approaches. Code is available: https://github.com/AnonymousPony/adap-edit.

1 Introduction

Text-conditioned image synthesis has recently received tremendous success with various magnificent and creative text-to-image cases being produced, where diffusion probability models (DPMs) (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021), as the backbone network that plays an important role, has been developed into a variety of text-to-image generation applications (Saharia et al. 2022; Ramesh et al. 2022; Ma et al. 2023a). However, the controllability of image synthesis and the support for user-customization are still challenging issues and have received more and more researchers' attention. To solve these problems, Stable Diffusion (Rombach et al. 2022) and ControlNet (Zhang and Agrawala 2023), as two leading studies, have proposed two efficient text-conditioned learning frameworks for training, and subsequently, a series of text-guided image editing methods (Hertz et al. 2022; Ruiz et al. 2023; Wallace, Gokul, and Naik 2023; Kawar et al. 2023; Mokady et al. 2023) have been proposed, which achieve user-customized image output by editing or modifying the original image given the human-written instructions, and have witnessed a huge improvement in terms of controllability.

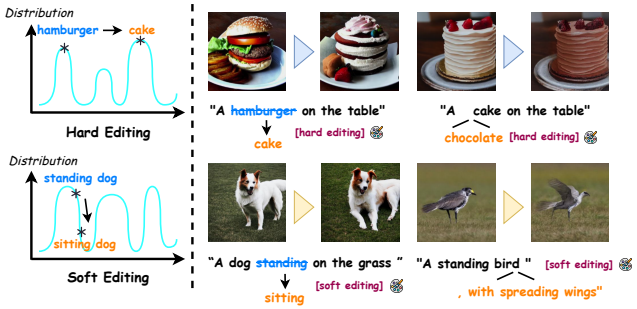

Figure 1: Example of image editing to show discreteness-sensitive image manipulations (*i.e., hard editing*) and continuity-sensitive image manipulations (*i.e., soft editing*).

Though achieving remarkable progress, existing text-based image editing methods still suffer from the following three limitations. (1) *Soft editing dilemma.* **Firstly**, as illustrated in Figure 1, we notice that the previous approaches mainly focus on relatively simple editing instructions such as adding, removing or replacing specific objects, background elements or global styles (*i.e., hard editing*), while generally ignoring subject-binding but semantically fine-changing complex instructions such as actions, poses or adjectives, and so on (*i.e., soft editing*), which hampers generative AI from generating user-customized images. (2) *Spatio-temporal continuity dilemma.* **Secondly**, we notice that the previous methods generally adopt the same guide degree during the text-to-image diffusion process for each pixel of the original image (*i.e., guided target*) and each word of the textual condition (*i.e., guiding source*), which may result in artifacts emerged in the unexpected areas, as well as may also cause semantic discontinuities with surrounding pixels to appear around the edited objects. (3) *Training dilemma.* **Thirdly**, to improve model's controllable generation ability under given editing-instructions, most of the previous methods need to feed a large number of labeled

*Equal contributions.
†Corresponding author.

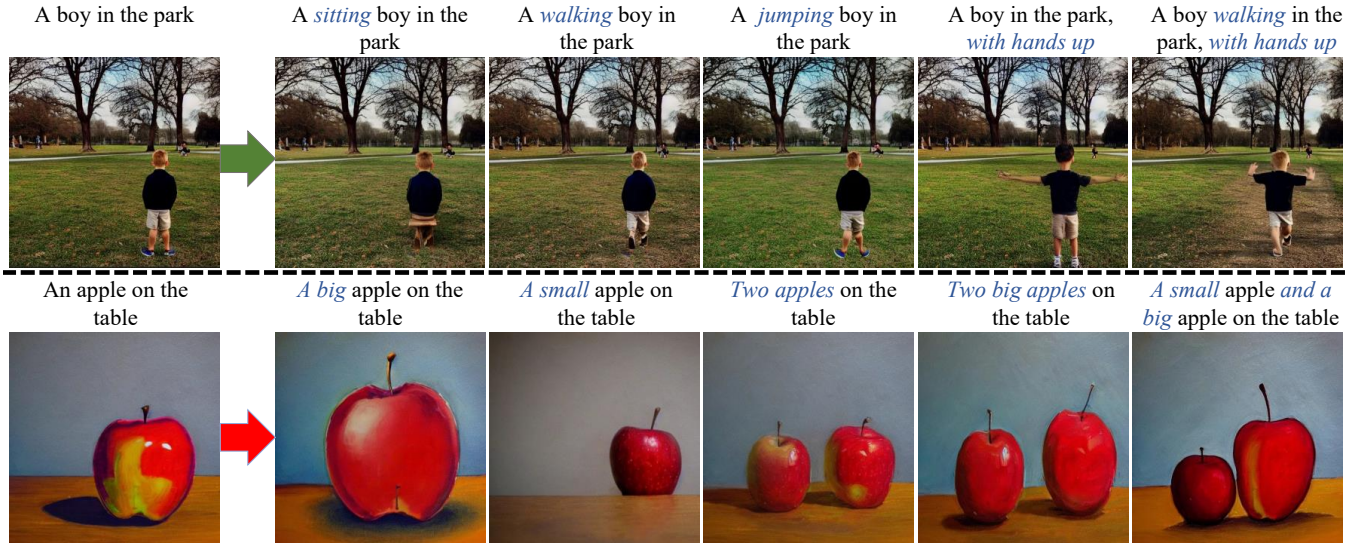

Figure 2: The performance of AdapEdit with soft editing instructions. The leftmost images are directly generated by the original condition, images on other lines are edited by the original and editing conditions.

image-text pairs as training data, and also need to design additional optimization objectives for training. However, the costly training often causes the *language drift* phenomenon appears in image priors conditioned on the editing instructions training, whereas forget the priors under the original unedited instructions.

To address the aforementioned limitations, we propose *AdapEdit*, a spatio-temporal guided adaptive editing algorithm for complex continuity-sensitive image editing tasks. Specifically, to address the first limitation, we introduce a soft attention strategy into algorithm to adaptively guide the attention degrees from textual words to visual pixels for controllable text-to-image editing. Based on the soft attention strategy, we further address the second limitation of *continuity dilemma* by respectively designing a flexible word-level temporal (FWT) adjustment module and a dynamic pixel-level spatial (DPS) weighting module to achieve spatio-temporal guided adaptive editing through spatial guiding scales $\mathbf{s}_{[V]}$ and temporal guiding scales τ_{c^*}. Moreover, we introduce a spatial interpolation weight λ_S, which is a hyper-parameter and is adopted to finely control the editing amplitude at the pixel level. Finally, the entire AdapEdit algorithm is executed in only one forward diffusion process and does not require any training or optimization with the algorithm time-overhead is as low as $O(n)$, which naturally solves the third *training dilemma*. Our main contributions are summarized as follows:

- **Soft editing capabilities enhancement.** The ability to support more complex editing tasks, such as soft editing tasks, has been enhanced for better user-customized image generation scenarios.
- **Editing naturalness and contextual continuity improvement.** By assigning variable spatio-temporal guidance scales to adaptively guide the attentions from textual tokens to visual pixels, the naturalness and contextual semantic coherence of the edited images are significantly improved.
- **No training required and no damage to model priors.** Since a deterministic adaptive algorithm is developed, it requires almost no training costs such as training data, annotations, additional optimization targets, and huge GPU overheads. Moreover, more importantly, our algorithm does not break the priors of the DPMs themselves.

2 Related Work

Diffusion-Based Generative Models. Recent years has witnessed the remarkable success of diffusion-based generative models, due to their excellent performance in the diversity and impressive generative capabilities. These previous efforts mainly focus in sampling procedure (Song et al. 2020; Liu et al. 2022), conditional guidance (Dhariwal and Nichol 2021; Nichol et al. 2021), likelihood maximization (Kingma et al. 2021; Kim et al. 2022) and generalization ability (Kawar et al. 2022; Ma et al. 2023b) and have enabled state-of-the-art image synthesis.

Text-Guided Image Synthesis Models. Text-guided image synthesis aims to generate images conforming to the text semantics by sampling and predicting the noise distributions in diffusion models. Take advantage of guidance diffusion techniques, recent large text-to-image models such as Imagen (Saharia et al. 2022), DALL-E2 (Ramesh et al. 2022), Parti (Yu et al. 2022), CogView2 (Ding et al. 2022) and Stable Diffusion (Rombach et al. 2022), have synthesized a wide variety of unprecedented images and shown excellent performance. However, it is worth noting that these models only support sampling in a random Gaussian distribution conditioned on given text prompts, whereas cannot achieve

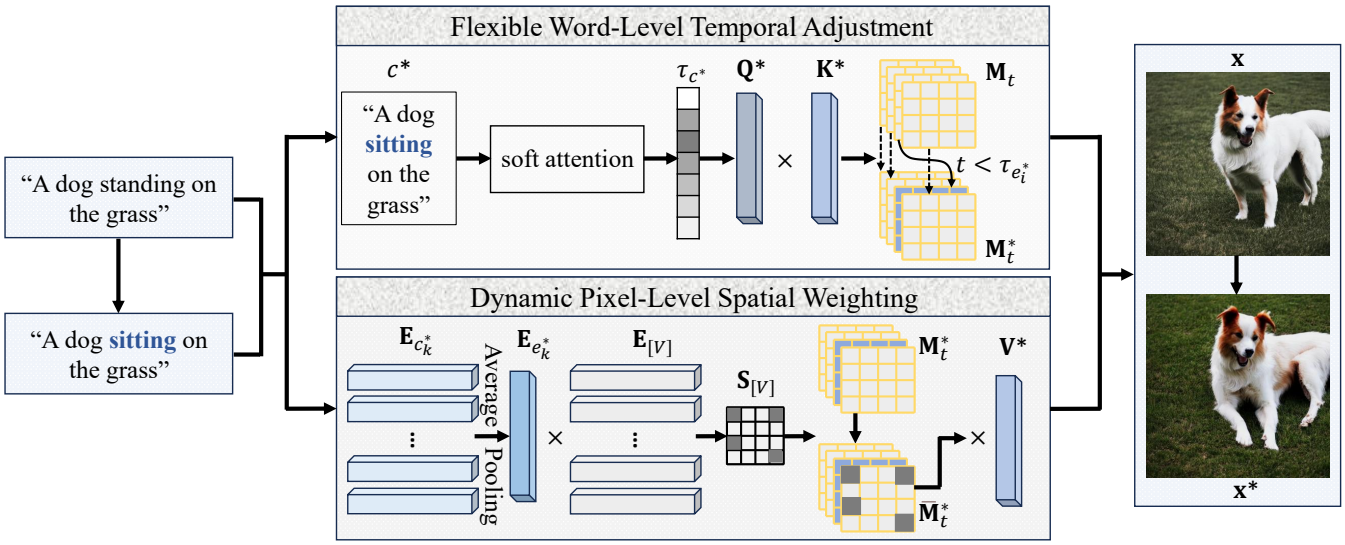

Figure 3: The framework overview of the proposed AdapEdit algorithm.

customized image generation through directly editing on the user-uploaded images, which hampers from supporting more complex and challenging image customizations tasks.

Text-Guided Image Editing Models. To solve the problem of controllability and support customizations, a series of diffusion-based image editing models have been proposed. As a leading research in controllable text-to-image generation, **ControlNet** (Zhang and Agrawala 2023) proposes an end-to-end neural architecture that controls aforementioned large diffusion models (e.g., Stable Diffusion) to learn task-specific input conditions for efficiently supporting condition-guided image synthesis. Based on this effective fine-tuning framework, a series of image editing methods have been proposed and have advanced the controllability research. **Prompt-to-Prompt** (Hertz et al. 2022) is one of the most representative models, which implements directed editing by performing a series of hard operations such as insert, replace, and reweight over attention maps of the cross-attention layer. While it still suffers from the accumulated errors caused by classifier-free guidance. To address the issue, **Null-Text-Inversion** (Mokady et al. 2023) introduces pivotal inversion technique to rectify the accumulated errors and proposes a null text optimization scheme for better performance. Subsequently, to implement subject-binding editing, **DreamBooth** (Ruiz et al. 2023) proposes to introduce a unique identifier for subject-specific binding, which is fine-tuned on a pretrained diffusion model with a class-specific prior preservation objective. Meanwhile, **Imagic** also utilizes a pre-trained text-to-image diffusion model and first uses it to find a optimal text embedding for the input image by fine-tuning the diffusion model, then adopts linearly interpolate technique to obtain a semantically meaningful mixture for effective editing.

Despite remarkable success, these recent image editing efforts still suffer from the aforementioned dilemmas, *soft editing*, *spatio-temporal continuity*, and *training dilemma*, which prevent generative models from moving toward more fine-grained control and more universal customization. To the end, we propose AdapEdit algorithm, which can be applied to all of the above models to improve their editing performance and can be applied to cope with more complex editing instructions (i.e., continuity-sensitive instructions) such as postures, actions, adjectives, and other subject-binding but semantically fine-changing instructions.

3 Methodology

3.1 Preliminaries

Diffusion models. The diffusion model is modeled as: 1) a deterministic forward noising process $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I})$ from the original image $\mathbf{x}_0$ to a pure-Gaussian distribution $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, which can be formulated in an accumulated form as follows:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

2) and a iteratively predictable reverse denoising process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$, which can be trained in a simplied denoising objective $\mathcal{L}_{\text{simple}}$ by merging $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ into predicting noise $\boldsymbol{\epsilon}_\theta$,

$$\mathcal{L}_{\text{simple}} = E_{\mathbf{x}_0,t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)||_2^2] \tag{2}$$

where $t \sim \mathcal{U}[1, T]$ is time parameters, $\mathcal{U}(\cdot)$ denotes uniform distribution. Moreover, in Stable Diffusion (Rombach et al. 2022), the image $\mathbf{x}_t$ is compressed into a latent variable $\mathbf{z}_t$ by encoder $\mathcal{E}$ for more efficient training, i.e., $\mathbf{z}_t = \mathcal{E}(\mathbf{x}_t)$, thus this preliminary objective is usually defined as making $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$ as close to $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ as possible.

Text-guided diffusion models. The core of text-guided diffusion models is to integrate the semantics of text condition $\boldsymbol{c}$ into noise prediction model $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$ to generate visual contents conforming to text semantics, i.e., $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \boldsymbol{c})$. The

Algorithm 1: AdapEdit algorithm

Input: The text condition $\boldsymbol{c}$, editing instruction $\boldsymbol{c}^*$
Output: Image $\mathbf{x}$, edited image $\mathbf{x}^*$

Initialize: Random seed and parameters
Function FWT:
if last diffusion step **then**
  Calculate the average of the cross-attention maps $\mathbf{M}_t^{\boldsymbol{c}^*}$;
  Multiply($\mathbf{M}_t^{\boldsymbol{c}^*}, \mathbf{E}_v) \rightarrow \mathbf{E}_{\boldsymbol{c}^*}$ (Eq.5);
  Average Pooling($\mathbf{E}_{\boldsymbol{e}_i^*}, i \in \boldsymbol{c}_k^*) \rightarrow \mathbf{E}_{\boldsymbol{e}_k^*}$;
  Multiply($\mathbf{E}_{\boldsymbol{c}^*}, \mathbf{E}_{\boldsymbol{e}_k^*}) \rightarrow \boldsymbol{A}_{\boldsymbol{c}^*}$ (Eq.6);
  for $i \in N_{\boldsymbol{c}^*}$ **do**
    Update $\tau_{\boldsymbol{e}_i^*}$ (Eq.7);
  end for
end if
Function DPS:
for t = $T, T-1$,...,1 **do**
  Multiply($\mathbf{E}_{\boldsymbol{e}_k^*}, E_v) \rightarrow \mathbf{S}_{[V]}$ (Eq.9);
  Update $\boldsymbol{\mathcal{C}}$ ($\mathbf{M}_t^{\boldsymbol{c}}, \mathbf{M}_t^{\boldsymbol{c}^*}$) (Eq.8);
  Update($\overline{\mathbf{M}}_t^{\boldsymbol{c}^*}$) (Eq.4);
end for
Generate $\{\mathbf{x}, \mathbf{x}^*\}$;
return $\{\mathbf{x}, \mathbf{x}^*\}$.

classifier-free guidance technique has recently been widely adopted in text-guided image generation as,

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, t, \boldsymbol{c}, \varnothing) = w \cdot \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \boldsymbol{c}) + (1-w) \cdot \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \varnothing) \quad (3)$$

where $w = 7.5$ is default linear parameter for weighting the unconditional guidance objective and conditional guidance objective in Stable Diffusion, t is time input, $\boldsymbol{c}$ is text condition, $\varnothing$ denotes null text embedding initialized by zero vector and θ is model parameters. Note that all of these parameters will be individually or jointly optimized for controlled image editing in diffusion-based variants.

3.2 AdapEdit Algorithm Framework

The pipeline of the proposed AdapEdit is illustrated in Figure 3. To facilitate user-friendly editing, the model's input consists of two textual conditions $\{\boldsymbol{c}, \boldsymbol{c}^*\}$, which is the same with prompt-to-prompt (Hertz et al. 2022), where $\boldsymbol{c}$ is the original text condition, and $\boldsymbol{c}^*$ denotes the edited condition. Given the image $\mathbf{x}$ that guidance-generated by $\boldsymbol{c}$, AdapEdit aims to manipulate hidden attention weights to modify the image $\mathbf{x}$ to $\mathbf{x}^*$ with guidance condition $\boldsymbol{c}^*$. Most previous methods mainly focus *hard editing* instructions such as adding, removing, or replacing the visual contents while basically ignore the more challenging *soft editing* instructions (i.e., continuity-sensitive prompts). To the end, we propose AdapEdit, a spatio-temporal guided adaptive editing algorithm.

Specifically, we first propose a flexible word-level temporal (FWT) adjustment module to support different guidance scales being adaptively assigned to sequential words for temporal-guided editing, which is driven by the impacts of each word $\boldsymbol{e}^*$ in the editing instruction $\boldsymbol{c}^*$ is variational.

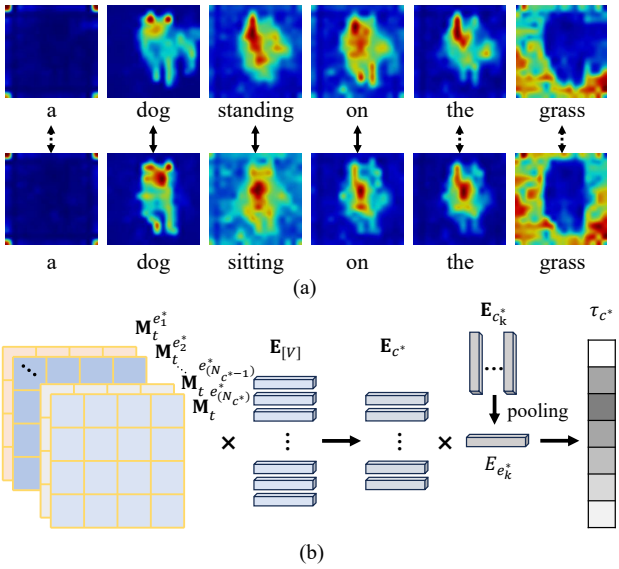

Figure 4: The illustration of our proposed soft attention strategy, in which (a) shows the cross-attention maps from $\boldsymbol{c}$ and $\boldsymbol{c}^*$ and (b) details the specific calculating process.

For further explanation, take the Figure 4(a) as an example, when only editing the attention map from $\mathbf{M}_t^{[\texttt{standing}]}$ into $\mathbf{M}_t^{[\texttt{sitting}]}$, the attention map from "dog" still remains the salient features of "standing", which hampers from the effective editing of the above instruction "a dog sitting...".

To address this issue, we design a soft attention strategy to calculate the $\tau_{\boldsymbol{c}^*}$ for fine-grained word-level editing in FWT module. Further, we subsequently design a dynamic pixel-level spatial (DPS) weighting module to adaptively integrate the edited visual features into original image for spatial-guided editing. Note the spatial-guidance scales $\mathbf{A}_{[S]}$ are determined by calculating the semantic similarities between the replaced words' average representations and the visual features. Based on the above two modules, AdapEdit can reliably control the DPMs (e.g., SD-v1.4) to effectively perform continuity-sensitive soft editing tasks with as many details as possible preserved from the original image $\mathbf{x}$. Next, we detail the separate modules.

Flexible Word-Level Temporal Adjustment Flexible word-level temporal (FWT) adjustment module is adopted to adaptively calculate the temporal-guidance scales $\tau_{\boldsymbol{c}^*}$ for achieving temporal-guided editing. Instead using identical time truncation τ to diffusion models for the key words (here refer to the words that have been modified in $\boldsymbol{c}^*$ compared to $\boldsymbol{c}$) and general words, we first propose a soft attention strategy, depicted in Figure 4 (b), to flexibly calculate $\tau_{\boldsymbol{e}_i^*}$ for each $\{\boldsymbol{e}_i^*\}_{i=1}^{N_{c^*}}$ (N_{c^*} denotes the number of words in $\boldsymbol{c}^*$) in condition $\boldsymbol{c}^*$ to obtain updated attention map $\overline{\mathbf{M}}_t^{\boldsymbol{e}_i^*}$. Formally,

$$\overline{\mathbf{M}}_t^{\boldsymbol{e}_i^*} = \begin{cases} \mathbf{M}_t^{\boldsymbol{e}_i} & t > \tau_{\boldsymbol{e}_i^*}, \\ \boldsymbol{\mathcal{C}}(\mathbf{M}_t^{\boldsymbol{e}_i}, \mathbf{M}_t^{\boldsymbol{e}_i^*}) & \text{otherwise.} \end{cases} \quad (4)$$

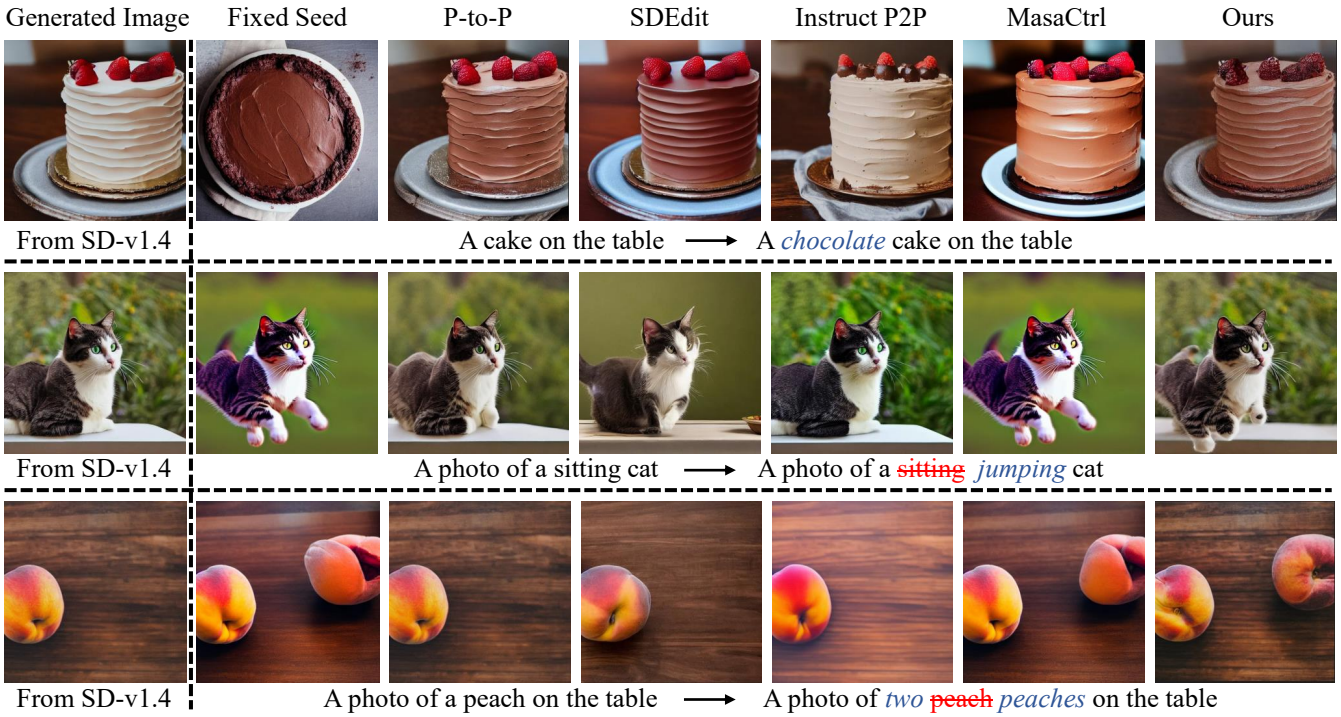

Figure 5: The qualitative comparisons of with the previous SOTA methods. The generated image (the leftmost column) denotes the original image $\mathbf{x}$ conditioned on $\boldsymbol{c}$ generated by SD-v1.4, other columns present the editing results conditioned on $\boldsymbol{c}^*$. Note the fixed seed denotes generating a new image conditioned on $\boldsymbol{c}^*$ by directively using SD-v1.4 with the same random seed.

where $\boldsymbol{e}_i$ and $\boldsymbol{e}_i^*$ respectively denotes i^{th} word in $\boldsymbol{c}$ and $\boldsymbol{c}^*$, $\mathbf{M}_t$ are their corresponding attention maps. When $t \geq \tau_{\boldsymbol{e}_i^*}$, the updated attention map $\overline{\mathbf{M}}_t^{\boldsymbol{e}_i^*}$ equals to $\mathbf{M}_t^{\boldsymbol{e}_i}$ for feature preservation of image $\mathbf{x}$. While when $t < \tau_{\boldsymbol{e}_i^*}$, the updated map $\overline{\mathbf{M}}_t^{\boldsymbol{e}_i^*}$ is obtained by the combination of $\mathbf{M}_t^{\boldsymbol{e}_i}$ and $\mathbf{M}_t^{\boldsymbol{e}_i^*}$, where $\mathcal{C}(\cdot)$ represents the interpolation operation described in Eq. 8, which will be detailed in the next DPS module.

Then, when running to the last diffusion step, we calculate the average of the cross-attention maps among all the layers based on 32×32 resolution. After that, when given the cross-attention map $\{\mathbf{M}_t^{\boldsymbol{e}_i^*}\}_{i=1}^{N_{\boldsymbol{c}^*}}$ of each word and visual embeddings $\mathbf{E}_{\texttt{[V]}} \in \mathcal{R}^{N_{\texttt{[V]}} \times d}$, each $\mathbf{M}_t^{\boldsymbol{e}_i^*} \in \mathcal{R}^{1 \times N_{\texttt{[V]}}}$, $N_{\texttt{[V]}}$ is the number of visual pixels. Followed by the softmax normalization layer, the $\{\mathbf{M}_t^{\boldsymbol{e}_i^*}\}_{i=1}^{N_{\boldsymbol{c}^*}}$ are multiplied with the visual embeddings $\mathbf{E}_{\texttt{[V]}}$ to obtain the textual embeddings $\mathbf{E}_{\boldsymbol{c}^*} \in \mathcal{R}^{N_{\boldsymbol{c}^*} \times d}$ corresponding to the edited text condition $\boldsymbol{c}^*$ as,

$$\mathbf{E}_{\boldsymbol{c}^*} = \mathbf{M}_t^{\boldsymbol{c}^*} \times \mathbf{E}_{\texttt{[V]}}, \tag{5}$$

Next, the embeddings of key words are fused by average pooling, which is defined as $\mathbf{E}_{\boldsymbol{e}_k^*}$. Then, after normalizing the embeddings, the correlation vector $\boldsymbol{A}_{\boldsymbol{c}^*}$ between the words $\{\boldsymbol{e}_i^*\}_{i=1}^{N_{c^*}}$ and the embeddings of replaced words $\mathbf{E}_{\boldsymbol{e}_k^*}$ can be calculated by:

$$\boldsymbol{A}_{\boldsymbol{c}^*} = \mathbf{E}_{\boldsymbol{c}^*} \times \mathbf{E}_{\boldsymbol{e}_k^*}, \boldsymbol{A}_{\boldsymbol{c}^*} \in R^{1 \times N_{\boldsymbol{c}^*}}, \tag{6}$$

Finally, the calculation of $\tau_{\boldsymbol{e}_i^*}$ can be formally defined as:

$$\tau_{\boldsymbol{e}_i^*} = \begin{cases} 0, & \boldsymbol{e}_i^* \in \boldsymbol{c}_k^*, \\ \lambda_{\tau_{\boldsymbol{c}^*}}[1 - exp(\boldsymbol{A}_{\boldsymbol{e}_i^*} - 1)], & \text{otherwise.} \end{cases} \tag{7}$$

To conveniently control the range of $\tau_{\boldsymbol{c}^*}$, we leverage a hyper-parameter $\lambda_{\tau_{\boldsymbol{c}^*}}$ to adjust each $\tau_{\boldsymbol{e}_i^*} \in \tau_{\boldsymbol{c}^*}$. In practice, we observe the $\tau_{\boldsymbol{e}_i^*}$ calculated by the above process is somewhat far from the expected result since the impacts of the background regions. Taking Figure 4 (a) as an example, the background region of "sitting" has large amount of overlaps with the foreground region of "grass". The large amount of non-zero pixels in background region would interfere the $\boldsymbol{A}_{\texttt{[grass]}}$. Therefore, a simple but effective strategy is utilized to obtain a mask for each $\mathbf{M}_t^{\boldsymbol{e}_i^*}$. When the values in $\mathbf{M}_t^{\boldsymbol{e}_i^*} < \alpha_m$, we set them as 0. The α_m is empirically set as 0.03. Eventually, we can obtain the discriminant embeddings to calculate the temporal-guidance scales $\tau_{\boldsymbol{c}^*}$.

Dynamic Pixel-Level Spatial Weighting Dynamic pixel-level spatial (DPS) weighting module is adopted to adaptively calculate the spatial-guidance scales $\mathbf{S}_{\texttt{[V]}}$ corresponding to visual content [V] for achieving spatial-guided editing, here $\mathbf{S}_{\texttt{[V]}}$ denotes the similarity matrix between the visual pixels of [V] and editing tokens of $\boldsymbol{c}^*$. Note since all the words $\boldsymbol{e}_i^* \in \boldsymbol{c}^*$ share the same $\mathbf{S}_{\texttt{[V]}}$, the subscript i will be omitted and the attention maps will be denoted as $\mathbf{M}_t^{\boldsymbol{c}}$ and $\mathbf{M}_t^{\boldsymbol{c}^*}$ as a whole if no confusion occurs in the following discussion. To adaptively integrate the original characteristics,

	LPIPS ↓	CLIP score ↑	CLIP directional ↑
P-to-P (2022)	0.33	26.05	0.11
SDEdit (2021)	0.55	26.52	0.11
Instruct P2P (2023)	**0.27**	24.59	0.04
MasaCtrl (2023)	0.41	27.00	**0.13**
AdapEdit (Ours)	0.29	**27.08**	**0.13**

Table 1: The quantitative evaluation results.

inspired by (Tumanyan et al. 2023), we introduce a hyper-parameter λ_S to weight $\mathbf{M}_t^{\boldsymbol{c}}$ and $\mathbf{M}_t^{\boldsymbol{c}^*}$ and subsequently utilize an interpolation operation to integrate them as:

$$\mathcal{C}(\mathbf{M}_t^{\boldsymbol{c}}, \mathbf{M}_t^{\boldsymbol{c}^*}) = \lambda_S[\mathbf{S}_{[\mathtt{V}]} \cdot \mathbf{M}_t^{\boldsymbol{c}^*} + (1 - \mathbf{S}_{[\mathtt{V}]}) \cdot \mathbf{M}_t^{\boldsymbol{c}}] + (1 - \lambda_S)\mathbf{M}_t^{\boldsymbol{c}}. \quad (8)$$

where after normalizing the $\mathbf{E}_{\boldsymbol{e}_k^*}$ (filtered by α_m) and $\mathbf{E}_{[\mathtt{V}]}$, the similarity matrix $\mathbf{S}_{[\mathtt{V}]}$ is calculated as:

$$\mathbf{S}_{[\mathtt{V}]} = \lambda_{\mathbf{S}_{[\mathtt{V}]}} \mathbf{E}_{\mathrm{e}_k^*} \times \mathbf{E}_{[\mathtt{V}]}. \quad (9)$$

Similar to $\lambda_{\tau_{\boldsymbol{c}^*}}$, we utilize $\lambda_{\mathbf{S}_{[\mathtt{V}]}}$ to adaptively adjust the range of values in $\mathbf{S}_{[\mathtt{V}]}$. In sum, based on two modules FWT and DPS, AdapEdit can perform continuity-sensitive editings with the spatio-temporal guidance scales $\tau_{\boldsymbol{c}^*}$ and $\mathbf{S}_{[\mathtt{V}]}$ in a training-free manner.

4 Experiments

4.1 Experimental Setup

Implementation Details We adopt the Stable Diffusion v1.4 version (SD-v1.4) (Rombach et al. 2022) as backbone for all experiments. Moreover, we follow (Song, Meng, and Ermon 2020) to set the default denoising iterations as 50. The classifier-free guidance scale is set to 7.5 and the maximum length of text instructions is set to 77. All experiments are conducted on 2 NVIDIA RTX3090 GPUs with PyTorch.

Baselines We compare AdapEdit versus the two groups of SOTA image editing methods. 1) The first group is real-image manipulation methods, which includes SDEdit (Meng et al. 2021) and InstructPix2Pix (Brooks, Holynski, and Efros 2023). In the two baselines, the target text-condition $\boldsymbol{c}^*$ and the original real image $\mathbf{x}$ server as common inputs. To be consistent with them, we utilize prompt-to-prompt (Hertz et al. 2022) to generate the original image for comparison. 2) The second group is text-generated methods, which includes Prompt-to-Prompt (Hertz et al. 2022) and MasaCtrl (Cao et al. 2023). Both of them use the original text-condition $\boldsymbol{c}$ and the target text-condition $\boldsymbol{c}^*$ as inputs, to generate the original image $\mathbf{x}$ and edited image $\mathbf{x}^*$, we keep the same settings. Note we have fixed the same random seed in our experiments for obtaining the same original images.

4.2 Qualitative Evaluation

The qualitative evaluation results are shown in Figure 5. From Figure 5, we obtain the following three observations. **1)** For the hard editing instruction from *"cake"* to *"chocolate cake"*, all the methods show excellent performance,

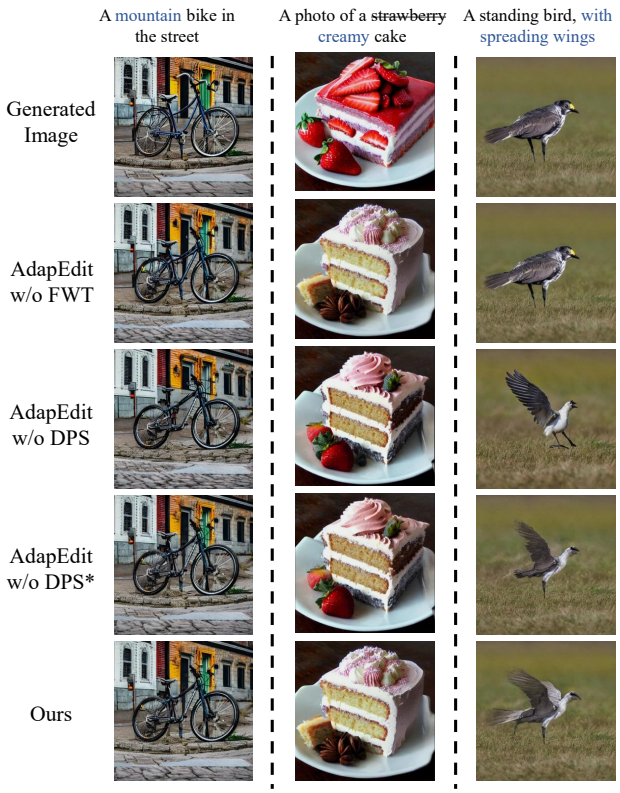

Figure 6: The ablation study of AdapEdit.

which is relatively easier to be achieved for diffusion-based methods. **2)** Further, when editing the posture of "cat" in the second line of Figure 5 and the number of "peach" in the third line of Figure 5, Prompt-to-Prompt, SDEdit, and InstructP2P all fail to accomplish the posture- or number-oriented editing task. Benefiting from the adaptive editing capability of the FWT and DPS modules, our approach effectively copes with all hard or soft editing scenes. Moreover, as shown in Figure 2, AdapEdit can also perform adaptive editing to control that the *"two apples"* have different sizes with certain reasoning capability emerging. **3)** MasaCtrl achieves excellent object editing performance but fails to retain more details from the original image. It is worth noting that both of these are well addressed by our AdapEdit, which proves its effectiveness.

4.3 Quantitative Evaluation

Following Wang et al. (2023), we further evaluate our approach in LPIPS, CLIP score and CLIP directional metrics. Specifically, we use the dataset from InstructPix2Pix (Brooks, Holynski, and Efros 2023), which contains 700 human annotated samples and 454,445 editing instructions generated by GPT-3. To avoid the potential noises exist in the machine-generated instructions, we random select 700 samples annotated by human for evaluation. From Table 1, we can find the AdapEdit achieves the highest CLIP score and CLIP directional similarity, while its LPIPS is

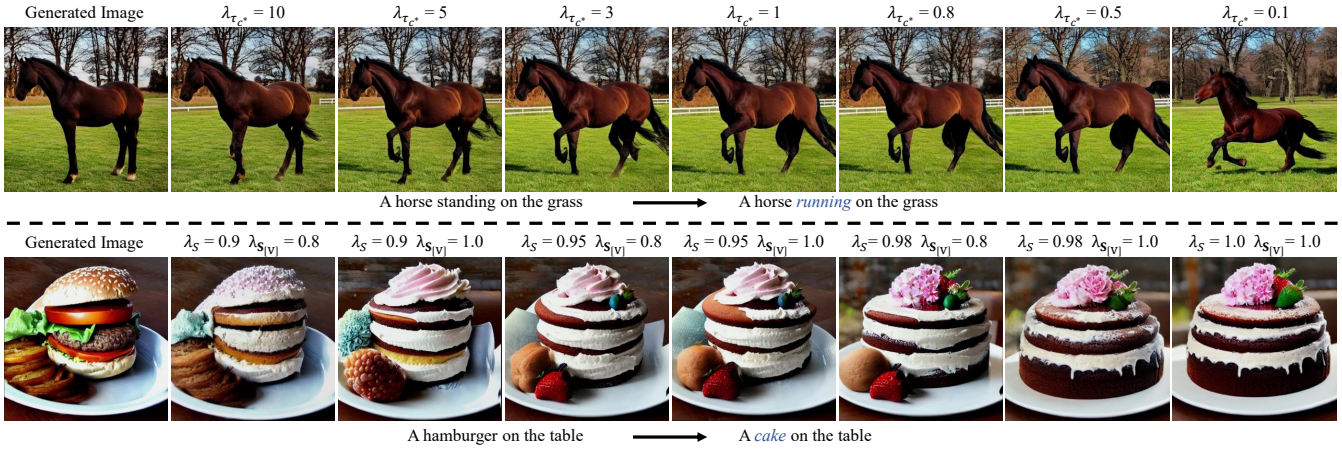

Figure 7: The hyper-parameter analysis of our proposed AdapEdit. We additionally introduce $\lambda_{\tau_{c^*}}$ and $\lambda_{\mathbf{S}_{[V]}}$ to control the range of τ_{c^*} and $\mathbf{S}_{[V]}$, and presents the results with different λ_S.

only slightly higher than the InstructPix2Pix. Unlike other methods, such as MasaCtrl, which performs well in CLIP score and CLIP directional similarity, but lacks consistency in editing results. InstructPix2Pix has a low LPIPS value but performs poorly in CLIP score. The above phenomenon demonstrates that our method strikes a good balance between the continuity and consistency of editing.

4.4 Ablation Study

In this part, we perform ablation experiments to evaluate the effectiveness of each component. We focus on four crucial settings: 1) w/o FWT denotes that we remove the flexible word-level temporal (FWT) adjustment module; 2) w/o DPS denotes that we remove the dynamic pixel-level spatial (DPS) weighting module; 3) w/o DPS* denotes that we only remove the above-mentioned pixel-level masking strategy in DPS. The visualized cases under given different settings are shown in the Figure 6. From Figure 6, we can observe that removing each component will result in a performance degradation. Particularly, when removing DPS or DPS* module in column 2, our AdapEdit will be difficult to remove the *"strawberries"* on the plate, which verfies the effectiveness of the DPS module. Further, when remove the pixel-level masking strategy in DPS (i.e., w/o DPS*), the algorithm will face a slightly inconsistency issue with unexpectedly changes to the original characteristics, which shows the proposed pixel-level masking strategy can effectively perceive the crucial regions corresponding to the target editing instruction and weaken the impacts of irrelevant regions. Moreover, when removing the FWT module in column 3, the bird standing on the grass will be difficult to be edit to *"spread wings"*, which proves the effectiveness of our proposed FWT module in facing continuity-sensitive editing tasks.

4.5 Hyper-parameter Analysis

To analyze the impacts of temporal scales τ_{c^*} and spatial scales $\mathbf{S}_{[V]}$,we conduct the hyper-parameter experiments on $\lambda_{\tau_{c^*}}$, $\lambda_{\mathbf{S}_{[V]}}$, and λ_S. Note λ_S is spatial interpolation weight. The exprimental results are presented in the Figure 7. As can be observed in the top row of Figure 7, the temporal hyper-parameter $\lambda_{\tau_{c^*}}$ reveals the temporal continuity-editing process. Specifically, with the $\lambda_{\tau_{c^*}}$ gradually decreases, the posture and movement of the horse are gradually becoming semantically close to the *"running horse"*. Further, as illustrated by second and third edited images in the bottom row of Figure 7, we can observe that when keeping $\lambda_S = 0.9$ unchanged and only vary $\lambda_{\mathbf{S}_{[V]}}$ from 0.8 to 1.0, the *"hamburger"* on the table shows significant variations with the cream on the top of *"cake"* has emerged. Moreover, when the λ_S reaches 0.98 and $\lambda_{\mathbf{S}_{[V]}}$ reaches 1.0, the *"hamburger"* shown in the picture has been significantly edited to *"cake"*, which verifies the effectiveness of adaptive spatial editing and indicates it is necessary to choose appropriate hyper-parameters for controllable spatio-temporal guidance.

5 Conclusion

In this work, we first propose to focus on continuity-sensitive "soft editing" tasks and based on the task, we further propose a spatio-temporal guided adaptive editing algorithm, AdapEdit for short. Specifically, the algorithm mainly includes two elaborately designed modules, flexible word-level temporal (FWT) adjustment module and dynamic pixel-level spatial (DPS) weighting module, to achieve spatio-temporal guided adaptive editing. Based on the promising insight that adopting variable guidance degrees into forward diffusion process would facilitate fine-grained image editing at the pixel-level, we adopt a soft-attention strategy to achieve the challenging continuity-sensitive "soft editing" tasks. Moreover, we additional introduce a spatial interpolation weight for adaptively retaining the characteristics of the original images. Experiments on the real image editing dataset demonstrate the effectiveness and superior performance of our AdapEdit in both qualitative and quantitative evaluation.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2022ZD0160603), the Project funded by China Postdoctoral Science Foundation (No. 2023M741950), and the Nationally Funded Postdoctoral Researcher Program (GZB20230347).

References

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. *arXiv preprint arXiv:2304.08465*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in NeurIPS*, 34: 8780–8794.

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35: 16890–16902.

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*.

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

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*.

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.

Kim, D.; Na, B.; Kwon, S. J.; Lee, D.; Kang, W.; and Moon, I.-C. 2022. Maximum Likelihood Training of Implicit Nonlinear Diffusion Models. *arXiv preprint arXiv:2205.13699*.

Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational diffusion models. *Advances in NeurIPS*, 34: 21696–21707.

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

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

Ma, Z.; Li, J.; Zhou, B.; et al. 2023b. LMD: Faster Image Reconstruction with Latent Masking Diffusion. *arXiv preprint arXiv:2312.07971*.

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. In *International Conference on Learning Representations*.

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.

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

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *ICML*, 8162–8171.

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*.

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.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

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. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*.

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

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wallace, B.; Gokul, A.; and Naik, N. 2023. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22532–22541.

Wang, Q.; Zhang, B.; Birsak, M.; and Wonka, P. 2023. InstructEdit: Improving Automatic Masks for Diffusion-based Image Editing With User Instructions. *arXiv preprint arXiv:2305.18047*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.