LayerDiffusion: Layered Controlled Image Editing with Diffusion Models

Pengzhi Li
Tsinghua University
China
lpz21@mails.tsinghua.edu.cn

Qinxuan Huang
Tsinghua University
China
hqx21@mails.tsinghua.edu.cn

Yikang Ding
Tsinghua University
China
dyk20@mails.tsinghua.edu.cn

Zhiheng Li*
Tsinghua University
China
zhhli@mail.tsinghua.edu.cn

ABSTRACT

Text-guided image editing has recently experienced rapid development. However, simultaneously performing multiple editing actions on a single image, such as background replacement and specific subject attribute changes, while maintaining consistency between the subject and the background remains challenging. In this paper, we propose *LayerDiffusion*, a semantic-based layered controlled image editing method. Our method enables non-rigid editing and attribute modification of specific subjects while preserving their unique characteristics and seamlessly integrating them into new backgrounds. We leverage a large-scale text-to-image model and employ a layered controlled optimization strategy combined with layered diffusion training. During the diffusion process, an iterative guidance strategy is used to generate a final image that aligns with the textual description. Experimental results demonstrate the effectiveness of our method in generating highly coherent images that closely align with the given textual description. The edited images maintain a high similarity to the features of the input image and surpass the performance of current leading image editing methods. *LayerDiffusion* opens up new possibilities for controllable image editing.

CCS CONCEPTS

• **Computing methodologies** → Image manipulation.

KEYWORDS

image editing, diffusion model

ACM Reference Format:
Pengzhi Li, Qinxuan Huang, Yikang Ding, and Zhiheng Li. 2023. LayerDiffusion: Layered Controlled Image Editing with Diffusion Models. In *SIGGRAPH Asia 2023 Technical Communications (SA Technical Communications '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3610543.3626172

1 INTRODUCTION

Given a single image of your pet, it can be imagined embarking on a worldwide journey and performing specific actions in any location. Generating such an image is a challenging and fascinating task in image editing. It entails preserving the specific subject's unique characteristics in new backgrounds and ensuring their seamless integration into the scene, harmoniously and naturally, while simultaneously accommodating multiple editing actions.

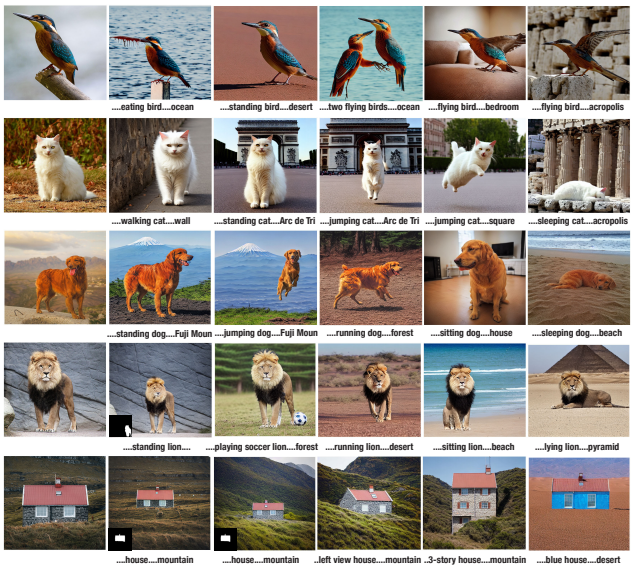

Figure 1: Given a complex text description, the original image (left) is capable of performing multiple editing actions and maintaining similar characteristics of a specific subject.

Recently, significant progress has been made in the development of deep learning-based large-scale text-to-image models [Rombach et al. 2022]. These models can generate high-quality synthetic images based on text prompts, enabling text-guided image editing and producing impressive results. As a result, numerous text-based image editing methods [Hertz et al. 2022; Tumanyan et al. 2022] have emerged and evolved. However, such models cannot mimic specific subject characteristics. Even with the most detailed textual descriptions of an object, they may generate instances with different appearances and still struggle to maintain background consistency. Thus, the current leading image editing methods encounter several challenges, including rigid editing limited to specific domain images [Hertz et al. 2022; Patashnik et al. 2021], the inability to simultaneously edit both the background and specific subjects, and the requirement for additional auxiliary input information [Avrahami et al. 2022]. These issues hinder the advancement of controllable image editing.

In this paper, we propose a semantic-based layered controlled image editing method, which we call *LayerDiffusion*, to alleviate these

*Corresponding author.

SA Technical Communications '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0314-0/23/12.
https://doi.org/10.1145/3610543.3626172

issues. By simply inputting textual descriptions of multiple editing actions, along with the target image and a reference image, we can perform non-rigid editing and attribute modification of specific subjects, generating images consistent with the textual descriptions while maintaining the consistency of the specific subject and background features with the input image. As shown in Figure 1, we can make a cat jump in a square or a dog sit in the house or modify their shapes and attributes in the original scene.

To implement our method, we leverage the robust and high-quality image generation capabilities of a large-scale text-to-image model [Rombach et al. 2022].

Our method comprises a well-defined sequence of steps. Initially, we utilize a mask to eliminate interference from foreground objects effectively. Subsequently, we apply a layered controlled optimization strategy to optimize the text embeddings acquired from the text encoders [Raffel et al. 2020], following the segmentation of the target text. This process aims to generate image backgrounds that exhibit a remarkable similarity to the reference images. Next, we employ a layered diffusion training strategy to fine-tune the model, thereby augmenting its ability to preserve the similarity between the specific subjects, backgrounds, and input images. Finally, during the diffusion process with the fine-tuned model, we adopt an iterative guidance strategy, where a highly constrained text embedding is iteratively employed to denoise the images. Consequently, this generates a final image aligning with the textual description.

We emphasize the contributions of each component in our method through ablation studies and compare our approach with other relevant image editing methods [Kawar et al. 2022; Meng et al. 2021; Tumanyan et al. 2022], clearly demonstrating superior editing quality. We summarize our main contributions as follows:

- We propose *LayerDiffusion*. To the best of our knowledge, this is the first image editing method that enables simultaneous editing of specific subjects and backgrounds using a single input image.
- We introduce a novel layered diffusion training framework that enables arbitrary and controllable editing of specific subjects and backgrounds.
- Experimental results demonstrate that our method generates images with highly similar features to the input images.

2 METHOD

As illustrated in Figure 2, our method begins by separating the background. We apply a layered controlled optimization strategy to refine the segmentation text embeddings acquired from the text encoders, which come from the target text. Then we identify the optimal text embedding that aligns with the desired target background in proximity to the target text embedding. Subsequently, we employ a layered diffusion strategy to fine-tune the diffusion model. This approach enhances the model's capability to maintain similarity between specific subjects, backgrounds, and input images, allowing for finer control and precision in image editing through parameter adjustments. During the inference stage, we utilize an iterative guidance strategy to directly generate images that align with the multiple image editing actions described in the input text without the text embedding interpolation. Each step of the process is outlined in detail below.

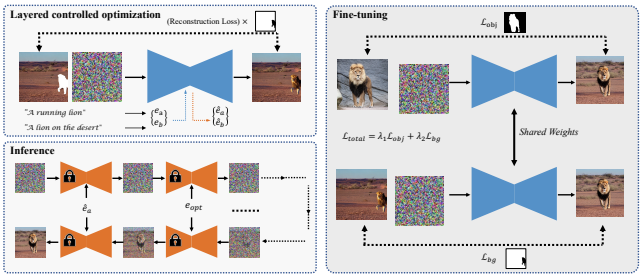

Figure 2: Our method utilizes a layered controlled optimization strategy to refine text embeddings and a layered diffusion strategy to fine-tune the diffusion model. During inference, an iterative guidance strategy is employed to directly generate images aligning with the multiple editing actions described in the input text.

2.1 Layered controlled optimization

Due to the potential interference of multiple text descriptions, optimizing text embeddings can be unstable during image editing. As a result, previous methods for image editing have often struggled to effectively modify selected object property and backgrounds simultaneously.

To this end, we aim to separate the background and foreground to reduce interference between different textual information. The target text T is first fed into the Stable Diffusion model [Rombach et al. 2022] to obtain the target image O_t. Then T is decomposed into T_a and T_b, which describe object properties and background separately and sent to the text encoder [Raffel et al. 2020] to output the corresponding text embeddings $\boldsymbol{e}_a \in \mathbb{R}^{C\times N}$ and $\boldsymbol{e}_b \in \mathbb{R}^{C\times N}$, where C is the number of tokens, and N is the token embedding dimension. However, $\boldsymbol{e}_a$ and $\boldsymbol{e}_b$ are in the distant embedding space, so we cannot directly perform linear interpolation on them. To make $\boldsymbol{e}_a$ and $\boldsymbol{e}_b$ match our input image background as much as possible and be in a close embedding space, we freeze the parameters of the diffusion model and optimize $\boldsymbol{e}_a$ and $\boldsymbol{e}_b$ simultaneously using the diffusion model objective [Ho et al. 2020]. In fact, we can optimize the initial text embedding to make it closer to the target image (modify the background) space or reference image (modify object properties) space. This process is controlled by the object mask M and can be represented as follows:

$$\left[\hat{\boldsymbol{e}}_a, \hat{\boldsymbol{e}}_b\right] = \arg\min \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|M * \left(\boldsymbol{\epsilon} - f_\theta\left(\boldsymbol{x}_t, t, \left[\boldsymbol{e}_a, \boldsymbol{e}_b\right]\right)\right)\right\|^2\right], \tag{1}$$

where M is computed by Segment Anything Model (SAM) [Kirillov et al. 2023], and $\boldsymbol{x}_t$ is the noisy version of the input image, and f_θ means the forward diffusion process using pre-trained diffusion model. The optimized text embeddings make it meaningful to modify the linear interpolation weights of $\hat{\boldsymbol{e}}_a$ and $\hat{\boldsymbol{e}}_b$ as follows:

$$\boldsymbol{e}_{opt} = \alpha * \hat{\boldsymbol{e}}_a + (1-\alpha) * \hat{\boldsymbol{e}}_b, \tag{2}$$

according to the experimental analysis of text embedding interpolation in Imagic [Kawar et al. 2022], we tend to set the weight α that describes object properties to 0.7.

2.2 Model fine-tuing

We obtain new text embeddings $\boldsymbol{e}_{opt}$ by linearly interpolating multiple optimized text embeddings. Due to the limited number of optimization steps, the resulting embeddings may not lead to a consistent representation of the selected objects or background in the input image. Therefore, we propose a layered diffusion strategy to optimize model parameters while freezing the optimized text embeddings $\boldsymbol{e}_{opt}$. This enables the model to fit the desired image at optimized text embedding points. To achieve the arbitrary modification and combination of foreground object properties and backgrounds, we employ SAM [Kirillov et al. 2023] to derive M_s (object) and $1 - M_s$ (background) from O_t and subsequently obtain M_r (object) and $1 - M_r$ (background) from the reference image O_r. The aforementioned can be achieved by optimizing the following equations:

$$\mathcal{L}_{obj} = \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{\epsilon} \sim \mathcal{N}(0,\boldsymbol{I})} \left[\left\| M_s * \left(\boldsymbol{\epsilon} - f_\theta \left(\boldsymbol{x}_t, t, e_{opt} \right) \right) \right\|^2 \right], \quad (3)$$

$$\mathcal{L}_{bg} = \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{\epsilon} \sim \mathcal{N}(0,\boldsymbol{I})} \left[\left\| (1 - M_r) * \left(\boldsymbol{\epsilon} - f_\theta \left(\boldsymbol{x}_t, t, e_{opt} \right) \right) \right\|^2 \right], \quad (4)$$

The total loss can be represented as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{obj} + \lambda_2 \mathcal{L}_{bg}, \quad (5)$$

This approach enables us to manipulate the foreground object and background independently, allowing for precise control over the final output image.

2.3 Iterative guidance strategy

We first represent the diffusion process of a pre-trained model as follows:

$$I_T, I_{T-1}, \ldots, I_0.\ I_{t-1} = D(I_t | y) \quad (6)$$

where D represents an update process: $\mathcal{I} \times C \rightarrow \mathcal{I}$, $\mathcal{I} \in \mathbb{R}^{H \times W \times C}$ is the image space, and C is the condition space, and $y \in C$ is a text prompt. From T to 0, I_T gradually changes from a Gaussian noise distribution to a desired image by y. Nonetheless, due to the significant gap between the initial image and the desired image in our task, applying the base generative diffusion process with fine-tuned models under condition $y(i.e., e_{opt})$ may still result in failures in modifying object properties in sometimes, such as modifications of actions.

This issue in image editing is due to the lack of a strong constraint corresponding to the text description of the edited attributes in the diffusion process. The network bias leads the diffusion model to favor object properties in the initial image. Furthermore, we observe that the formed clear target image layout occurs in the early stages of the diffusion process. To address this, we strengthen the object properties by utilizing the decomposed $\hat{\boldsymbol{e}}_{\boldsymbol{a}}$ at the beginning and end of the diffusion process (half of the total number of steps T and it is even) with $I_{t-1} = D(I_t | \hat{\boldsymbol{e}}_{\boldsymbol{a}})$. Subsequently, we utilize $I_{t-1} = D(I_t | \boldsymbol{e}_{\boldsymbol{opt}})$ to iterate the diffusion process.

2.4 Implementation details

We adopt the Stable Diffusion v1.4 model [Rombach et al. 2022] as the baseline. We first fine-tune the text embeddings with a learning rate of 1e-3 and perform 500 steps in most of our experiments. Subsequently, we fine-tune the diffusion model by using a learning rate of 2e-6 and executing 250 steps. We employ an iterative guidance strategy throughout the diffusion process, starting from random noise. This iterative process consist of 50 iterations by default, resulting in more refined results. For one image, it takes about 2 minutes to run on a single NVIDIA A100 GPU.

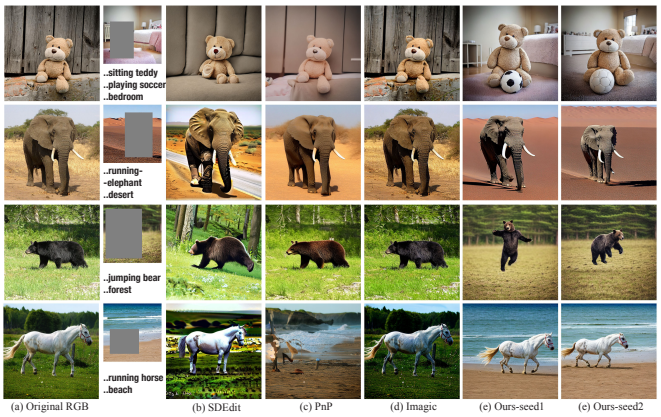

Figure 3: We present several edited images and compare them with similar image editing algorithms, such as SDEdit [Meng et al. 2021], Imagic [Kawar et al. 2022], and PnP [Tumanyan et al. 2022]. Our method generates the best results.

Table 1: Quantitative results with different settings. We report the CLIP score [Hessel et al. 2021] over 300 images.

	Settings			CLIP score
	$\mathcal{L}_{obj}$	$\mathcal{L}_{bg}$		
(a)	×			0.28
(b)		×		0.32
(c)	$\lambda_1 = 1$			0.29
(d)	$\lambda_1 = 3$			0.32
(e)	$\lambda_1 = 1$	$\lambda_2 = 3$		0.28
(f)				**0.35**

	Settings				CLIP score
	L-e-o	Fine-tune	I-g	α	
(g)	×				0.33
(h)		×			0.32
(i)			×		0.29
(j)				= 1	0.32
(k)				= 0	0.29
(f)					**0.35**

3 EXPERIMENTS

3.1 Qualitative Evaluation

We extensively evaluate our approach using images from various domains and categories. We employ a layered editing strategy to ensure robustness and controllability in the editing process. This approach enables multiple editing actions simultaneously on the images, demonstrating excellent editing controllability. By employing our layered diffusion strategy, we can generate images that closely match the provided text descriptions while preserving the critical attributes of the original image in most cases.

In Figure 1, we present some edited images. These images preserve the distinct characteristics of the input image, and they are altered based on text prompts to accommodate a range of editing actions that go beyond mere background replacement and property modification. Our method can execute precise editing actions on the images by leveraging reference background or foreground objects. For instance, we can alter foreground objects based on reference foreground object maps or implement background modifications guided by reference background maps. More results can be found in the supplementary material.

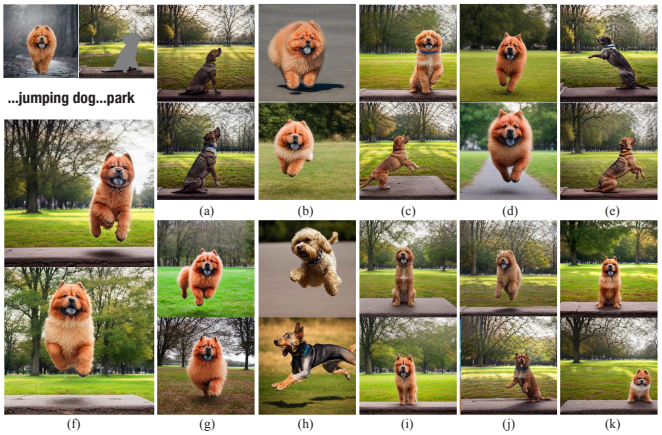

Figure 4: We present the edited images with different settings. For each setting, we show two generated images using different random seeds. (f) illustrates the final edited results.

3.2 Comparisons

We primarily compare our proposed image editing method with previous text prompt methods, such as SDEdit [Meng et al. 2021], Imagic [Kawar et al. 2022], and PnP [Tumanyan et al. 2022]. It is worth noting that Imagic [Kawar et al. 2022] necessitates fine-tuning of both the network and text embeddings, while our method adopts a similar fine-tuning approach.

As shown in Figure 3, non-rigid editings, such as jumping and rotation, have significant challenges in image editing tasks. This complexity leads to the failure of both PnP [Tumanyan et al. 2022] and SDEdit [Meng et al. 2021] in performing editing actions. In contrast, our approach adopts a layered strategy that allows for the simultaneous execution of multiple editing actions. As a result, our method achieves impressive results in real image editing tasks and outperforms other methods. The last two columns of Figure 3 show the edited results generated by employing different random seeds. We can also choose our reference image as long as it is close to the perspective of the original image.

3.3 Ablation Study

In this section, we present a comprehensive analysis of the three modules employed in our method. We utilize the *TEdBench* [Kawar et al. 2022] dataset and generate over 300 images using 20 different random seeds. As an auxiliary objective evaluation metric, we employ the text-image CLIP score [Hessel et al. 2021]. Furthermore, we present the specific performance of each component in Tab. 1. As mentioned previously, the CLIP score may not fully capture the suitability of our method as it primarily focuses on the alignment between images and text. For instance, the results of (b), (g), and (i) show high CLIP scores, but their object features significantly differ from the reference images.

As shown in Figure 4 and Tab. 1, (a) does not utilize $\mathcal{L}_{obj}$, resulting in a background that matches the reference image, while the properties of the foreground objects differ substantially. On the other hand, (b) demonstrates that $\mathcal{L}_{bg}$ preserves a more similar background. (c), (d), and (e) analyze the impact of different weights assigned to the two losses, which affect the similarity of the background and foreground objects. In this paper, we mostly set λ_1 to 2 and λ_2 to 1, except when λ_1 is set to 3 for smaller foreground objects. (g), (h), and (i) validate the effectiveness of each of the three modules in our method. (g) enhances the similarity of the background, (h) controls the global features, and (i) significantly increases the percentage of image generation results that satisfy the description text, rising from 43% to 81%.

4 LIMITATIONS

Dealing with fine-grained tasks is still challenging for our method while we rely on a pre-trained text-to-image diffusion model and the problem of overfitting that occurs during model fine-tuning. We need to fine-tune the model to accommodate the reference image.

5 CONCLUSION

We propose *LayerDiffusion*, a semantic-based layered image editing method that simultaneously edits specific subjects and backgrounds using a single input image. *LayerDiffusion* preserves the unique characteristics of the subjects while integrating them seamlessly into new scenes. Extensive experimentation demonstrates that our method generates images closely resembling the feature of the input images, surpassing existing approaches in editing quality and controllability. Our contributions include introducing *LayerDiffusion* as the first method for simultaneous editing of specific subjects and backgrounds. We develop a layered diffusion training framework for controllable image editing, which opens up new possibilities for text-guided image editing tasks. We may focus on preserving complex textures and facial features in the future.

REFERENCES

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2022. SpaText: Spatio-Textual Representation for Controllable Image Generation. *arXiv preprint arXiv:2211.14305* (2022).

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276* (2022).

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proc. IEEE Int. Conf. Comp. Vis.* 2085–2094.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* 10684–10695.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. *arXiv preprint arXiv:2211.12572* (2022).