Noname manuscript No.
(will be inserted by the editor)

StyleDiffusion: Prompt-Embedding Inversion for Text-Based Editing

Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang* and Jian Yang

Received: date / Accepted: date

Abstract A significant research effort is focused on exploiting the amazing capacities of pretrained diffusion models for the editing of images. They either finetune the model, or invert the image in the latent space of the pretrained model. However, they suffer from two problems: (1) Unsatisfying results for selected regions and unexpected changes in non-selected regions. (2) They require careful text prompt editing where the prompt should include all visual objects in the input image. To address this, we propose two improvements: (1) Only optimizing the input of the value linear network in the cross-attention layers is sufficiently powerful to reconstruct a real image. (2) We propose attention regularization to preserve the object-like attention maps after reconstruction and editing, enabling us to obtain accurate style editing without invoking significant structural changes. We further improve the editing technique that is used for the unconditional branch of classifier-free guidance as used by P2P (Hertz et al., 2022). Extensive experimental prompt-editing results on a variety of images demonstrate qualitatively and quantitatively that our method has superior editing capabilities compared to existing and concurrent works. See our accompanying code in Stylediffusion: `https://github.com/sen-mao/StyleDiffusion`.

*The corresponding author.

S. Li, T. Hu, Q. Hou, Y. Wang and J. Yang are CS, Nankai University, Tianjin, China. E-mail: senmaonk@mail.nankai.edu.cn, hutaihang00@gmail.com, {houqb,yaxing, csjyang}@nankai.edu.cn.
J. van de Weijer is with the Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona 08193, Spain. E-mail: joost@cvc.uab.es.
F. Shahbaz Khan is Mohamed bin Zayed University of Artificial Intelligence (UAE), and Linkoping University (Sweden). E-mail: fahad.khan@liu.se

1 Introduction

Text-based deep generative models have achieved extensive adoption in the field of image synthesis. Notably, GANs (Patashnik et al., 2021; Gal et al., 2021; Kang et al., 2023; Sauer et al., 2023), Diffusion models (Saharia et al., 2022; Ramesh et al., 2022; Gafni et al., 2022), autoregressive models (Yu et al., 2022), and their hybrid counterparts have been prominently utilized in this domain.

Diffusion models (Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2021) have made remarkable progress due to their exceptional realism and diversity. It has seen rapid applications in other domains such as video generation (Khachatryan et al., 2023; Wu et al., 2022; Zhou et al., 2022), 3D generation (Poole et al., 2022; Lin et al., 2023; Wang et al., 2023) and speech synthesis (Jeong et al., 2021; Huang et al., 2022; Koizumi et al., 2022). In this work, we focus on Stable Diffusion (SD) models for real image editing. Researchers basically perform real image editing with two steps: *projection* and *manipulation*. The former aims to either adapt the weights of the model (Song et al., 2020; Liu et al., 2023; Kim et al., 2022; Xiao et al., 2023) or project given images into the latent code or embedding space of SD models (Mokady et al., 2022; Gal et al., 2022; Avrahami et al., 2023; Han et al., 2023a).

The latter aims to edit the latent code or embedding to further manipulate real images (Kawar et al., 2022; Meng et al., 2021; Cao et al., 2023; Zhang and Agrawala, 2023; Couairon et al., 2022).

In the first *projection* step, some works finetune the whole model (Kawar et al., 2022; Valevski et al., 2022; Ruiz et al., 2022; Kim et al., 2022; Gal et al., 2023; Xiao et al., 2023) or partial weights of the pretrained model (Kumari et al., 2022; Xie et al., 2023). Yet, fine-

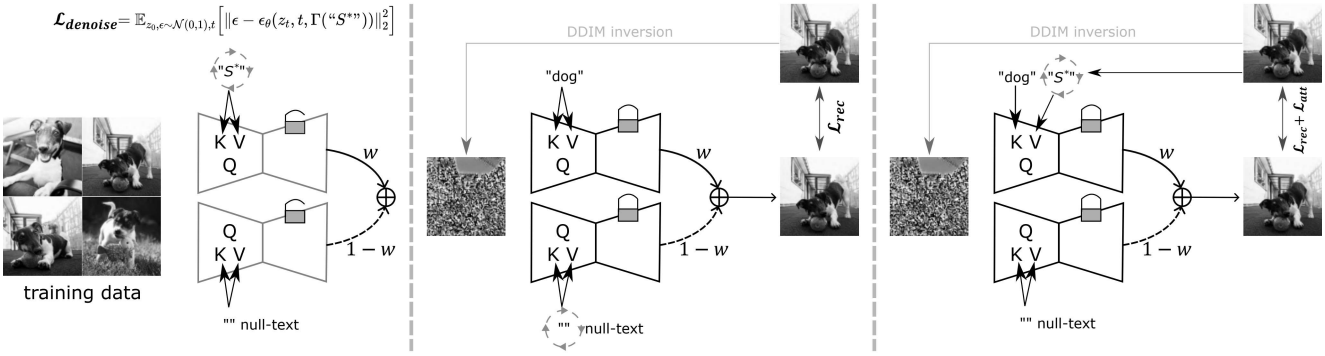

Fig. 1: Three different optimization methods for inverting real image(s). (Left) Some works invert the image(s) into a new textual embedding "S^*" by finetuning the pretrained diffusion model (Kawar et al., 2022; Valevski et al., 2022; Ruiz et al., 2022) or freezing the model (Kumari et al., 2022) and applying a denoise loss $\mathcal{L}_{denoise}$. Here, w is the classifier-free guidance parameter. These methods require a few training images. (Middle) Null-text inversion (Mokady et al., 2022) optimizes the null-text embedding with the reconstruction loss. (Right) we propose Stylediffusion which maps the real image to the input embedding of the *value* of the cross-attention, which enables us to obtain accurate style editing without invoking significant structural changes.

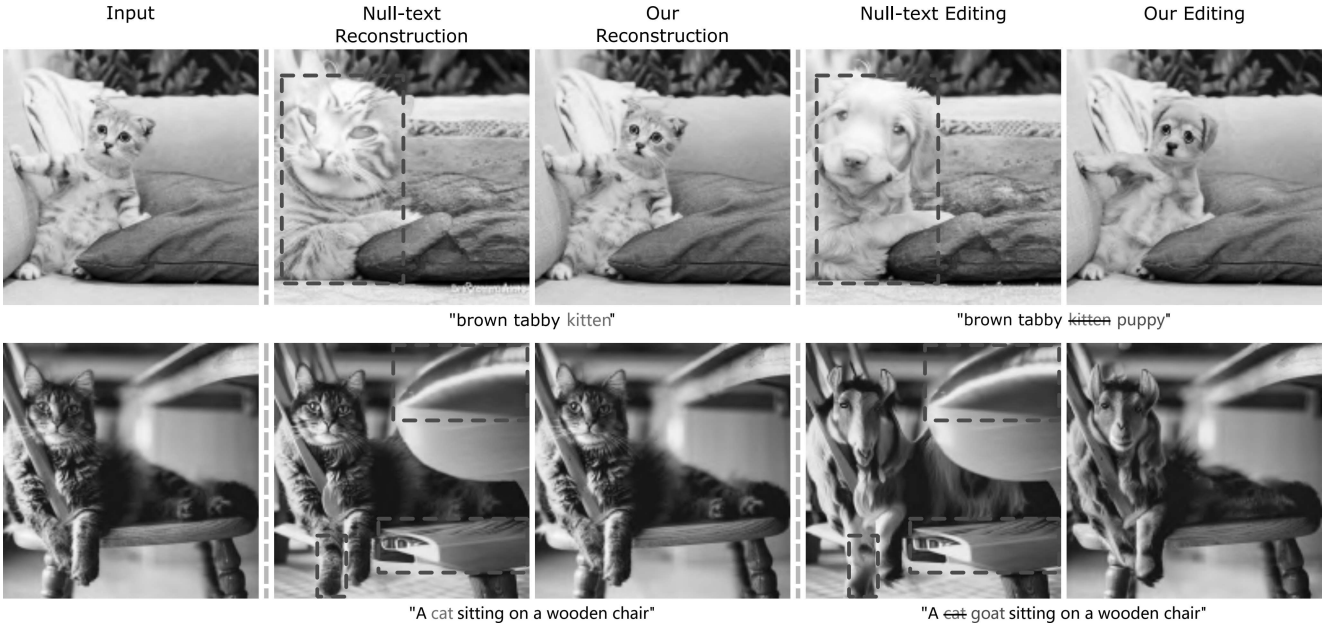

Fig. 2: Our method takes as an input a real image (leftmost column) and an associated caption. Here we demonstrate more accurate reconstruction and editing capabilities compared to Null-text (Mokady et al., 2022). We manipulate the inverted image using the editing technique P2P.

tuning both the entire or part of the generative model with only a few real images suffers from both the cumbersome tuning of the model's weights and catastrophic forgetting (Kumari et al., 2022; Wu et al., 2018; Xie et al., 2023). Other works on *projection* attempt to learn a new embedding vector which represents given real images (keeping the SD model frozen) (Ho and Salimans, 2021; Gal et al., 2022; Avrahami et al., 2023; Han et al., 2023a; Tewel et al., 2023; Zhang et al., 2023a; Dong et al., 2023). They focus on optimizing conditional or unconditional inputs of the cross-attention layers of the classifier-free diffusion model (Ho and Salimans, 2021). Textual Inversion (Gal et al., 2022) uses the denoising loss to optimize the textual embedding of the conditional branch given a few content-similar images. Null-text optimization (Mokady et al., 2022) firstly inverts the real image into a series of timestep-related latent codes, then leverages a reconstruction loss to learn the null-text embedding of the unconditional branch (see Fig. 1 (middle)). However, these methods suffer from the following challenges: (I) They lead to unsatisfying results for selected regions, and unexpected changes

Fig. 3: Null-text (Mokady et al., 2022) editing results of the real image with different prompts. A satisfactory result requires a carefully selected prompt.

in non-selected regions, both when reconstruction and editing (see Fig. 2 (*Null-text*)). (II) They require a user to provide an accurate text prompt that describes every visual object, and the relationship between them in the input image (see Fig. 3). Finally, Pix2pix-zero (Parmar et al., 2023) requires the textual embedding directions (e.g, cat → dog in Fig. 9 (up, the fourth column)) with thousands of sentences with GPT-3 (Brown et al., 2020) before editing, which lacks scalability and flexibility.

To overcome the above-mentioned challenges, we analyze the role of the attention mechanism (and specifically the roles of keys, queries and values) in the diffusion process. This leads to the observation that the key dominates the output image structure (where) (Hertz et al., 2022), whereas the value determines the object style (what). We perform an effective experiment to demonstrate that the value determines the object style (what). As shown in Fig. 4 (top), we generate two sets of images with prompt embeddings $\mathbf{c}_1$ and $\mathbf{c}_2$.

We use two different embeddings for the input of both key and value in the same attention layer. When swapping the input of the keys and fixing the one of the values, we observe that the content of generated images are similar, see Fig. 4 (middle). For example, The images of Fig. 4 (middle) are similar to the ones of Fig. 4 (top). When exchanging the input of the values and fixing the one of the keys, we find that the content swaps while preserving much of the structure, see Fig. 4 (bottom). For example, the images of the last row of Fig. 4 (bottom) have similar semantic information with the ones of the first row of Fig. 4 (top). It should be noted that their latent codes are all shared [1], so the structure of the results in Fig. 4 (middle) does not change significantly. This experiment indicates that the value determines the object's style (what).

Therefore, to improve the projection of a real image, we introduce **Stylediffusion** which maps a real image to the input embedding for the value computation (we refer to this embedding as **prompt-embedding**), which enables us to obtain accurate style editing without invoking significant structural changes. We propose to map the real image to the input of the *value* linear layer in the cross-attention layers (Bahdanau et al., 2014; Vaswani et al., 2017) providing freedom to edit effectively the real image in the manipulation step.

We take the given textual embedding as the input of the *key* linear layer, which is frozen (see Fig. 1 (right)). Using frozen embedding contributes to preserving the well-learned attention map from DDIM inversion, which guarantees the initial editability of the inverted image.

We observe that the system often outputs unsatisfactory reconstruction results (greatly adjusting the input image structure) (see Fig. 2 (first row, second columns) and Fig. 14 (first row, second columns)) due to locally less accurate attention maps (see Fig. 14 (first row, and third columns)).

Hence, to further improve our method, we propose an attention regularization to obtain more precise reconstruction and editing capabilities.

In the second manipulation step, researchers propose a series of outstanding techniques (Kawar et al., 2022; Meng et al., 2021; Cao et al., 2023; Zhang and Agrawala, 2023; Couairon et al., 2022; Mou et al., 2023; Jia et al., 2023; Zhang et al., 2023b; Qiu et al., 2023; Levin and Fried, 2023). Among them, P2P (Hertz et al.,

[1] The first row of top, middle, and bottom shares a common latent code, and the second row also shares a common latent code.

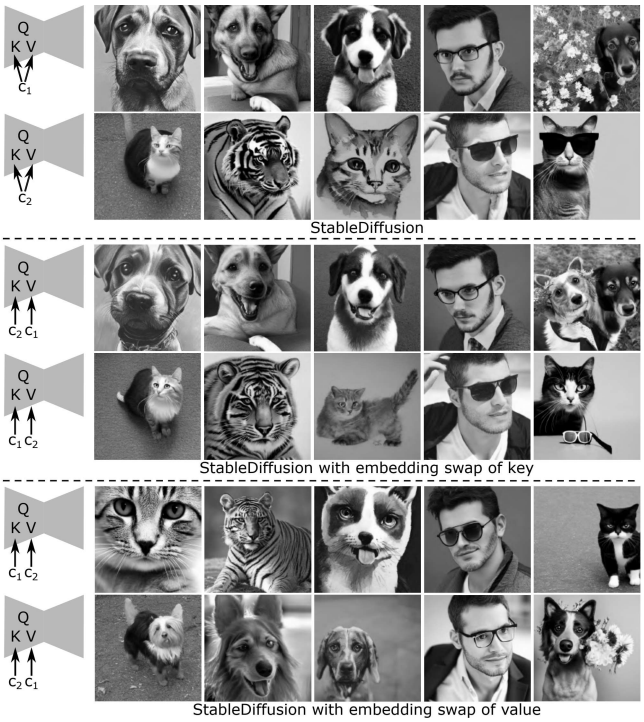

Fig. 4: (Top) from second to last columns, the prompts corresponding to both $\mathbf{c}_1$ and $\mathbf{c}_2$ are [*"Dog"*, *"Dog"*, *"Dog"*, *"A man in glasses"*, *"A dog holding flowers"*] and [*"Cat"*, *"Tiger"*, *"Watercolor drawing of a cat"*, *"A man in sunglasses"*, *"A cat wearing sunglasses"*], respectively. (Middle) we swap the input embedding of the key and fix the one of the value. This result shows that the generated images of both the third and fourth rows are similar to the ones of both the first and second rows, respectively. (Bottom) we fix the key and swap the one of the value. This experiment indicates that the value determines the object style (what).

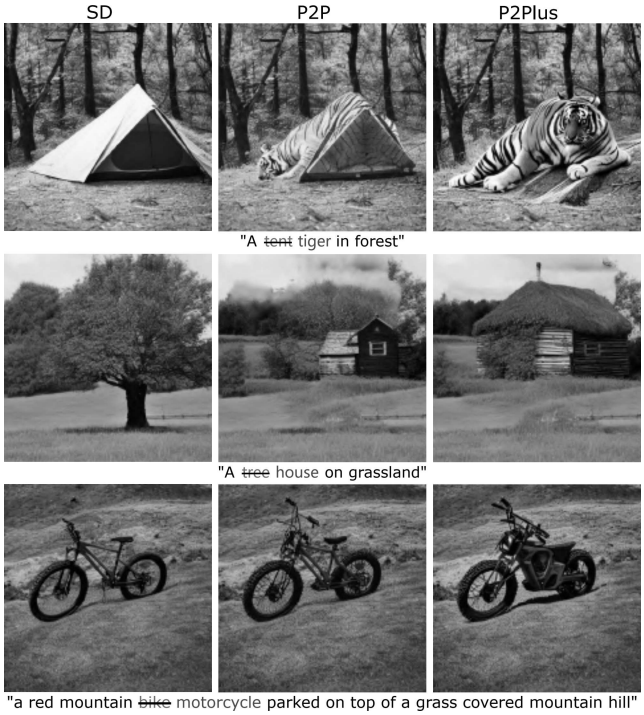

Fig. 5: Comparison of both P2P and P2Plus. P2P fails when the desired editing changes require large structural changes (e.g., tent, tree, and bike). By reducing the dependence on the source prompt (by also involving the unconditional branch), P2Plus can better handle such cases.

2022) is one of the most widely used image editing methods.

However, P2P only operates on the conditional branch, and ignores the unconditional branch. This leads to less accurate editing capabilities for some cases, especially where the structural changes before and after editing

are relatively large (e.g., "...tree..."→"...house..." in Fig. 5). To address this problem, we need to reduce the dependence of the structure on the source prompt and provide more freedom to generate the structure following the target prompt. Since the unconditional branch allows us to edit out concepts (Armandpour et al., 2023; Tumanyan et al., 2023). Thus, we propose to further perform the self-attention map exchange in the unconditional branch based on P2P (called *P2Plus*), as well as in the conditional branch like P2P (Hertz et al., 2022). This technique enables us to obtain more accurate editing capabilities (see Fig. 5 (the third column)). We build our method on SD models (Rombach et al., 2021) and experiment on a variety of images and several ways of prompt editing.

2 Related work

2.1 Transfer learning for diffusion models

A series of recent works investigated knowledge transfer on diffusion models (Song et al., 2020; Liu et al., 2023; Mokady et al., 2022; Gal et al., 2022; Avrahami et al., 2023; Kwon et al., 2022; Kim et al., 2022; Gal et al., 2023; Xiao et al., 2023; Han et al., 2023a; Tewel et al., 2023; Kumari et al., 2023; Zhang et al., 2023a; Dong et al., 2023) with one or a few images. Recent work (Kawar et al., 2022; Meng et al., 2021; Cao et al., 2023; Zhang and Agrawala, 2023; Couairon et al., 2022; Mou et al., 2023; Jia et al., 2023; Zhang et al., 2023b; Qiu et al., 2023; Levin and Fried, 2023; Chen et al., 2023) either finetune the pretrained model or invert the image in the latent space of the pretrained model. Dreambooth (Ruiz et al., 2022) shows

that training a diffusion model on a small data set (of $3 \sim 5$ images) largely benefits from a pre-trained diffusion model, preserving the textual editing capability. Similarly, Imagic (Kawar et al., 2022) and UniTune (Valevski et al., 2022) rely on the weights of the interpolation or the classifier-free guidance at the inference stage, except when finetuning the diffusion model during training. Kumari et al. (Kumari et al., 2022) study only updating part of the parameters of the pre-trained model, namely the key and value mapping from text to latent features in the cross-attention layers. However, updating the diffusion model unavoidably loses the text editing capability of the pre-trained diffusion model. In this paper, we focus on real image editing with a frozen text-guilded diffusion model.

2.2 GAN inversion

Image inversion aims to project a given real image into the latent space, allowing users to further manipulate the image. There exist several approaches (Bermano et al., 2022; Creswell and Bharath, 2018; Goetschalckx et al., 2019; Jahanian et al., 2020; Lipton and Tripathi, 2017; Xia et al., 2021; Yeh et al., 2017; Zhu et al., 2016) which focus on image manipulation based on pre-trained GANs, following literately optimization of the latent representation to restructure the target image. Given a target semantic attribute, they aim to manipulate the output image of a pretrained GAN. Several other methods (Abdal et al., 2019; Zhu et al., 2020) reverse a given image into the input latent space of a pretrained GAN (e.g., StyleGAN), and restructure the target image by optimization of the latent representation. They mainly consist of fixing the generator Abdal et al. (2019, 2020); Richardson et al. (2020); Tov et al. (2021) and updating the generator Alaluf et al. (2021); Roich et al. (2022).

2.3 Diffusion model inversion

Diffusion-based inversion can be performed naively by optimizing the latent representation. Dhariwal and Nichol (2021) show that a given real image can be reconstructed by DDIM sampling (Song et al., 2020). DDIM provides a good starting point to synthesize a given real image. Several works (Avrahami et al., 2022a,b; Nichol et al., 2021) assume that the user provides a mask to restrict the region in which the changes are applied, achieving both meaningful editing and background preservation. P2P (Hertz et al., 2022) proposes a mask-free editing method. However, it leads to unexpected results when editing the real image (Mokady et al., 2022). Recent work investigates the text embedding of the conditional input (Gal et al., 2022), or the null-text optimization of the unconditional input (i.e., Null-text inversion (Mokady et al., 2022)). Although having the editing capability by combining the new prompts, they suffer from the following challenges: (I) They lead to unsatisfying results for the selected regions, and unexpected changes in non-selected regions. (II) They require careful text prompt editing where the prompt should include all visual objects in the input image.

Concurrent work (Parmar et al., 2023) proposes pix2pix-zero, also aiming to increase more accurate editing capabilities of the real image. However, it firstly needs to compute the textual embedding direction with thousand sentences in advance.

3 Method

Problem setting. DDIM inversion proposes an inversion scheme for unconditional diffusion models. However, this method fails when applied to text-guided diffusion models. This was observed by Mokady et al. (Mokady et al., 2022) and they propose Null-text inversion to address this problem. However, their methods has some drawbacks: (1) unsatisfying results for selected regions and unexpected changes in non-selected regions. (2) they require careful text prompt editing where the prompt should include all visual objects in the input image.

Therefore, our goal is to obtain a more accurate editing capability based on an accurate reconstruction of the real image $\mathbf{x}$ guided by the source prompt $\mathbf{p}^{src}$. Our method, called StyleDiffusion, is based on the observation that the *keys* of the cross-attention layer dominate the output image structure (where), whereas the *values* determine the object style (what). After projecting faithfully the real image, we propose P2Plus, which is an improved version of P2P (Hertz et al., 2022).

Next, we introduce SD models in Sec. 3.1, followed by the proposed StyleDiffusion in Sec. 3.2 and P2Plus in Sec. 3.3. A general overview is provided in Fig. 6.

3.1 Preliminary: Diffusion Model

Generally, diffusion models optimize a UNet-based denoiser network ϵ_θ to predict Gaussian noise ϵ, following the objective:

$$\min_\theta E_{\mathbf{z}_0,\epsilon\sim N(0,I),t\sim[1,T]} \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2 , \quad (1)$$

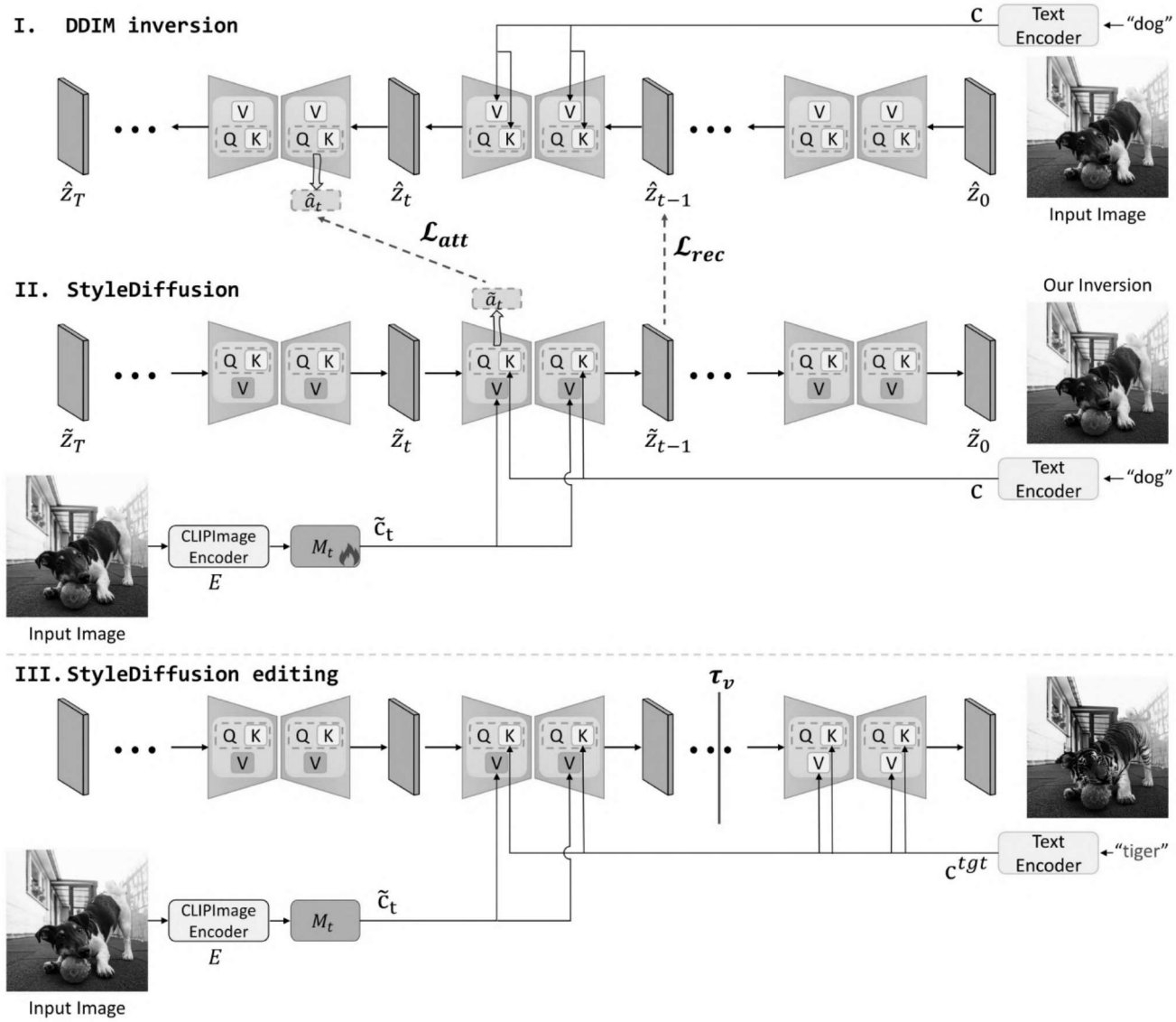

Fig. 6: Overview of the proposed method. (**I**) DDIM inversion: the diffusion process is performed to generate the latent representations: $(\hat{\mathbf{z}}_t, \hat{\mathbf{a}}_t)(t = 1, ..., T)$, where $\hat{\mathbf{z}}_0 = \mathbf{z}_0$, which is the extracted feature of the input real image $\mathbf{x}$. $\mathbf{c}$ is the extracted textual embedding by a Clip-text Encoder with a given prompt $\mathbf{p}^{src}$. (**II**) Stylediffusion: we take the input image $\mathbf{x}$ as input, and extract the prompt-embedding $\widetilde{\mathbf{c}}_t = M_t\left(E(\mathbf{x})\right)$, which is taken as the input value matrix $\mathbf{v}$ of the linear layer Ψ_V. The input of the linear layer Ψ_K is the given textual embedding $\mathbf{c}$. We get both the latent code $\widetilde{\mathbf{z}}_{\mathbf{t-1}}$ and the attention map $\widetilde{\mathbf{a}}_t$, which are aligned with both the latent code $\hat{\mathbf{z}}_{\mathbf{t-1}}$ and the attention map $\hat{\mathbf{a}}_{\mathbf{t}}$, respectively. Note $\widetilde{\mathbf{z}}_T = \hat{\mathbf{z}}_T$. (**III**) StyleDiffusion editing. From T to τ_v timestep, the input of the linear network Ψ_v comes from the learned textual embedding $\widetilde{\mathbf{c}}_{\mathbf{t}}$ produced by the trained M_t. From $\tau_v - 1$ to 1 the corresponding input comes from the prompt-embedding $\mathbf{c}^{tgt}$ of the target prompt. We use P2Plus to perform the attention exchange.

where z_t is a noise sample according to timestep $t \sim [1, T]$, and T is the number of the timesteps. The encoded text embedding $\mathbf{c}$ is extracted by a Clip-text Encoder Γ with given prompt $\mathbf{p}^{src}$: $\mathbf{c} = \Gamma(\mathbf{p}^{src})$. In this paper, we build on SD models (Rombach et al., 2021). These first trains both encoder and decoder. Then the diffusion process is performed in the latent space. Here the encoder maps the image $\mathbf{x}$ into the latent representation $\mathbf{z_0}$, and the decoder aims to reverse the latent representation $\mathbf{z_0}$ into the image. The sampling process is given by:

$$\mathbf{z}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\mathbf{z}_t + \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}), \quad (2)$$

where α_t is a scalar function.

DDIM inversion. For real-image editing with a pre-trained diffusion model, a given real image is to be reconstructed by finding its initial noise. We use the deterministic DDIM model to perform image inversion. This process is given by:

$$\mathbf{z}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\mathbf{z}_t + \sqrt{\alpha_{t+1}}\left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}). \quad (3)$$

DDIM inversion synthesizes the latent noise that produces an approximation of the input image when fed to the diffusion process. While the reconstruction based on DDIM is not sufficiently accurate, it still provides a good starting point for training, enabling us to efficiently achieve high-fidelity inversion (Hertz et al., 2022). We use the intermediate results of DDIM inversion to train our model, similarly as (Couairon et al., 2022; Mokady et al., 2022).

Cross-attention. SD models achieve text-driven image generation by feeding a prompt into the cross-attention layer. Given both the text embedding $\mathbf{c}$ and the image feature representation $\mathbf{f}$, we are able to produce the key matrix $\mathbf{k} = \Psi_K(\mathbf{c})$, the value matrix $\mathbf{v} = \Psi_V(\mathbf{c})$ and the query matrix $\mathbf{q} = \Psi_Q(\mathbf{f})$, via the linear networks: Ψ_K, Ψ_V, Ψ_Q. The attention maps are then computed with:

$$\mathbf{a} = \text{Softmax}(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{\mathbf{d}}}), \quad (4)$$

where $\mathbf{d}$ is the projection dimension of the keys and queries. Finally, the cross-attention output is $\hat{\mathbf{f}} = \mathbf{a}\mathbf{v}$, which is then taken as input in the following convolution layers.

Intuitively, P2P (Hertz et al., 2022) performs prompt-to-prompt image editing with cross attention control. P2P is based on the idea that the attention maps largely control where the image is drawn, and the values decide what is drawn (mainly defining the style). Improving the accuracy of the attention maps leads to more powerful editing capabilities (Mokady et al., 2022). We experimentally observe that DDIM inversion generates satisfying attention maps (e.g., Fig. 14 (second row, first column)), and provides a good starting point for the optimization. Next, we investigate the attention maps to guide the image inversion.

3.2 StyleDiffusion

Method overview. As illustrated in Fig. 6 (I), given a pair of a real image $\mathbf{x}$ and a corresponding prompt $\mathbf{p}^{src}$ (i.e., "dog"), we perform DDIM inversion (Dhariwal and Nichol, 2021; Song et al., 2020) to synthesize a series of latent noises $\{\hat{\mathbf{z}}_t\}$ and attention maps $\{\hat{\mathbf{a}}_t\}(t = 1, ..., T)$, where $\hat{\mathbf{z}}_0 = \mathbf{z}_0$, which is the extracted latent code of the input image $\mathbf{x}$ [2]. Fig. 6 (II) shows that our method reconstructs the latent noise $\hat{\mathbf{z}}_t$ in the order of the diffusion process $T \to 0$, where $\widetilde{\mathbf{z}}_T = \hat{\mathbf{z}}_T$. Our framework consists of three networks: a frozen ClipImageEncoder E, a learnable mapping network M_t and a denoiser network ϵ_θ. For a specific timestep t, the ClipImageEncoder E takes the input image $\mathbf{x}$ as an input. The output $E(\mathbf{x})$ is fed into the mapping network M_t, producing the prompt-embedding $\widetilde{\mathbf{c}}_\mathbf{t} = M_t(E(\mathbf{x}))$, which is fed into the value network Ψ_V of the cross-attention layers. The input of the linear layer Ψ_K is the given textual embedding $\mathbf{c}$. We generate both the latent code $\widetilde{\mathbf{z}}_{\mathbf{t-1}}$ and the attention map $\widetilde{\mathbf{a}}_t$. Our full algorithm is presented in algorithm 1.

Algorithm 1 Our algorithm

Require: the features of the training image and the prompt embeddings: $\{\mathbf{z}_0, \mathbf{c}\}$. $K_t = e^{-t} * K$, where $K = 100$ which is the starting number of inner iterations. K_t is the number of training iteration for each timestep t. The mapping network $\{M_t\}(t = 1, ..., T)$ with initialization parameters ω.
Temporary results: With guidance scale $w = 1$ for the classifier-free diffusion model, we use DDIM inversion to produce $\{\hat{\mathbf{z}}_j, \hat{\mathbf{a}}_j\}(j = 1, ..., T)$.
Output: Mapping network $\{M_t\}(t = 1, ..., T)$.

```
Set guidance scale w = 7.5;
Initializing z̃_T ← ẑ_T;
for t = T, T − 1, ..., 1 do
    for k = 0, ..., K_t − 1 do
        a_t, z_{t−1} ← z̃_t;(Eqs. 4 and 6)
        ω ← ω − η∇_ω L ;(Eq. 8)
    end
    Synthesizing z̃_{t−1};(Eq. 6)
end
Return Mapping network {M_t}(t = 1, ..., T)
```

The full loss function consists of two losses: *reconstruction loss* and *attention loss*, which guarantee that both the denoised latent code $\widetilde{\mathbf{z}}_{\mathbf{t-1}}$ and the corresponding attention map $\widetilde{\mathbf{a}}_\mathbf{t}$ at inference time are close to the ones: $\hat{\mathbf{z}}_{\mathbf{t-1}}$ and $\hat{\mathbf{a}}_\mathbf{t}$ from DDIM inversion, respectively.

Reconstruction Loss. Since the noise representations ($\{\hat{\mathbf{z}}_1, \cdots \hat{\mathbf{z}}_T\}$) provide an initial trajectory which is close to the real image, we train the network M_t to generate the prompt embedding $\widetilde{\mathbf{c}}_\mathbf{t} = M_t(E(\mathbf{x}))$, which is the input of the value network. We optimize the M_t in such a matter, that the output latent code ($\widetilde{\mathbf{z}}_t$) is close to the noise representations ($\hat{\mathbf{z}}_t$). The objective is:

$$\mathcal{L}_{rec} = \min_{M_t} \|\hat{\mathbf{z}}_{t-1} - \widetilde{\mathbf{z}}_{t-1}\|^2, \quad (5)$$

[2] Note when generating the attention map $\hat{\mathbf{a}}_T$ in the last timestep $t = T$, we throw out the synthesized latent code $\hat{\mathbf{z}}_{T+1}$.

$$\widetilde{\mathbf{z}}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\widetilde{\mathbf{z}}_t + \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_{t-1}}-1} - \sqrt{\frac{1}{\alpha_t}-1}\right)\cdot \epsilon_\theta(\widetilde{\mathbf{z}}_t, t, \mathbf{c}, \mathbf{c_t}). \quad (6)$$

Attention loss. It is known that a more accurate attention map is positively correlated to the editing capability (Mokady et al., 2022). The attention map, which is synthesized during the DDIM inversion, provides a good starting point. Thus, we introduce attention regularization

when optimizing the mapping network M_t to further improve its quality. The objective is the following:

$$\mathcal{L}_{att} = \min_{M_t} \|\hat{\mathbf{a}}_t - \widetilde{\mathbf{a}}_t\|^2, \quad (7)$$

where $\hat{\mathbf{a}}_t$ and $\widetilde{\mathbf{a}}_t$ can be obtained with Eqs. 4.

Full objective. The full objective function of our model is:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{att}. \quad (8)$$

In conclusion, in this section, we have proposed an alternative solution to the inversion of text-guided diffusion models which aims to improve over existing solutions by providing more accurate editing capabilities, and without requiring careful prompt engineering.

3.3 P2Plus

After having inverted the text-guided diffusion model, we can now perform prompt-based image editing (see Figs. 2 and 9). We will here outline our approach, which is an improvement on the popular P2P (Hertz et al., 2022).

P2P performs the replacement of both the cross-attention map and the self attention map of the conditional branch, aiming to maintain the structure of the source prompt, see Fig. 7 (middle). However, it ignores the replacement in the unconditional branch. This leads to less accurate editing capabilities for some cases, especially when the structural changes before and after editing are relatively large (e.g., "...tent..."→"...tiger..."). To address this problem, we need to reduce the dependence of the structure on the source prompt and provide more freedom to generate the structure following the target prompt. Thus, as shown in Fig. 7 (bottom) we propose to further perform the self-attention map replacement in the unconditional branch based on P2P (called *P2Plus*), as well as in the conditional branch like P2P (Hertz et al., 2022). This technique enables us to obtain more accurate editing capabilities. Like P2P, we introduce a timestep parameter τ_u that determines until which step the injection is applied. Fig. 8 shows the results with different τ_u values.

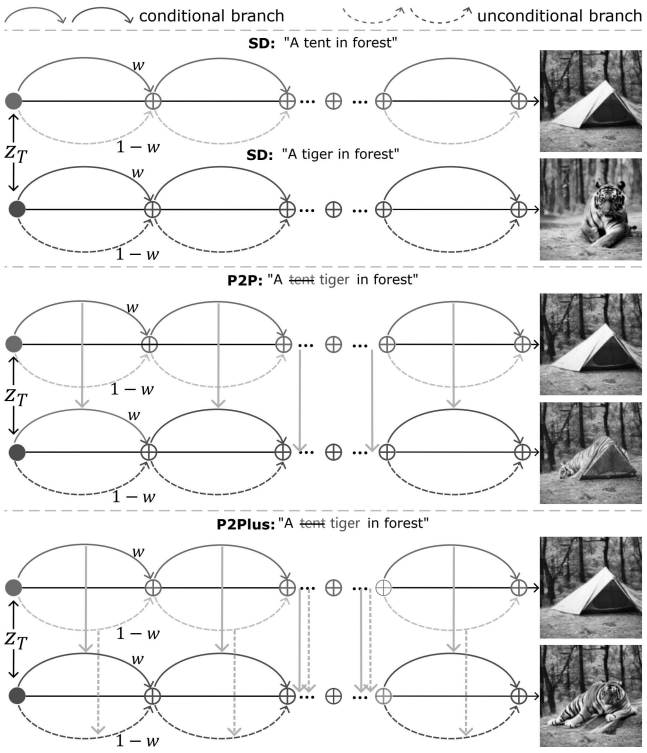

Fig. 7: (Top) Given the same latent code $\mathbf{z}_T$ and the classifier-free guidance parameter w, we generate two images with two prompts. The arrows indicate the diffusion process from T to 0. Here green is used for the generation with the source prompt: 'A tent in forest', and red for the generation with the target prompt: 'A tiger in forest'. (Middle) P2P copies both the self-attention maps and the cross-attention maps of the conditional branch of the SD models from the source prompt into the corresponding self-attention and cross-attention maps of the conditional branch of the SD models with target prompt. This process is indicated with yellow arrows. (Bottom) P2Plus additionally replaces the self-attention map of the unconditional branch of SD models, in addition to the ones of the conditional branch. The dashed yellow arrows indicate this technique.

4 Experimental setup

Training details and datasets. We use the pre-trained Stable Diffusion model. The mapping network M_t configuration is provided in Tab. 1. We set $T = 30$. τ_v is a timestep parameter that determines which timestep is used by the output of the mapping network in the StyleDiffusion editing phase. Similarly, we can set the timestep τ_u (as in the conditional branch in P2P (Hertz et al., 2022)) to control the number of diffusion steps in which the injection of the unconditional branch is applied. We use Adam Kingma and Ba (2014)

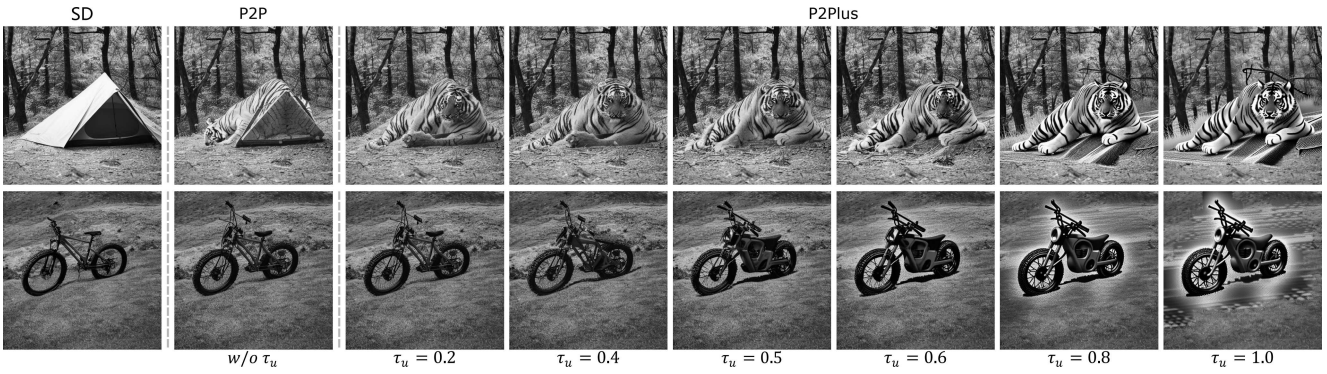

Fig. 8: *P2P* (the second column: w/o τ_u) has not succeeded in replacing the tent with the tiger. Adding the injection parameter τ_u can help to edit successfully, especially if $\tau_u = 0.5$. We also use the classifier-free guidance parameter $w = 7.5$ like SD, when the weight $1 - w$ of the unconditional branch is negative (Armandpour et al., 2023; Tumanyan et al., 2023), which can gradually weaken the influence of the "tent" in the unconditional branch as τ_u increases from 0.2 to 1.0 (third to eighth columns).

with a batch size of 1 and a learning rate of 0.0001. The exponential decay rates are $(\beta_1, \beta_2) = (0, 0.999)$. We randomly initialize the weights of the mapping network following a Gaussian distribution centered at 0 with 0.01 standard deviation. We use one Quadro RTX 3090 GPUs (24 GB VRAM) to conduct all our experiments. We randomly collect a real image dataset of 100 images and caption pairs (of 512×512 resolution) from Unsplash(`https://unsplash.com/`) and COCO (Chen et al., 2015).

Evaluation metrics. *Clipscore* (Hessel et al., 2021) is a metric that evaluates the quality of a pair of a prompt and an edited image. To evaluate the preservation of the structure information after editing, we use Structure Dist (Tumanyan et al., 2022) to compute the structural consistency of the edited image. Furthermore, in this paper, we aim to modify the selected region, which corresponds to the target prompt, and preserve the non-selected region. Thus, we need to evaluate the change in the non-selected region after editing. To get automatically the non-selected region of the edited image, we use a binary method to generate the raw mask from the attention map. Then we reverse it to get the non-selected region mask. Using the non-selected region mask, we compute the non-selected region LPIPS (Zhang et al., 2018) between a pair of real and edited images, named *NS-LPIPS*. A lower score on NS-LPIPS means that the non-selected region is more similar to the input image. We also use both PSNR and SSIM to evaluate image reconstruction.

Baselines. We compare against the following baselines. *Null-text* (Mokady et al., 2022) inverts real images with corresponding captions into the text embedding of the unconditional part of the classifier-free diffusion model. *SDEdit* (Meng et al., 2021) introduces a stochastic differential equation to generate realistic images through an iterative denoising process. *Pix2pix-zero* (Parmar et al., 2023) edits the real image to find the potential direction from the source to the target words. We compare with *DDIM + word swap* (Parmar et al., 2023) which performs DDIM sampling with an edited prompt generated by swapping the source word with the target. We use the official codes of the baselines to compare with StyleDiffusion. We also ablate variants of StyleDiffusion.

Metric	(Input channel, Output channel)	(Kernal size,Stride)	Input Dimension	Output dimension
Conv0	(197, 77)	(1, 1)	(197, 768)	(77, 768)
Conv1	(77, 77)	(1, 1)	(77, 768)	(77, 768)
BatN1	-	-	(77, 768)	(77, 768)
LeakyRelu1	-	-	(77, 768)	(77, 768)
Conv2	(77, 77)	(1, 1)	(77, 768)	(77, 768)

Table 1: Mapping network M_t configuration.

Metric	Structure-dist↓	NS-LPIPS↓	Clipscore↑
*DDIM	0.092	0.4131	**81.9**%
SDEit	0.046	0.2473	78.0%
Null-text	0.027	0.1480	75.2%
Ours	**0.026**	**0.1165**	77.9%

Table 2: Comparison with baselines on three metrics. NS-LPIPS: non-selected LPIPS. *DDIM: DDIM inversion with word swap. Although *DDIM with word swap* achieves the best Clipscore, it not only changes the background, but also modifies the structure of the selected region, see also Fig. 9 (last three rows, fourth column).

	Inference Time↓	PSNR↑/SSIM↑
Null-text	**0.28s**	31.314/0.730
Ours	0.30s	**31.523/0.751**

Table 3: We report the inference time of each timestep and PSNR/SSIM. We have better reconstruction quality with a small computational overhead.

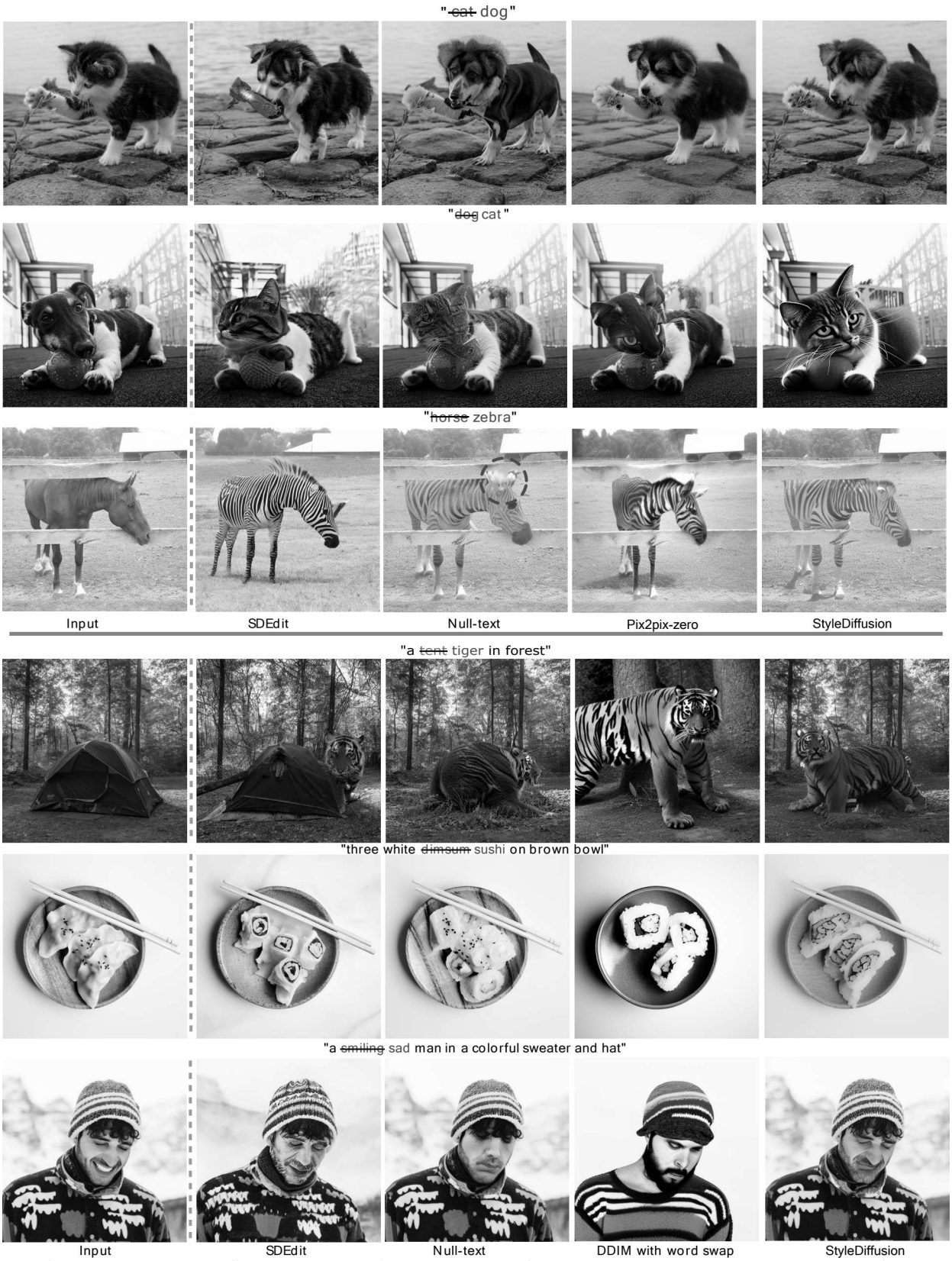

Fig. 9: Comparisons with different baselines for real images. Our method, achieves realistic editing of both style and structured objects, while preserving the structure of the input image (last column).

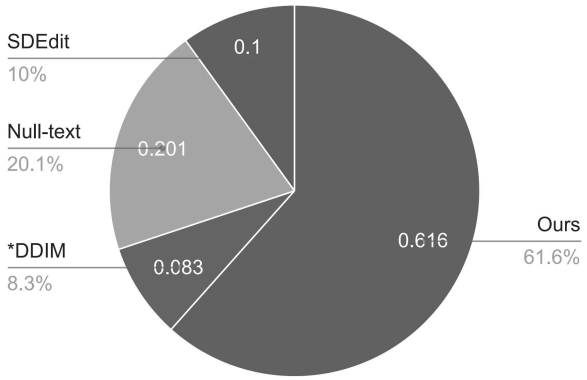

Fig. 10: We conduct a forced-choice user study and ask subjects to select the results according to *'Which figure does best preserve the input image structure and matches target prompt style?'*.

5 Experiments

Qualitative and quantitative results. Fig. 9 presents a comparison between the baselines and our method. *SDEit* (Meng et al., 2021) fails to generate high-quality images, such as dog or cat faces (second column). *Pix2pix-zero* (Parmar et al., 2023) synthesizes better results, but it also modifies the non-selected region, such as removing the plant when translating cat → dog. The official implementation of *pix2pix-zero* (Parmar et al., 2023) provides the editing directions (e.g, cat ↔ dog), and we directly use them. Note, *pix2pix-zero* (Parmar et al., 2023) firstly requires that the editing directions are calculated in advance, while our method does not require this. Fig. 9 (last three rows, fourth column) shows that *DDIM with word swap* largely modifies both the background and the structure information of the foreground. Our method successfully edits the target-specific object, resulting in a high-quality image, indicating that the proposed method has more accurate editing capabilities.

We evaluate the performance of the proposed method on the collected dataset. As reported in Tab. 2, in terms of both Structure distance and NS-LPIPS the proposed method achieves the best score, which indicates that we have superior capabilities to preserve structure information. In terms of Clipscore, we get a better score than Null-text (i.e., 77.9% vs 75.2%), and a comparative result with SDEdit. *DDIM with word swap* achieves the best Clipscore. We observe that *DDIM with word swap* not only changes the background, but also modifies the structure of the selected-regions (see Fig. 9 (last three rows, fourth column)). Note that we do not compare with pix2pix-zero (Parmar et al., 2023) in Fig. 9 (last three rows), since it first needs to compute the textual embedding directions with thousands of sentences with GPT-3 (Brown et al., 2020). We also evaluate the reconstruction quality and the inference time for each timestep. As reported in Tab. 3, we achieve the best PSNR/SSIM scores, with an acceptable time overhead.

Furthermore, we conduct a user study and ask subjects to select the results that best match the following statement: *which figure preserves the input image structure and matches the target prompt style* (Figure 10). We apply quadruplet comparisons (forced choice) with 18 users (30 quadruplets/user). Experiments are performed on images from the collected dataset. Fig. 10 shows that our method considerably outperforms the other methods.

Fig 11 shows that we can manipulate the inverted image with attention injection and prompt refinement. For example, we translate *glasses* into *sunglasses* (Fig. 11 (first row)). Fig. 11 (last row) we add *Chinese style* (new prompts) to the source prompt. These results indicate that our approach manages to invert real images with corresponding captions into the latent space, while maintaining its powerful editing capabilities.

We observe that StyleDiffusion (Fig. 12 (last column)) allows for object structure modifications while preserving the identity within the range given by the input image cross-attention map, resembling the capabilities demonstrated by Imagic (Kawar et al., 2022) (the third column)). In contrast, Null-text (Mokady et al., 2022) does not possess the capacity to accomplish such changes (Fig. 12 (the second column)).

Ablation study. Here, we evaluate the effect of each independent contribution to our method and their combinations.

Attention injection in the unconditional branch. Although P2P (Hertz et al., 2022) obtains satisfactory editing results with attention injection in the conditional branch, this method ignores attention injection in the unconditional branch (as proposed by our P2Plus in Sec. 3.3). We experimentally observe that the self-attention maps in the unconditional branch play an important role in obtaining more accurate editing capabilities, especially when the object structure changes before and after editing of the real image are relatively large, e.g., translating *bike* to *motorcycle* in Fig. 13 (left, the third row). It also shows that the unconditional branch contains a lot of useful texture and structure information, allowing us to reduce the influence of the unwanted structure of the input image.

Prompt-embedding in cross-attention layers. We evaluate variants of our method, namely (1) learning the input prompt-embedding for the key linear layer

Fig. 11: Examples of StyleDiffusion for editing with attention injection or prompt refinement.

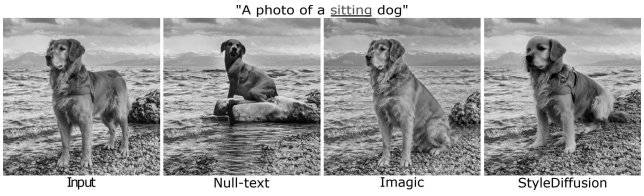

Fig. 12: StyleDiffusion can achieve object structure changes within the range of the input image cross-attention map (e.g., "...dog"→"...sitting dog").

and freezing the input of the value linear layer with the one provided by the user, and (2) learning the prompt-embedding for both key and value linear layers. As shown in Fig. 13 (right), the two variants fail to edit the image according to the target prompt. Our method successfully modifies the real image with the target prompt, and produces realistic results.

Attention regularization. We perform an ablation study of attention regularization. Fig. 14 shows that the system fails to reconstruct partial object information (e.g., the nose in Fig. 14 (first row, second column)), and learns a less accurate attention map (e.g., the nose attention map in Fig. 14 (first row, third column). Our method not only synthesizes high-quality images, but also learns a better attention map even compared to the one generated by DDIM inversion (Fig. 14 (second row, first column)).

Value-embedding optimization. Fig. 15 illustrates the reconstruction and editing results of value-embedding optimization, that is, similar to our method extracting the prompt-embedding from the input image but directly optimizing the input textual embedding. Value-embedding optimization fails to reconstruct the input image. Null-text (Mokady et al., 2022) draws a similar conclusion that optimizing both the input textual embedding for the value and key linear layers results in lower editing accuracy.

StyleDiffusion with SDEdit. After inverting a real image with StyleDiffusion, we leverage SDEdit to edit it. Only using SDEdit, the results suffer from unwanted changes, such as the orientation of the dog (Fig. 17 (first row, second column) and the texture detail of the leg of the dog (Fig. 17 (second row, second col-

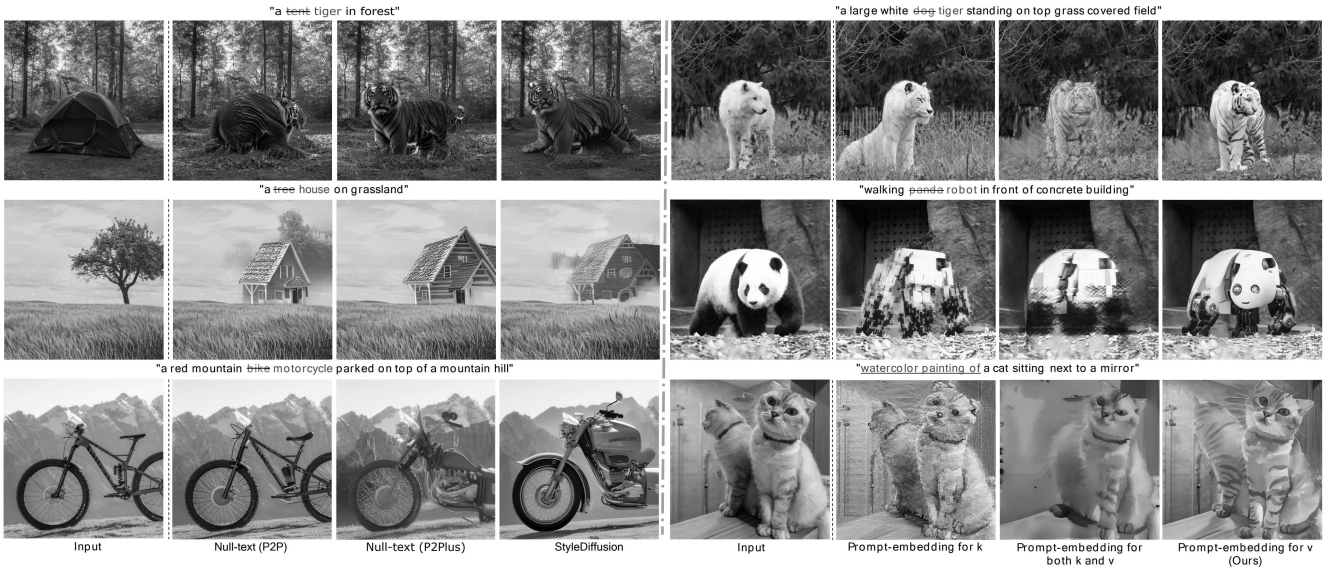

Fig. 13: (Left) Using additionally the attention injection in unconditional branch improves the real image editing ability of *Null-text* (Mokady et al., 2022) (*P2P*). (Right) Comparison of variants of our method.

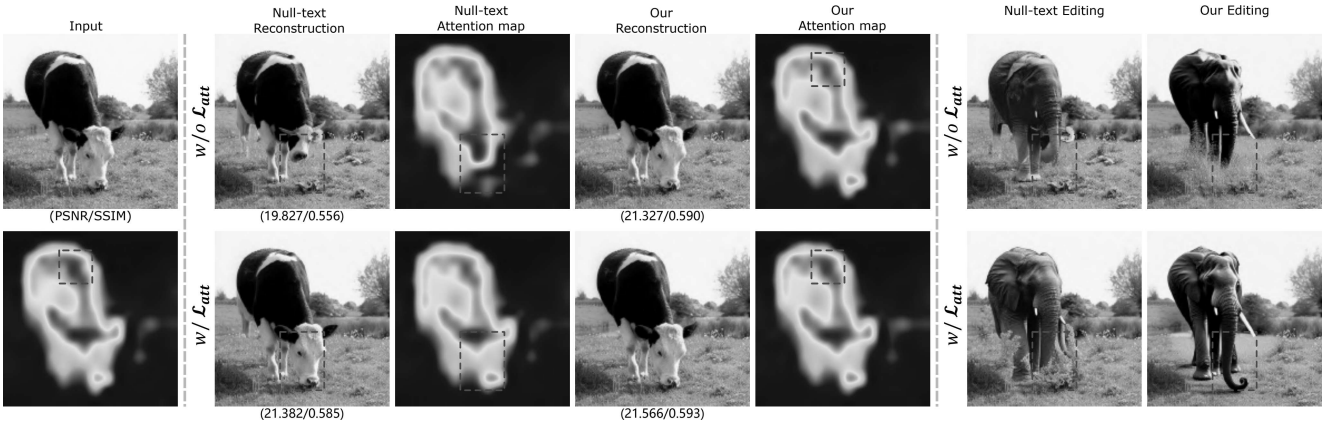

Fig. 14: The reconstruction effect of attention regularization. The cross-attention map of our reconstruction image (second row, fifth column) more closely matches the one of the input image (second row, first column). Meanwhile, $\mathcal{L}_{att}$ can improve the reconstruction quality of Null-text (second row, second column). In the last two columns, the *cow* is replaced by an *elephant* (source prompt: "cow", target prompt: "elephant").

umn)). While combining StyleDiffusion and SDEdit significantly enhances the fidelity to the input image (see Fig. 17 (the third column)). This indicates our method exhibits robust performance when combining different editing techniques (e.g., SDEdit and P2Plus).

Comparison with optimization-free methods. Recently, some methods without optimization have been proposed (Miyake et al., 2023; Han et al., 2023b). Negative-prompt inversion (NPI) (Miyake et al., 2023) replaces the null-text embedding of the unconditional branch with the textural embedding in SD to implement reconstruction and editing. Proximal Negative-Prompt Inversion (ProxNPI) (Han et al., 2023b) attempts to enhance NPI by introducing regularization terms using proximal function and reconstruction guidance based on the foundation of NPI. While these methods do not require optimizing parameters to achieve the inversion of real images, similar to the method shown in Fig. 1, they suffer from challenges when reconstructing and editing images containing intricate content and structure (see Fig. 16 (the second and third columns, the sixth and seventh columns)). Due to the absence of an optimization process in these methods, it is not feasible to utilize attention loss to refine the attention maps like Null-text+$\mathcal{L}_{att}$ (Fig. 14), consequently limiting the potential for enhancing reconstruction and editing quality.

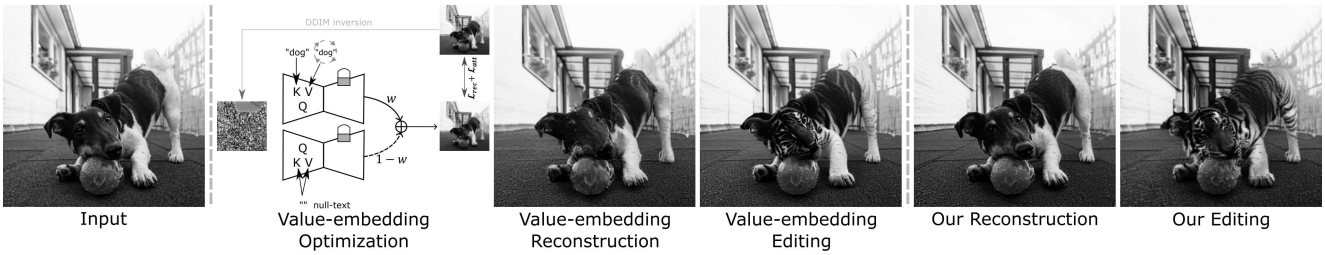

Fig. 15: ***Value-embedding optimization.*** We compare our method (right) to the one (middle), where we directly optimize the input textual embedding for the value linear layer while freezing the input of the key linear layer. As can be seen (middle), this approach leads to an inaccurate reconstruction, resulting in the dog's face not being completely reconstructed.

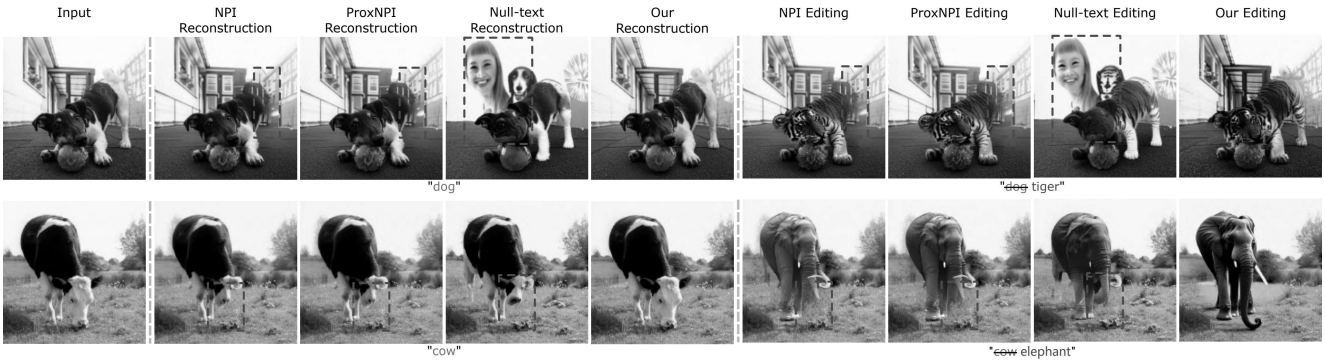

Fig. 16: Both the optimization-free methods NPI and ProxNPI (the second and third columns, the sixth and seventh columns) show limitations in reconstructing and editing real images that contain complex structures and content.

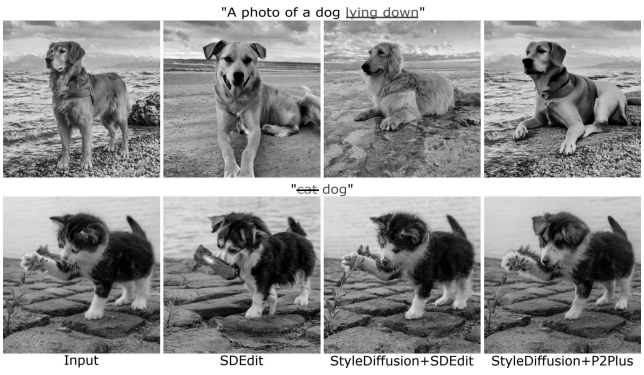

Fig. 17: ***StyleDiffusion with SDEdit.*** From left to right: input image, SDEdit, applying SDEdit after StyleDiffusion inversion, and applying P2Plus after StyleDiffusion inversion. It is evident that StyleDiffusion+SDEdit significantly improves the fidelity of the input image compared to SDEdit alone.

StyleDiffusion for style transfer. As a final illustration, we show that StyleDiffusion can be used to perform style transfer. As shown in Fig. 19(left), given a content image, we use DDIM inversion to generate a series of timestep-related latent codes. They are then progressively denoised using DDIM sampling. During this process, we extract the spatial features from the decoder layers. These spatial features are injected into the corresponding layers of StyleDiffusion model. Note that we first optimize StyleDiffusion to reconstruct the style image, then use both the well-learned M_t and the extracted content feature to perform the style transfer task. Fig. 19(right) exhibits that we can successfully combine both content and style images, and perform style transfer.

Fig. 18: Some examples of failure cases.

6 Conclusions and Limitations

We propose a new method for real image editing. We invert the real image into the input of the value linear mapping network in the cross-attention layers, and

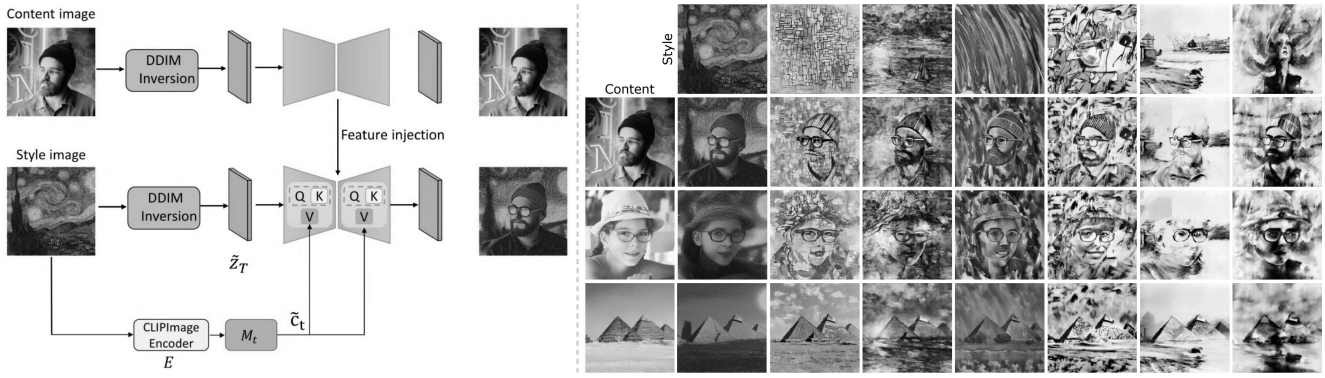

Fig. 19: ***StyleDiffusion for style transfer.*** (Left) our framework for style transfer. Given a content image, we use DDIM inversion to generate a series of timestep-related latent codes. They are then progressively denoised using DDIM sampling. During this process, we extract the spatial features from the decoder layers. These spatial features are injected into the corresponding layers of StyleDiffusion model. Note we first optimize StyleDiffusion to reconstruct the style image, then use both the learned-well M_t and the extracted content feature to perform the style transfer task. This approach allows us to efficiently transfer the desired artistic style to the content image without the need for additional optimization on the content image. (Right) results of style transfer with StyleDiffusion.

freeze the input of the key linear layer with the textual embedding provided by the user. This allows us to learn initial attention maps, and an approximate trajectory to reconstruct the real image. We introduce a new attention regularization to preserve the attention maps after editing, enabling us to obtain more accurate editing capabilities. In addition, we propose attention injection in the unconditional branch of the classifier-free diffusion model (P2Plus), further improving the editing capabilities, especially when both source and target prompts have a large domain shift.

While StyleDiffusion successfully modifies the real image, it still suffers from some limitations. Our method fails to generate satisfying images when the object of the real image has a rare pose (Fig. 18 (left)), or both the source and the target prompts have a large semantic shift (Fig. 18 (right)).

References

Abdal R, Qin Y, Wonka P (2019) Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4432–4441

Abdal R, Qin Y, Wonka P (2020) Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8296–8305

Alaluf Y, Tov O, Mokady R, Gal R, Bermano AH (2021) Hyperstyle: Stylegan inversion with hypernetworks for real image editing. 2111.15666

Armandpour M, Zheng H, Sadeghian A, Sadeghian A, Zhou M (2023) Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. arXiv preprint arXiv:230404968

Avrahami O, Fried O, Lischinski D (2022a) Blended latent diffusion. arXiv preprint arXiv:220602779

Avrahami O, Lischinski D, Fried O (2022b) Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 18208–18218

Avrahami O, Aberman K, Fried O, Cohen-Or D, Lischinski D (2023) Break-a-scene: Extracting multiple concepts from a single image. arXiv preprint arXiv:230516311

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473

Bermano AH, Gal R, Alaluf Y, Mokady R, Nitzan Y, Tov O, Patashnik O, Cohen-Or D (2022) State-of-the-art in the architecture, methods and applications of stylegan. In: Computer Graphics Forum, Wiley Online Library, vol 41, pp 591–611

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. (2020) Language models are few-shot learners. Advances in neural information processing

systems 33:1877–1901

Cao M, Wang X, Qi Z, Shan Y, Qie X, Zheng Y (2023) Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:230408465

Chen M, Laina I, Vedaldi A (2023) Training-free layout control with cross-attention guidance. arXiv preprint arXiv:230403373

Chen X, Fang H, Lin TY, Vedantam R, Gupta S, Dollár P, Zitnick CL (2015) Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:150400325

Couairon G, Verbeek J, Schwenk H, Cord M (2022) Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:221011427

Creswell A, Bharath AA (2018) Inverting the generator of a generative adversarial network. IEEE transactions on neural networks and learning systems 30(7):1967–1974

Dhariwal P, 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, Han S (2023) Prompt tuning inversion for text-driven image editing using diffusion models. arXiv preprint arXiv:230504441

Gafni O, Polyak A, Ashual O, Sheynin S, Parikh D, Taigman Y (2022) Make-a-scene: Scene-based text-to-image generation with human priors. arXiv preprint arXiv:220313131

Gal R, Patashnik O, Maron H, Chechik G, Cohen-Or D (2021) Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint arXiv:210800946

Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano AH, Chechik G, Cohen-Or D (2022) An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:220801618

Gal R, Arar M, Atzmon Y, Bermano AH, Chechik G, Cohen-Or D (2023) Encoder-based domain tuning for fast personalization of text-to-image models. arXiv preprint arXiv:230212228

Goetschalckx L, Andonian A, Oliva A, Isola P (2019) Ganalyze: Toward visual definitions of cognitive image properties. In: ICCV, pp 5744–5753

Han I, Yang S, Kwon T, Ye JC (2023a) Highly personalized text embedding for image manipulation by stable diffusion. arXiv preprint arXiv:230308767

Han L, Wen S, Chen Q, Zhang Z, Song K, Ren M, Gao R, Chen Y, Liu D, Zhangli Q, et al. (2023b) Improving negative-prompt inversion via proximal guidance. arXiv preprint arXiv:230605414

Hertz A, Mokady R, Tenenbaum J, Aberman K, Pritch Y, Cohen-Or D (2022) Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:220801626

Hessel J, Holtzman A, Forbes M, Bras RL, Choi Y (2021) CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP

Ho J, Salimans T (2021) Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications

Huang R, Lam MW, Wang J, Su D, Yu D, Ren Y, Zhao Z (2022) Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. arXiv preprint arXiv:220409934

Jahanian A, Chai L, Isola P (2020) On the"steerability" of generative adversarial networks. In: ICLR

Jeong M, Kim H, Cheon SJ, Choi BJ, Kim NS (2021) Diff-tts: A denoising diffusion model for text-to-speech. arXiv preprint arXiv:210401409

Jia X, Zhao Y, Chan KC, Li Y, Zhang H, Gong B, Hou T, Wang H, Su YC (2023) Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:230402642

Kang M, Zhu JY, Zhang R, Park J, Shechtman E, Paris S, Park T (2023) Scaling up gans for text-to-image synthesis. In: CVPR

Kawar B, Zada S, Lang O, Tov O, Chang HT, Dekel T, Mosseri I, Irani M (2022) Imagic: Text-based real image editing with diffusion models. ArXiv abs/2210.09276

Khachatryan L, Movsisyan A, Tadevosyan V, Henschel R, Wang Z, Navasardyan S, Shi H (2023) Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:230313439

Kim G, Kwon T, Ye JC (2022) Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2426–2435

Kingma D, Ba J (2014) Adam: A method for stochastic optimization. ICLR

Koizumi Y, Zen H, Yatabe K, Chen N, Bacchiani M (2022) Specgrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping. arXiv preprint arXiv:220316749

Kumari N, Zhang B, Zhang R, Shechtman E, Zhu JY (2022) Multi-concept customization of text-to-image diffusion. arXiv preprint arXiv:221204488

Kumari N, Zhang B, Zhang R, Shechtman E, Zhu JY (2023) Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1931–1941

Kwon M, Jeong J, Uh Y (2022) Diffusion models already have a semantic latent space. arXiv preprint

arXiv:221010960

Levin E, Fried O (2023) Differential diffusion: Giving each pixel its strength. arXiv preprint arXiv:230600950

Lin CH, Gao J, Tang L, Takikawa T, Zeng X, Huang X, Kreis K, Fidler S, Liu MY, Lin TY (2023) Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 300–309

Lipton ZC, Tripathi S (2017) Precise recovery of latent vectors from generative adversarial networks. arXiv preprint arXiv:170204782

Liu G, Sun H, Li J, Yin F, Yang Y (2023) Accelerating diffusion models for inverse problems through shortcut sampling. arXiv preprint arXiv:230516965

Meng C, Song Y, Song J, Wu J, Zhu JY, Ermon S (2021) Sdedit: Image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:210801073

Miyake D, Iohara A, Saito Y, Tanaka T (2023) Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. arXiv preprint arXiv:230516807

Mokady R, Hertz A, Aberman K, Pritch Y, Cohen-Or D (2022) Null-text inversion for editing real images using guided diffusion models. arXiv preprint arXiv:221109794

Mou C, Wang X, Song J, Shan Y, Zhang J (2023) Dragondiffusion: Enabling drag-style manipulation on diffusion models. arXiv preprint arXiv:230702421

Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M (2021) Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:211210741

Parmar G, Singh KK, Zhang R, Li Y, Lu J, Zhu JY (2023) Zero-shot image-to-image translation. arXiv preprint arXiv:230203027

Patashnik O, Wu Z, Shechtman E, Cohen-Or D, Lischinski D (2021) Styleclip: Text-driven manipulation of stylegan imagery. arXiv preprint arXiv:210317249

Poole B, Jain A, Barron JT, Mildenhall B (2022) Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:220914988

Qiu Z, Liu W, Feng H, Xue Y, Feng Y, Liu Z, Zhang D, Weller A, Schölkopf B (2023) Controlling text-to-image diffusion by orthogonal finetuning. arXiv preprint arXiv:230607280

Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:220406125

Richardson E, Alaluf Y, Patashnik O, Nitzan Y, Azar Y, Shapiro S, Cohen-Or D (2020) Encoding in style: a stylegan encoder for image-to-image translation. arXiv preprint arXiv:200800951

Roich D, Mokady R, Bermano AH, Cohen-Or D (2022) Pivotal tuning for latent-based editing of real images. ACM Transactions on Graphics (TOG)

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2021) High-resolution image synthesis with latent diffusion models. 2112.10752

Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K (2022) Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:220812242

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, Ghasemipour SKS, Ayan BK, Mahdavi SS, Lopes RG, Salimans T, Salimans T, Ho J, Fleet DJ, Norouzi M (2022) Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:220511487

Sauer A, Karras T, Laine S, Geiger A, Aila T (2023) StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. In: International Conference on Machine Learning, vol abs/2301.09515

Song J, Meng C, Ermon S (2020) Denoising diffusion implicit models. In: International Conference on Learning Representations

Tewel Y, Gal R, Chechik G, Atzmon Y (2023) Key-locked rank one editing for text-to-image personalization. arXiv preprint arXiv:230501644

Tov O, Alaluf Y, Nitzan Y, Patashnik O, Cohen-Or D (2021) Designing an encoder for stylegan image manipulation. arXiv preprint arXiv:210202766

Tumanyan N, Bar-Tal O, Bagon S, Dekel T (2022) Splicing vit features for semantic appearance transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10748–10757

Tumanyan N, Geyer M, Bagon S, 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 (CVPR), pp 1921–1930

Valevski D, Kalman M, Matias Y, Leviathan Y (2022) Unitune: Text-driven image editing by fine tuning an image generation model on a single image. arXiv preprint arXiv:221009477

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems, vol 30

Wang Z, Lu C, Wang Y, Bao F, Li C, Su H, Zhu J (2023) Prolificdreamer: High-fidelity and diverse text-to-3d

generation with variational score distillation. arXiv preprint arXiv:230516213

Wu C, Herranz L, Liu X, Van De Weijer J, Raducanu B, et al. (2018) Memory replay gans: Learning to generate new categories without forgetting. Advances in Neural Information Processing Systems 31

Wu JZ, Ge Y, Wang X, Lei W, Gu Y, Hsu W, Shan Y, Qie X, Shou MZ (2022) Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:221211565

Xia W, Zhang Y, Yang Y, Xue JH, Zhou B, Yang MH (2021) Gan inversion: A survey. 2101.05278

Xiao G, Yin T, Freeman WT, Durand F, Han S (2023) Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:230510431

Xie E, Yao L, Shi H, Liu Z, Zhou D, Liu Z, Li J, Li Z (2023) Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. arXiv preprint arXiv:230406648

Yeh RA, Chen C, Lim TY, Schwing AG, Hasegawa-Johnson M, Do MN (2017) Semantic image inpainting with deep generative models. 1607.07539

Yu J, Xu Y, Koh JY, Luong T, Baid G, Wang Z, Vasudevan V, Ku A, Yang Y, Ayan BK, et al. (2022) Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:220610789

Zhang L, Agrawala M (2023) Adding conditional control to text-to-image diffusion models. 2302.05543

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

Zhang Y, Dong W, Tang F, Huang N, Huang H, Ma C, Lee TY, Deussen O, Xu C (2023a) Prospect: Expanded conditioning for the personalization of attribute-aware image generation. arXiv preprint arXiv:230516225

Zhang Z, Huang Z, Liao J (2023b) Continuous layout editing of single images with diffusion models. arXiv preprint arXiv:230613078

Zhou D, Wang W, Yan H, Lv W, Zhu Y, Feng J (2022) Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:221111018

Zhu J, Shen Y, Zhao D, Zhou B (2020) In-domain gan inversion for real image editing. arXiv preprint arXiv:200400049

Zhu JY, Krähenbühl P, Shechtman E, Efros AA (2016) Generative visual manipulation on the natural image manifold. In: European conference on computer vision, Springer, pp 597–613