This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing

Bingyan Liu[1,2], Chengyu Wang[2]*, Tingfeng Cao[1,2], Kui Jia[3]*, Jun Huang[2]
[1]South China University of Technology, [2] Alibaba Group,
[3]School of Data Science, The Chinese University of Hong Kong, Shenzhen
{eeliubingyan, setingfengcao}@mail.scut.edu.cn,
{chengyu.wcy,huangjun.hj}@alibaba-inc.com, kuijia@cuhk.edu.cn

Abstract

Deep Text-to-Image Synthesis (TIS) models such as Stable Diffusion have recently gained significant popularity for creative text-to-image generation. However, for domain-specific scenarios, tuning-free Text-guided Image Editing (TIE) is of greater importance for application developers. This approach modifies objects or object properties in images by manipulating feature components in attention layers during the generation process. Nevertheless, little is known about the semantic meanings that these attention layers have learned and which parts of the attention maps contribute to the success of image editing. In this paper, we conduct an in-depth probing analysis and demonstrate that cross-attention maps in Stable Diffusion often contain object attribution information, which can result in editing failures. In contrast, self-attention maps play a crucial role in preserving the geometric and shape details of the source image during the transformation to the target image. Our analysis offers valuable insights into understanding cross and self-attention mechanisms in diffusion models. Furthermore, based on our findings, we propose a simplified, yet more stable and efficient, tuning-free procedure that modifies only the self-attention maps of specified attention layers during the denoising process. Experimental results show that our simplified method consistently surpasses the performance of popular approaches on multiple datasets. [1]

1. Introduction

Text-to-Image Synthesis (TIS) models, such as Stable Diffusion [26], DALL-E 2 [25], and Imagen [29], have demonstrated remarkable visual effects for text-to-image generation, capturing substantial attention from both academia and industry [6, 20, 37, 38]. These TIS models are trained on vast amounts of image-text pairs, such as Laion [30, 31], and employ cutting-edge techniques, including large-scale pre-trained language models [23, 24], variational auto-encoders [14], and diffusion models [11, 32] to achieve success in generating realistic images with vivid details. Specifically, Stable Diffusion stands out as a popular and extensively studied model, making significant contributions to the open-source community.

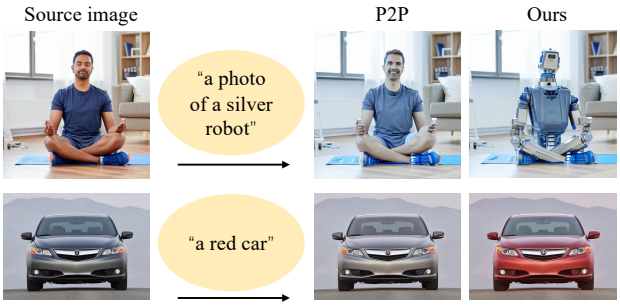

Figure 1. An example showing that our method can perform more consistent and realistic TIE compared to P2P [10].

In addition to image generation, these TIS models possess powerful image editing capabilities, which hold great importance as they aim to modify images while ensuring realism, naturalness, and meeting human preferences. Text-guided Image Editing (TIE) involves modifying an input image based on a descriptive prompt. Existing TIE methods [1, 2, 2, 5, 10, 17, 18, 21, 22, 34] achieve remarkable effects in image translation, style transfer, and appearance replacement, as well as preserving the input structure and scene layout. To this end, Prompt-to-Prompt (P2P) [10] modifies image regions by replacing cross-attention maps corresponding to the target edit words in the source prompt. Plug-and-Play (PnP) [34] first extracts the spatial features

*Co-corresponding authors.
[1]Source code and datasets are available at https://github.com/alibaba/EasyNLP/tree/master/diffusion/FreePromptEditing.

and self-attention of the original image in the attention layers and then injects them into the target image generation process. Among these methods, attention layers play a crucial role in controlling the image layout and the relationship between the generated image and the input prompt. However, inappropriate modifications to attention layers can yield varied editing outcomes and even lead to editing failures. For example, as depicted in Figure 1, editing authentic images on cross-attention layers can result in editing failures; converting a man into a robot or changing the color of a car to red fails. Moreover, some operations in the above-mentioned methods can be revised and optimized.

In our paper, we explore attention map modification to gain comprehensive insights into the underlying mechanisms of TIE using diffusion-based models. Specifically, we focus on the attribution of TIE and ask the fundamental question: *how does the modification of attention layers contribute to diffusion-based TIE?* To answer this question, we carefully construct new datasets and meticulously investigate the impact of modifying the attention maps on the resulting images. This is accomplished by probe analysis [3, 16] and systematic exploration of attention map modification with different blocks in the diffusion model. We find that (1) editing cross-attention maps in diffusion models is optional for image editing. Replacing or refining cross-attention maps between the source and target image generation process is dispensable and can result in failed image editing. (2) The cross-attention map is not only a weight measure of the conditional prompt at the corresponding positions in the generated image but also contains the semantic features of the conditional token. Therefore, replacing the target image's cross-attention map with the source image's map may yield unexpected outcomes. (3) Self-attention maps are crucial to the success of the TIE task, as they reflect the association between image features and retain the spatial information of the image. Based on our findings, we propose a simplified and effective algorithm called Free-Prompt-Editing (FPE). FPE performs image editing by replacing the self-attention map in specific attention layers during denoising, without needing a source prompt. It is beneficial for real image editing scenarios. The contributions of our paper are as follows:

- We conduct a comprehensive analysis of how attention layers impact image editing results in diffusion models and answer why TIE methods based on cross-attention map replacement can lead to unstable results.
- We design experiments to prove that cross-attention maps not only serve as the weight of the corresponding token on the corresponding pixel but also contain the characteristic information of the token. In contrast, self-attention is crucial in ensuring that the edited image retains the original image's layout information and shape details.
- Based on our experimental findings, we simplify currently popular tuning-free image editing methods and propose FPE, making the image editing process simpler and more effective. In experimental tests over multiple datasets, FPE outperforms current popular methods.

Overall, our paper contributes to the understanding of attention maps in Stable Diffusion and provides a practical solution for overcoming the limitations of inaccurate TIE.

2. Related Works

Text-guided Image Editing (TIE) [39] is a crucial task involving the modification of an input image with requirements expressed by texts. These approaches can be broadly categorized into two groups: tuning-free methods and fine-tuning based methods.

2.1. Tuning-free Methods

Tuning-free TIE methods aim to control the generated image in the denoising process. To achieve this goal, SDEdit [17] uses the given guidance image as the initial noise in the denoising step, which leads to impressive results. Other methods operate in the feature space of diffusion models to achieve successful editing results. One notable example is P2P [10], which discovers that manipulating cross-attention layers allows for controlling the relationship between the spatial layout of the image and each word in the text. Null-text inversion [18] further employs an optimization method to reconstruct the guidance image and utilizes P2P for real image editing. DiffEdit [5] automatically generates a mask by comparing different text prompts to help guide the areas of the image that need editing. PnP [34] focuses on spatial features and self-affinities to control the generated image's structure without restricting interaction with the text. Additionally, MasaCtrl [2] converts self-attention in diffusion models into a mutual and mask-guided self-attention strategy, enabling pose transformation. In this paper, we aim to provide in-depth insights into the attention layers of diffusion models and further propose a more streamlined tuning-free TIE approach.

2.2. Fine-tuning Based Methods

The core idea of fine-tuning-based TIE methods is to synthesize ideal new images by model fine-tuning over the knowledge of domain-specific data [8, 12, 13, 27] or by introducing additional guidance information [1, 19, 40]. DreamBooth [27] fine-tunes all the parameters in the diffusion model while keeping the text transformer frozen and utilizes generated images as the regularization dataset. Textual Inversion [8] optimizes a new word embedding token for each concept. Imagic [13] learns the approximate text embedding of the input image through tuning and then edits the posture of the object in the image by interpolating the approximate text embedding and the target text embedding. ControlNet [40] and T2I-Adapter [19] allow users to guide

the generated images through input images by tuning additional network modules. Instructpix2pix [1] fully fine-tunes the diffusion model by constructing image-text-image triples in the form of instructions, enabling users to edit authentic images using instruction prompts, such as "turn a man into a cyborg". In contrast to these works, our method focuses on tuning-free techniques without the fine-tuning process.

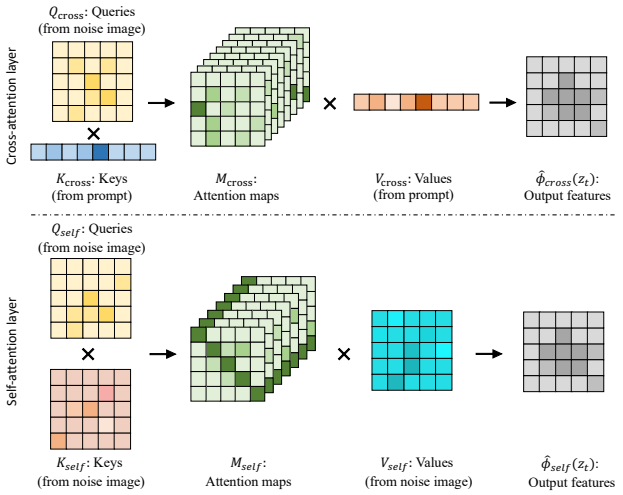

Figure 2. Cross and self-attention layers in Stable Diffusion.

3. Analysis on Cross and Self-Attention

In this section, we analyze how cross and self-attention maps in Stable Diffusion contribute to the effectiveness of TIE.

3.1. Cross-Attention in Stable Diffusion

In Stable Diffusion and other similar models, cross-attention layers play a crucial role in fusing images and texts, allowing T2I models to generate images that are consistent with textual descriptions. As depicted in the upper part of Figure 2, the cross-attention layer receives the query, key, and value matrices, i.e., Q_{cross}, K_{cross}, and V_{cross}, from the noisy image and prompt. Specifically, Q_{cross} is derived from the spatial features of the noisy image $\phi_{cross}(z_t)$ by learned linear projections ℓ_q, while K_{cross} and V_{cross} are projected from the textual embedding P_{emb} of the input prompt P using learned linear projections denoted as ℓ_k and ℓ_v, respectively. The cross-attention map is defined as:

$$Q_{cross} = \ell_q(\phi_{cross}(z_t)), \quad K_{cross} = \ell_k(P_{emb}) \tag{1}$$

$$M_{cross} = \text{Softmax}\left(\frac{Q_{cross}{K_{cross}}^{\mathsf{T}}}{\sqrt{d_{cross}}}\right) \tag{2}$$

where d_{cross} is the dimension of the keys and queries. The final output is defined as the fused feature of the text and image, denoted as $\hat{\phi}(z_t) = M_{cross}V_{cross}$, where $V_{cross} = \ell_v(P_{emb})$. Intuitively, each cell in the cross-attention map, denoted as M_{ij}, determines the weights attributed to the value of the j-th token relative to the spatial feature i of the image. The cross-attention map enables the diffusion model to locate/align the tokens of the prompt in the image area.

3.2. Self-Attention in Stable Diffusion

As depicted in Figure 2, unlike cross-attention, the self-attention layer receives the keys matrix K_{self} and the query matrix Q_{self} from the noisy image $\phi_{self}(z_t)$ through learned linear projections $\bar{\ell}_K$ and $\bar{\ell}_Q$, respectively. The self-attention map is defined as:

$$Q_{self} = \bar{\ell}_q(\phi_{self}(z_t)), \quad K_{self} = \bar{\ell}_K(\phi_{self}(z_t)), \tag{3}$$

$$M_{self} = \text{Softmax}\left(\frac{Q_{self}{K_{self}}^{\mathsf{T}}}{\sqrt{d_{self}}}\right) \tag{4}$$

where d_{self} is the dimension of K_{self} and Q_{self}. M_{self} determines the weights assigned to the relevance of the i-th and j-th spatial features in the image and can affect the spatial layout and shape details of the generated image. Consequently, the self-attention map can be utilized to preserve the spatial structure characteristics of the original image throughout the image editing process.

3.3. Probing Analysis

Yet, the semantics of cross and self-attention maps remain unclear. Are these attention maps merely weight matrices, or do they contain feature information of the image? To answer these questions, we aim to explore the meaning of attention maps in diffusion models. Inspired by probing analysis methods [3, 16] in the field of NLP, we propose building datasets and training classification networks to explore the properties of attention maps. Our fundamental idea is that if a trained classifier can accurately classify attention maps from different categories, then the attention map contains meaningful feature representation of the category information. Therefore, we introduce a task-specific classifier on top of the diffusion model's cross-attention and self-attention layers. This classifier is a two-layer MLP designed to predict specific semantic properties of the attention maps. To present the analysis results more visually, we utilize color adjectives and animal nouns to form prompt datasets, each containing ten categories. For the color adjective, there are two prompt formats: *a <color> car* and *a <color> <object>*. The prompt format for animal nouns is *a/an <animal> standing in the park*. After generating the prompts, we employ the probing method to extract the cross-attention maps corresponding to the words <color> and <animal>, along with the self-attention maps in the attention layers. Finally, by training and evaluating the performance of the classifiers, we gain insights into the semantic knowledge captured by the attention maps.

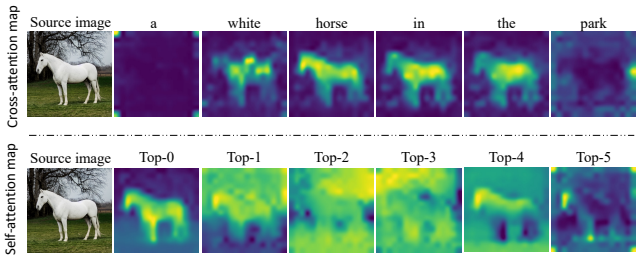

Figure 3. The heatmaps of cross-attention and self-attention maps in a generated image with the prompt "a white horse in the park". The visualization of the cross-attention map corresponds to each word in the prompt. The visualization of the self-attention map is the top-6 components obtained after SVD [36].

3.4. Probing Results on Cross-Attention Maps

What does the cross-attention map learn? We directly visualize the attention maps, as demonstrated in Figure 3. Each word in the prompt has a corresponding attention map associated with the image, indicating that the information related to the word exists in specific areas of the image. However, is this information exclusive to these areas? Referring to Equation 2, we observe that M_{cross} is derived from K_{cross} and Q_{cross}, indicating that M_{cross} carries information from both. To validate this hypothesis, we conduct probing experiments on M_{cross}, with the results presented in Table 1. Due to space limitations, we show only the probing results for five colors and five animals from the last layer of the down, middle, and up blocks. As evident in Table 1, the trained classifier achieves high accuracy in both the color and animal classification tasks. For instance, the average accuracy for classifying "sheep" reaches 98%, and that for "orange" reaches 93%. These results demonstrate that the cross-attention map acts as a reliable category representation, indicating that it reflects not only weight information but also contains category-related features. This explains the failure of image editing using cross-attention map replacement. The upper part of Figure 4 illustrates the editing results obtained by replacing the cross-attention map of the corresponding word ("rabbit" and "coral") at different cross-attention layers. It is apparent that when all layers are replaced, the editing results are the least satisfactory. The dog fails to transform completely into a rabbit, and the black car cannot turn into a coral car. Conversely, when the cross-attention map is left unaltered, correct editing results can be achieved. The complete and more additional experimental results are available in Section 8 in the Supplementary Material.

3.5. Probing Results on Self-Attention Maps

What does the self-attention map learn? Table 2 presents the results of the probing experiments. The results indicate that the trained classifier struggles to classify the

Class	Layer 3	Layer 6	Layer 9	Layer 10	Layer 12	Layer 14	Layer 16	Avg.
dog	1.00	1.00	1.00	1.00	0.89	0.76	1.00	0.95
horse	0.96	1.00	1.00	1.00	0.64	1.00	0.91	0.93
sheep	0.97	1.00	1.00	1.00	1.00	0.90	0.97	0.98
leopard	0.97	1.00	1.00	1.00	0.97	0.79	0.87	0.94
tiger	1.00	1.00	0.97	1.00	0.88	1.00	0.97	0.97
green	0.93	0.91	0.91	0.96	0.67	0.38	0.60	0.77
white	0.97	1.00	0.94	0.97	0.97	0.61	0.85	0.90
orange	0.97	1.00	0.94	0.92	0.89	0.94	0.83	0.93
yellow	0.96	0.77	1.00	0.98	1.00	0.36	0.68	0.82
red	0.97	0.97	0.93	0.85	0.70	0.23	0.65	0.76

Table 1. Probing accuracy of cross-attention map in difference layers.

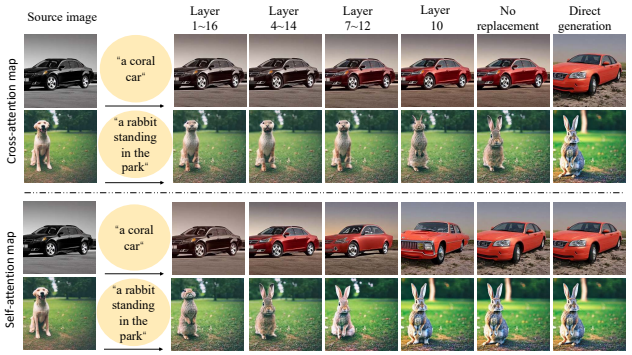

Figure 4. Results of cross-attention and self-attention map replacements in difference layers of the diffusion model.

Class	Layer 3	Layer 6	Layer 9	Layer 10	Layer 12	Layer 14	Layer 16	Avg.
dog	0.53	0.60	0.78	0.60	0.53	0.47	0.38	0.55
horse	0.50	0.70	0.82	0.65	0.68	0.53	0.28	0.59
sheep	0.53	0.45	0.25	0.45	0.62	0.53	0.25	0.44
leopard	0.47	0.65	0.57	0.60	0.47	0.65	0.60	0.57
tiger	0.23	0.12	0.55	0.20	0.45	0.42	0.53	0.36
green	0.00	0.00	0.05	0.00	0.05	0.00	0.12	0.03
white	0.00	0.05	0.30	0.55	0.03	0.15	0.25	0.19
orange	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
yellow	0.00	0.42	0.07	0.05	0.00	0.30	0.07	0.13
red	0.00	0.15	0.28	0.20	0.00	0.20	0.10	0.13

Table 2. Probing accuracy of self-attention map in difference layers.

self-attention map generated from images containing color prompts. For animals, the results are better, although not as precise as those using cross-attention maps. This discrepancy may be attributed to the irregular spatial structure present in the self-attention map corresponding to the color prompt. Conversely, the self-attention map corresponding to the animal prompt contains structural information of different animals, enabling the learning of category information through recognizing structural or contour features. As shown in the lower part of Figure 3, the first component of the horse's self-attention map clearly expresses the outline information of the horse. The lower part of Figure 4 showcases our experimental results of operating on the self-attention map across different attention layers. When the self-attention map of all layers in the source image is replaced during the generation process of the target image, the resulting target image retains

all the structural information from the original image but hinders successful editing. Conversely, if we do not replace the self-attention map, we obtain an image identical to that generated directly using the target prompt. As a compromise, replacing the self-attention map in Layers 4 to 14 allows for preserving the structural information of the original image to the greatest extent while ensuring successful editing. This experimental result further supports the idea that the self-attention map in Layers 4 to 14 does not serve as a reliable category representation but does contain valuable spatial structure information of the image.

3.6. Probing Results for Other Tokens

Do cross-attention maps corresponding to non-edited words contain category information? Furthermore, we explore the attention maps associated with non-edited words. This is relevant because within a text sequence, the text embedding for each word retains the contextual information of the sentence, particularly when a transformer-based text encoder [7, 23] is utilized. We employ the prompt data in the format of *a <color> car* for our probing experiments. The experimental results are presented in Table 3. The findings demonstrate that the article "a" does not encompass any category information of color. In contrast, the noun "car," when modified by the color adjective, does contain color category information. Consequently, if we replace the cross-attention map corresponding to a non-edited word with the cross-attention map of the target image, color information may be introduced, ultimately resulting in editing failures. This observation is also evident from the experimental results in Figure 5, where replacing the cross-attention maps of non-edited words likewise leads to editing failures.

Class	Layer 3	Layer 6	Layer 9	Layer 10	Layer 12	Layer 14	Layer 16	Avg.
green	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
white	0.20	0.00	0.70	0.00	0.72	0.30	0.05	0.28
orange	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
yellow	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.02
red	0.50	0.00	0.82	0.00	0.00	0.00	0.00	0.19
green	0.67	0.67	0.00	0.00	0.00	0.50	0.00	0.26
white	0.33	1.00	0.83	0.58	0.00	0.00	1.00	0.54
orange	0.60	1.00	0.80	1.00	1.00	0.40	0.80	0.80
yellow	0.50	0.25	0.00	0.00	0.12	0.25	0.00	0.16
red	0.38	0.88	0.75	0.12	0.12	0.38	0.00	0.38

Table 3. Probing analysis of cross-attention maps w.r.t. difference tokens. The upper part shows the classification results corresponding to the token "a", and the lower shows results for "car".

4. Our Approach

Based on our exploration of attention layers, we propose a more straightforward yet more stable and efficient approach named Free-Prompt-Editing (FPE). Let I_{src} be the image to be edited. Our goal is to synthesize a new desired image I_{dst} based on the target prompt P_{dst} while preserving the content and structure of the original image I_{src}. Current editing

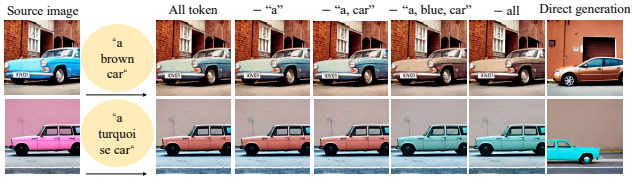

Figure 5. Editing results on replacing attention maps of different tokens in a prompt. "-" is a minus sign. - *"a"* represents subtracting the cross-attention map corresponding to "a".

methods like P2P [10] replace the cross-attention map in the source and target image generation process. This requires modifying the original prompt to find the corresponding attention map for replacement. However, this limitation prevents the direct application of P2P to editing real images, as they do not come with an original prompt.

Based on our exploration of attention layers, our core idea is to combine the layout and contents of I_{src} with the semantic information synthesized with the target prompt P_{dst} to synthesize the desired image I_{dst} that retains the structure and content information of the original image I_{src}. To achieve this, we adapt the self-attention hijack mechanism in the diffusion model's attention layers 4 to 14 during the denoising process between the source and target images. For generated image editing, we substitute the target image's self-attention map with the source image's self-attention map during the diffusion denoising process. When working with actual images, we first obtain the necessary latents for reconstructing the real image by employing the inversion operation [33]. Subsequently, during the editing process, we replace the self-attention map of the real image within the generation process of the target image. We can accomplish the TIE task for the following reasons: 1) the cross-attention mechanism [26, 35] facilitates the fusion of the synthetic image and the target prompt, allowing the target prompt and the image to be automatically aligned even without introducing the cross-attention map of the source prompt; 2) the self-attention map contains spatial layout and shape details of the source image, and the self-attention mechanism [35] allows for the injection of structural information from the original image into the generated target image. Algorithms 1 and 2 present the pseudocode for our simplified method applied to generated and real images, respectively. FPE can also be combined with null text inversion for real image editing (refer to Section 10 in the Supplementary Material).

5. Experiments

5.1. Experimental Settings

Since there are no publicly available datasets specifically designed to verify the effectiveness of image editing algorithms, we construct two types of image-prompt pairs datasets: one

Algorithm 1 Free-Prompt-Editing for a generated image.

Input: P_{src}: a source prompt; P_{dst}: a target prompt; S: random seed;
Output: I_{src}: source image; I_{dst}: edited image;
1: $z_T \sim \mathcal{N}(0,1)$, a unit Gaussian random value sampled with random seed S;
2: $z_T^* \leftarrow z_T$;
3: **for** $t = T, T-1, \ldots, 1$ **do**
4: $\quad z_{t-1}, M_{self} \leftarrow DM(z_t, P_{src}, t)$;
5: $\quad z_{t-1}^* \leftarrow DM(z_t^*, P_{dst}, t)\{M_{self}^* \leftarrow M_{self}\}$;
6: **end for**
7: **Return** $(I_{src} \leftarrow Decoder(z_0), I_{dst} \leftarrow Decoder(z_0^*))$;

Algorithm 2 Free-Prompt-Editing for a real image.

Input: P_{dst}: a target prompt; I_{src}: real image;
Output: I_{dst}: edited image; I_{res}: reconstructed image;
1: $\{z_t\}_{t=0}^{T} \leftarrow DDIM - inv(I_{src})$;
2: $z_T^* \leftarrow z_T$;
3: **for** $t = T, T-1, \ldots, 1$ **do**
4: $\quad z_{t-1}, M_{self} \leftarrow DM(z_t, t)$;
5: $\quad z_{t-1}^* \leftarrow DM(z_t^*, P_{dst}, t)\{M_{self}^* \leftarrow M_{self}\}$;
6: **end for**
7: **Return** $(I_{res} \leftarrow Decoder(z_0), I_{dst} \leftarrow Decoder(z_0^*))$;

for generated images and one for real images. The generated images dataset includes Car-fake-edit and ImageNet-fake-edit, where Car-fake-edit contains 756 prompt pairs, and ImageNet-fake-edit contains 1182 prompt pairs sampled from FlexIT [4] and ImageNet [28]. The real image datasets include Car-real-edit, sampled from the Stanford Car (CARS196) dataset [15], containing 3321 image-prompt pairs, and ImageNet-real-edit, which contains 1092 pairs. For more details, see section 7.2 in the Supplementary Material. In addition, we also use benchmarks constructed by PnP [34]. These benchmarks contain two datasets: Wild-TI2I and ImageNet-R-TI2I. For generated images, Wild-TI2I contains 70 prompt pairs, and ImageNet-R-TI2I contains 150 pairs. For real images, Wild-TI2I contains 78 image-prompt pairs, and ImageNet-R-TI2I includes 150 pairs.

We utilize Clip Score (CS) and Clip Directional Similarity (CDS) [9, 23] to quantitatively analyze and compare our method with currently popular image editing algorithms. The underlying model for our experiments is Stable Diffusion 1.5[2]. The experimental results of comparative methods are produced using the publicly disclosed codes from their original papers with unified random seeds.

5.2. Image Editing Results

We evaluate our method through quantitative and qualitative analyses. As illustrated in Figure 6, we showcase the editing outcomes of our method, demonstrating that it successfully transforms various attributes, styles, scenes, and categories of the original images.

Figure 6. Results of our method on image-text pairs from Wild-TI2I and ImageNet-R-TI2I.

5.2.1 Comparison to Prior/Concurrent Work

In this section, we compare our work with state-of-the-art image editing methods, including (i) P2P [10] (with null text inversion [18] for the real image scene), (ii) PnP [34], (iii) SDEdit [17] under two noise levels (0.5 and 0.75), (iv) DiffEdit [5], (v) MasaCtrl [2], (vi) Pix2pixzero [22], (vii) Shape-guided [21], and (viii) InstructPix2Pix [1]. We further present the image editing results using other Stable Diffusion-based models to demonstrate the universality of our method, including Realistic-V2[3], Deliberate[4], and Anything-V4[5].

Comparison to P2P We first compare our method with P2P [10] for synthetic image editing scenes and P2P combined with null text inversion [18] for real image scenes, both denoted as P2P. The experimental results are shown in Figure 7 and Table 4. In Figure 7, it is evident that when performing color transformation on a real image by modifying the cross-attention map, the editing fails. The editing

[2] https://huggingface.co/runwayml/stable-diffusion-v1-5

[3] https://huggingface.co/SG161222/Realistic_Vision_V2.0

[4] https://huggingface.co/XpucT/Deliberate

[5] https://huggingface.co/xyn-ai/anything-v4.0

results of P2P for car color tend to replicate the color (white) of the original image. Regarding the category conversion results for generated images, we observe that while P2P can accurately transform different animals, the edited results still retain appearances of sheep. This leads to an incomplete conversion for patterned animals such as giraffes, leopards, and tigers. Unlike P2P, our method operates only at the self-attention layers and is not susceptible to editing failures caused by modifications to the cross-attention map.

Dataset	CS ↑		CDS ↑	
	P2P	Ours	P2P	Ours
Car-fake-edit	25.96	**26.02**	0.2451	**0.2659**
Car-real-edit	24.64	**24.85**	0.2288	**0.2605**
ImageNet-fake-edit	27.42	**27.80**	0.2401	**0.2560**
ImageNet-real-edit	26.17	**26.35**	0.2426	**0.2468**

Table 4. Quantitative experimental results over Car-fake-edit, ImageNet-fake-edit, Car-real-edit and ImageNet-real-edit.

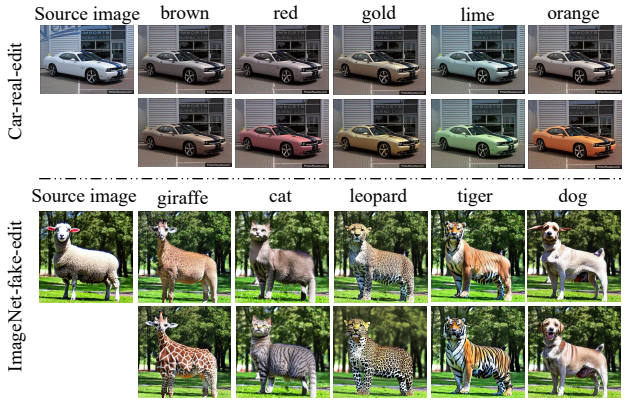

Figure 7. Comparisons results with P2P [10] on Car-real-edit and ImageNet-fake-edit. Upper part: P2P. Lower part: ours.

Comparison to Other Methods Further, we compare our method with other state-of-the-art (SOTA) image editing methods [1, 2, 5, 10, 17, 21, 22, 34] over the Wild-TI2I and ImageNet-R-TI2I benchmarks. The experimental results are presented in Figure 8 and Table 5. As shown in Figure 8, our method successfully converts different inputs for both real and synthetic images. In all examples, our method achieves high-fidelity editing that aligns with the target prompt while preserving the original image's structural information to the greatest extent possible. In contrast, SDEdit and InstructPix2Pix struggle to preserve the structural information of the original image. SDEdit aligns the editing results better with the target prompt when there is high-level noise but fails in the presence of low-level noise. InstructPix2Pix retains consistency with the target prompt but loses the original structural information. DiffEdit and Pix2pix-zero also struggle to perform better editing based on the target prompt. Similarly, PnP achieves good editing results, but it is a two-step method that leads to significant computational overhead; editing a single image in a generated image editing scenario takes approximately 335.65 seconds. In contrast, our method only requires around 6.30 seconds on an A100 GPU with 40GB memory, as Table 5 indicates.

Table 5 presents the quantitative experimental results of different editing algorithms on the Wild-TI2I and ImageNet-R-TI2I benchmarks. From Table 5, it is evident that our method outperforms all others in terms of the CDS metric. This indicates that our method excels in preserving the spatial structure of the original image and performing editing according to the requirements of the target prompt, yielding superior results. Meanwhile, our method achieves a good balance between time consumption and effectiveness, as demonstrated in Table 5.

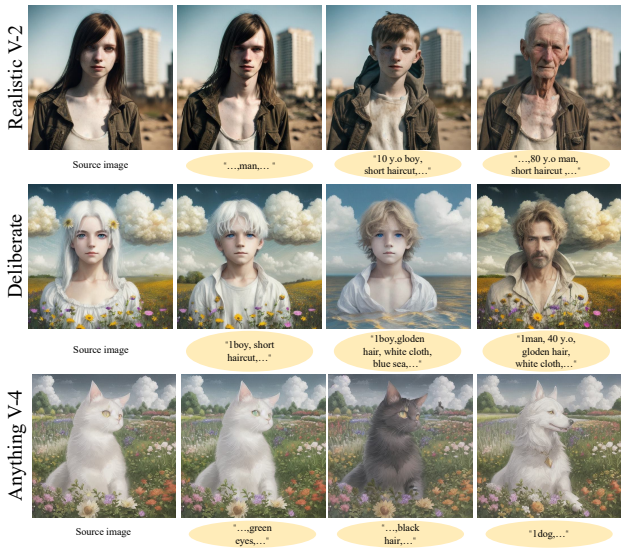

Figure 9. Experimental results of our method using other TIS models, including Realistic-V2, Deliberate and Anything-V4.

5.2.2 Results in Other TIS Models

We have applied our method to other TIS models based on Stable Diffusion-style frameworks to demonstrate its transferability. Figure 9 showcases the editing results of our method on Realistic-V2, Deliberate, and Anything-V4 TIS models. From these results, it can be observed that our method is capable of effectively editing images on other diffusion models as well. For example, it can transform a girl into a boy, change a boy's age to 10 or 80, modify hairstyles, change hair colors, alter backgrounds, and switch categories.

5.3. Limitations and Discussion

Although our method employs probe analysis to elucidate the role of attention layers in the TIS model and proposes a novel method for editing images in multiple scenarios

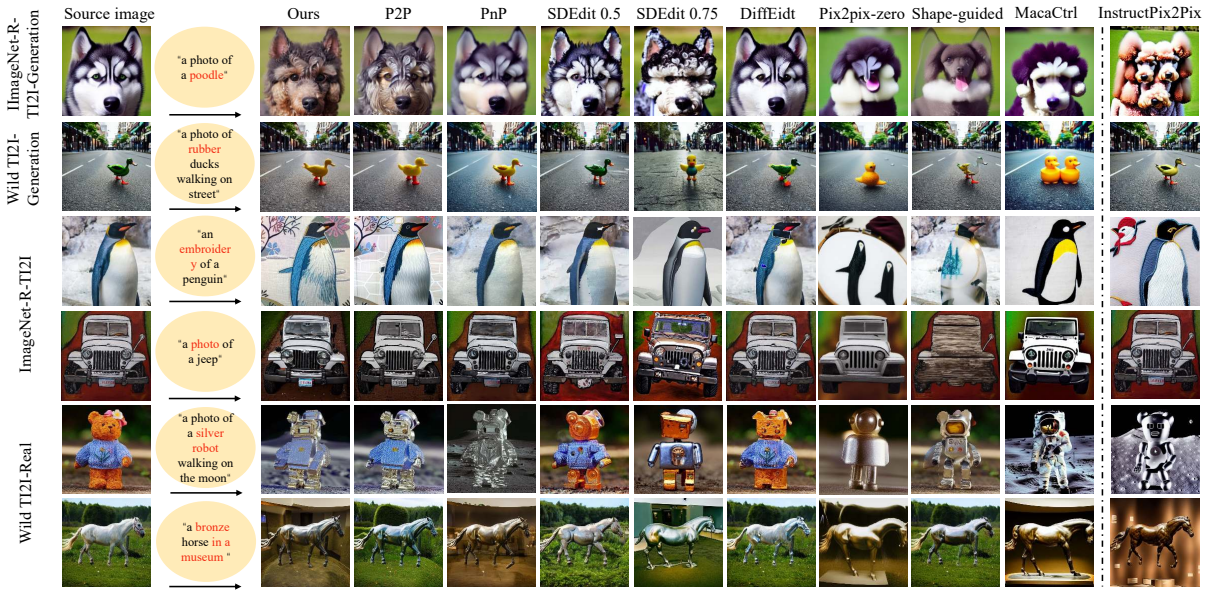

Figure 8. Comparison to prior works. Left to right: source image, target prompt, our result, P2P [10], PnP [34], SDEdit [17] w/ two noising levels, DiffEdit [5], Pix2pixzero [22], Shape-guided [21], MasaCtrl [2] and InstructPix2Pix [1] (fine-tuning based method.)

Method	ImageNet-R-TI2I fake		ImageNet-R-TI2I real		Wild-fake		Wild-real		Editing Times (s)	
	CS ↑	CDS ↑	CS ↑	CDS ↑	CS ↑	CDS ↑	CS ↑	CDS ↑	fake ↓	real ↓
SDEdit (0.5)	-	-	28.37	0.1415	-	-	27.48	0.1220	-	**2.59**
SDEdit (0.75)	-	-	30.17	0.2171	-	-	**29.79**	0.2007	-	3.35
Shape-Guided	-	-	26.01	0.1090	-	-	26.53	0.1330	-	16.02
DiffEdit	26.68	0.0748	26.50	0.0909	25.59	0.0794	26.33	0.0879	9.02	4.85
Pix2pixzero	27.94	0.2271	28.96	0.1415	28.19	0.2864	29.55	0.1462	24.92	36.76
P2P	28.88	0.3394	28.56	0.2146	27.85	0.2796	28.42	0.1930	6.41	55.32
PnP	28.83	0.2318	28.76	0.2073	28.20	0.2838	28.46	0.2020	335.65	384.26
MasaCtrl	29.66	0.3024	**31.40**	0.2170	**29.96**	**0.3474**	29.33	0.2101	**6.18**	10.90
Ours	**29.79**	**0.3559**	29.05	**0.2271**	27.88	0.3116	29.04	**0.2234**	6.30	10.75

Table 5. Quantitative experimental results over Wild-TI2I and ImageNet-R-TI2I benchmarks, including real and generated guidance images. CS: Clip score [23] and CDS: Clip Directional Similarity [9, 23]. Editing Times: per-image/second

without complex operations, it still has some limitations. Firstly, our method is constrained by the generative capabilities of the TIS model. Our editing method will fail if the generative model cannot produce images consistent with the target prompt description. When editing real images, the original image must first be reconstructed. Some detailed information, especially facial details, may be lost during the reconstruction process, primarily due to the limitations of the VQ autoencoder [14]. Optimizing the VQ autoencoder is beyond the scope of this paper, as our objective is to provide a simple and universal editing framework. Addressing these challenges will be part of our future work.

6. Conclusion

In this work, we utilized probe analysis and conducted experiments to elucidate the following insights on TIS models: the cross-attention map carries the semantic information of the prompt, which leads to the ineffectiveness of image editing methods that rely on it. On the contrary, the self-attention map captures the spatial structural information of the original image, playing an essential role in preserving the image's inherent structure during editing. Based on our comprehensive analysis and empirical evidence, we have streamlined current image editing algorithms and proposed an innovative image editing approach. Our approach does not require additional tuning or the alignment of target and source prompts to achieve effective object or background editing in images. In extensive experiments across multiple datasets, our simplified method has outperformed existing image editing algorithms. Furthermore, our algorithm can be seamlessly adapted to other TIS models.

Acknowledgements This work is partially supported by Alibaba Cloud through the Research Talent Program with South China University of Technology, and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (No. 2017ZT07X183).

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 2, 3, 6, 7, 8

[2] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 1, 2, 6, 7, 8

[3] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286. Association for Computational Linguistics, 2019. 2, 3

[4] Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Flexit: Towards flexible semantic image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18270–18279, 2022. 6, 2

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 6, 7, 8

[6] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 5

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[9] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 6, 8

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 5, 6, 7, 8

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[13] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. 1, 8

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6, 1

[16] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094. Association for Computational Linguistics, 2019. 2, 3

[17] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 1, 2, 6, 7, 8

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 1, 2, 6, 4, 5

[19] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[20] OpenAI. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023. 1

[21] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-guided diffusion with inside-outside attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4198–4207, 2024. 1, 6, 7, 8

[22] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 1, 6, 7, 8

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 5, 6, 8

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

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 5

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 2

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 6, 2

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[30] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[32] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 5

[34] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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*, pages 1921–1930, 2023. 1, 2, 6, 7, 8

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[36] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003. 4

[37] Chengyu Wang, Zhongjie Duan, Bingyan Liu, Xinyi Zou, Cen Chen, Kui Jia, and Jun Huang. Pai-diffusion: Constructing and serving a family of open chinese diffusion models for text-to-image synthesis on the cloud. *arXiv preprint arXiv:2309.05534*, 2023. 1

[38] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022. 1

[39] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2