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

Supplementary Material

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

The Supplementary Material of this paper consists of several sections that provide additional information and support for the main content.

- **Details of Data Collection:** Describes the data collection process used in the probing analysis and image editing experiments.
- **Probing Analysis Results:** Presents the complete results of the probing experiments, including supplementary experiments.
- **Impact of Replace Steps:** Provides ablation experimental results on replacing the self-attention map and cross-attention map at different denoising steps.
- **Real Image Editing with Null-text Inversion:** Introduces the method of using null-text inversion for editing real images and displays some experimental results.

7. Details of Data Collection

7.1. Data Collection for Probing Analysis

Cross-Attention Map Data: We have constructed five datasets by saving the cross-attention maps of the target word in the prompt, each containing 2,000 samples, for cross-attention map analysis. The prompts include color adjectives and animal nouns. Specifically, for color adjectives, we used two types of prompts: *"a <color> car"* and *"a <color> <object>"*, to construct the data. The prompt *"a <color> car"* consists of ten categories of color words using 200 random seeds. We obtained the cross-attention maps by averaging the steps during the generation process. Each cross-attention map consists of 16 attention maps corresponding to the 16 attention layers of the diffusion model. The same procedure was followed for the construction of cross-attention maps in the format *"a <color> <object>"*, but with two random seeds and 100 everyday objects. Similarly, we used the prompt format *"a <animal> standing in the park"*, sampled 200 random seeds, and constructed 2,000 samples for animal nouns. For the cross-attention maps corresponding to non-editing words, we used the prompt format *"a <color> car"* and saved the cross-attention maps for the words "a" and "car." The same approach was applied to the complex text templates.

Self-Attention Map Data: For the self-attention maps, we constructed two datasets, each containing 2,000 samples, using prompt formats *"a <color> car"* and *"a <animal> standing in the park"* and sampling 200 random seeds. However, due to the large size of the self-attention maps, which are 4096×4096, 1024×1024, 256×256, 64×64, and 8×8 in dimensions, we resized the layers with dimensions larger than or equal to 256×256 to 256×256 for the probing analysis experiments.

The 10 color, 10 animal, 100 object categories, and 12 more complex text templates are as follows:

```
color_words = [
    "red", "blue", "green", "yellow", "brown",
    "pink", "purple", "black", "white", "orange"
]
animal_list = [
    "dog", "giraffe", "horse", "lion", "rabbit",
    "sheep", "cat", "monkey", "leopard", "tiger"
]
object_list = [
    "apple", "banana", "carrot", "dog", "flower",
    "giraffe", "hat", "car", "train", "bicycle",
    ...
    "yak", "acorn", "bear", "caterpillar", "turtle",
    "dandelion", "elephant", "feather", "grape", "hedgehog",
    "inchworm", "jackfruit", "kiwi", "lemon", "zebra",
    "mushroom", "otter", "peacock", "rose", "strawberry",
    "volcano", "watermelon", "xenops", "yucca", "cup"
]
complex_templates = [
    "a photo of a {} car and a dog",
    "a photo of {} car",
    "the painting of a {} car",
    ...
    "a cool painting of an old {} car",
    "a man and a {} car",
    "a {} car and a dog",
]
```

7.2. Data Collection for Editing Experiments

Car-fake-edit: Using the prompt format *"a <color> car"* and 28 color words, we constructed Car-fake-edit, which contains 756 prompt pairs.

Car-real-edit: We sampled 123 real images from the Stanford Car dataset [15] with image sizes ranging from 512 to 768. We then used CLIP [23] to align these images with the 28 color words, resulting in the source prompts for the

original images in the format *"a <color> car"*. We constructed 27 target prompts for each image, resulting in 3,321 image-text pairs.

ImageNet-fake-edit: The paper FlexIT [4] proposes constructing a validation set from the ImageNet [28] validation set. The method for constructing the test set is as follows: A subset of 273 labeled categories is taken from ImageNet, and these categories are manually divided into 47 clusters. During testing, only transformations within the same cluster are allowed; for example, a cat to a dog, but not a laptop to a butterfly. For each label T, eight random categories are sampled from the cluster to serve as queries, resulting in 2,184 queries for the 273 categories. We utilized their test dataset, which consists of 1,092 queries, and added ten animal categories to construct 1,182 prompt pairs using prompt templates.

ImageNet-real-edit: Based on ImageNet-fake-edit, we used the prompt format "a photo of a/an {}" and constructed 1,092 image-text pairs for the real images. The color words, animal list, and prompt templates are as follows:

```
color_words = [
    "red", "green", "blue", "yellow",
    "orange", "purple", "pink", "black",
    "white", "gray", "brown", "beige",
    "cyan", "magenta", "teal", "lime",
    "olive", "navy", "maroon", "silver",
    "gold", "bronze", "peach", "coral",
    "indigo", "violet", "turquoise", "chocolate"
]
animal_list = [
    "dog", "giraffe", "horse", "lion", "rabbit",
    "sheep", "cat", "monkey", "leopard", "tiger"
]
imagenet_templates = [
    "a photo of a {}",
    "a rendering of a {}",
    "a cropped photo of the {}",
    "the photo of a {}",
    "a photo of a clean {}",
    "a photo of a dirty {}",
    "a dark photo of the {}",
    "a photo of my {}",
    "a photo of the cool {}",
    "a close-up photo of a {}",
    "a bright photo of the {}",
    "a cropped photo of a {}",
    "a photo of the {}",
    "a good photo of the {}",
    "a photo of one {}",
    "a close-up photo of the {}",
    "a rendition of the {}",
    "a photo of the clean {}",
    "a rendition of a {}",
    "a photo of a nice {}",
    "a good photo of a {}",
    "a photo of the nice {}",
    "a photo of the small {}",
    "a photo of the weird {}",
    "a photo of the large {}",
    "a photo of a cool {}",
    "a photo of a small {}",
    "a {} in the park"
]
```

8. Probing Analysis Results

In this section, we present the complete experimental results for probe analysis and supplementary experiments, shown in Tables 6, 7, 8 and 9.

Table 6 presents the probe analysis experiment results for cross-attention maps with four sub-tables. The first three sub-tables, from top to bottom, correspond to the prompt formats *"a/an <animal> standing in the park," "a <color><object>,"* and *"a <color> car."* The bottom sub-table corresponds to the experiment with training data *"a <color> car"* and test data *"a <color><object>."*

For the experiments within the distribution, regardless of the prompt format, the cross-attention maps corresponding to the words can be accurately classified by the classifier. Even on out-of-distribution data, an average accuracy of around 50% can be achieved. This indicates that the diffusion model's 16 layers of cross-attention maps can serve as good feature representations, containing semantic information about the corresponding words. This is consistent with the conclusion in the main text that the cross-attention maps are both weight matrices and rich in semantic information. Table 7 presents the probe experiment results for non-target words, aiming to verify whether the attention maps corresponding to words other than the target word in the prompt contain the semantics of the target word. We conducted experiments using the simple prompt format *"a <color> car."*

Table 8 presents the complete probe experiment results for self-attention maps. Compared to cross-attention maps, self-attention maps are not directly usable as feature representations for classification, especially for prompts with color adjectives where the classification performance could be improved. However, compared to color adjective prompts, higher classification accuracy is observed for prompts with animal adjectives, which may be related to the self-attention maps' ability to represent objects' appearance contours in animal class images.

We expanded our investigation through probing analyses utilizing a set of twelve intricate text templates to eliminate the potential experimental bias that may arise from using consistently simple and regular templates in previous experiments. Examples of these templates include phrases such as *"a painting of a wooden <color> car"* and *"a photo of a <color> car and a dog"*, among others. The findings, as presented in Table 9, corroborate the conclusions drawn from experiments using simpler text prompts, indicating consistency in results across varying levels of template complexity.

9. Impact of Replacement Steps

In this section, we conduct ablation experiments on different attention layers of cross-attention and self-attention maps

Class	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12	Layer 13	Layer 14	Layer 15	Layer 16	Avg.
dog	0.97	0.92	1.00	0.89	0.97	1.00	1.00	1.00	1.00	1.00	0.54	0.89	1.00	0.76	0.68	1.00	0.91
giraffe	0.94	0.97	1.00	0.97	1.00	1.00	1.00	1.00	0.92	0.97	0.97	1.00	0.97	0.92	0.92	0.89	0.97
horse	0.98	0.98	0.96	0.84	0.82	1.00	1.00	1.00	1.00	1.00	0.64	0.64	0.93	1.00	0.89	0.91	0.91
lion	0.98	1.00	1.00	1.00	1.00	1.00	0.93	0.93	0.95	0.98	0.66	0.90	0.98	0.93	0.98	0.98	0.95
rabbit	0.85	0.87	1.00	0.98	1.00	1.00	0.96	0.96	0.98	0.91	0.49	0.98	0.96	0.74	0.96	1.00	0.91
sheep	0.93	0.82	0.97	1.00	0.95	1.00	0.97	1.00	1.00	1.00	0.97	1.00	0.93	0.90	0.80	0.97	0.95
cat	0.95	0.95	0.90	0.95	0.88	1.00	1.00	0.95	0.95	0.82	0.85	0.78	0.97	0.75	0.70	0.78	0.89
monkey	1.00	1.00	1.00	1.00	0.95	1.00	0.98	1.00	1.00	1.00	0.67	1.00	1.00	1.00	0.95	1.00	0.97
leopard	1.00	0.89	0.97	1.00	0.97	1.00	0.97	0.95	1.00	1.00	0.76	0.97	0.97	0.79	0.97	0.87	0.94
tiger	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.97	1.00	0.91	0.88	1.00	1.00	1.00	0.97	0.98
green	0.91	0.84	0.93	0.89	0.98	0.91	0.96	0.96	0.91	0.96	0.93	0.67	0.98	0.38	0.47	0.60	0.83
white	0.88	0.15	0.97	0.91	1.00	1.00	1.00	0.97	0.94	0.97	0.97	0.97	1.00	0.61	0.88	0.85	0.88
purple	0.98	0.93	1.00	0.98	1.00	0.98	0.63	0.98	0.84	0.93	0.86	0.95	1.00	0.51	0.60	0.44	0.85
brown	0.97	0.82	0.97	0.95	1.00	0.92	0.95	0.95	0.89	0.92	0.95	0.97	0.95	0.34	0.84	0.79	0.89
blue	0.92	0.65	0.86	0.92	1.00	1.00	1.00	1.00	0.95	0.92	0.86	0.95	0.97	0.19	0.62	0.57	0.84
orange	0.94	0.72	0.97	1.00	1.00	1.00	0.92	0.89	0.94	0.92	0.86	0.89	0.94	0.94	1.00	0.83	0.92
black	0.93	0.80	0.85	0.78	1.00	0.95	0.97	0.90	0.95	0.93	0.60	0.75	1.00	0.65	0.85	0.82	0.86
pink	0.93	0.54	0.98	0.93	1.00	1.00	0.95	0.95	1.00	1.00	1.00	0.51	1.00	0.46	0.51	0.88	0.85
yellow	0.96	0.51	0.96	0.89	0.96	0.77	0.98	0.96	1.00	0.98	0.91	1.00	0.98	0.36	0.81	0.68	0.86
red	0.85	0.85	0.97	0.90	1.00	0.97	0.60	0.97	0.93	0.85	0.90	0.70	1.00	0.23	0.80	0.65	0.82
green	1.00	1.00	0.97	0.97	1.00	0.97	0.47	0.68	0.80	0.93	1.00	0.97	0.97	0.50	0.75	0.95	0.87
white	1.00	0.89	1.00	0.91	1.00	0.82	0.18	0.64	0.44	0.84	0.60	0.73	0.98	0.56	0.69	0.89	0.76
purple	1.00	1.00	1.00	1.00	1.00	1.00	0.77	0.69	0.90	0.87	0.05	0.97	0.97	0.67	0.90	1.00	0.86
brown	1.00	1.00	0.97	1.00	0.97	0.78	0.62	0.34	0.09	0.66	0.97	0.97	0.94	0.84	0.75	1.00	0.81
blue	1.00	0.97	0.97	0.94	1.00	1.00	0.94	0.89	0.72	1.00	0.42	0.97	1.00	0.86	0.94	0.94	0.91
orange	1.00	1.00	0.95	1.00	1.00	0.97	0.84	0.97	0.82	0.92	0.89	1.00	1.00	1.00	1.00	1.00	0.96
black	1.00	0.95	0.97	0.85	1.00	1.00	0.97	0.77	0.79	0.95	0.08	0.95	1.00	0.85	0.97	0.85	0.87
pink	1.00	0.95	0.95	0.95	1.00	1.00	0.95	0.88	0.76	0.90	1.00	0.95	1.00	0.71	0.90	0.95	0.93
yellow	1.00	1.00	1.00	1.00	1.00	1.00	0.74	0.88	0.85	0.97	0.74	0.94	1.00	0.76	0.82	0.74	0.90
red	1.00	1.00	1.00	1.00	1.00	0.95	0.15	0.87	0.87	0.74	0.44	0.95	1.00	0.95	0.90	0.79	0.85
green	0.82	0.58	0.84	0.13	0.84	0.16	0.55	0.03	0.16	0.08	0.92	0.42	0.34	0.29	0.29	0.84	0.46
white	0.74	0.11	0.42	0.82	0.16	0.18	0.39	0.42	0.68	0.63	0.58	1.00	0.63	0.32	0.08	0.89	0.50
purple	0.53	0.11	0.79	0.47	0.84	1.00	0.26	0.08	0.08	0.24	0.00	0.92	0.95	0.08	0.13	0.29	0.42
brown	0.97	0.68	0.74	0.66	0.84	0.18	0.13	0.08	0.00	0.03	0.92	0.76	0.68	0.45	0.61	0.18	0.50
blue	0.87	0.68	0.11	0.13	1.00	0.61	0.05	0.32	0.16	0.18	0.05	0.34	0.87	0.03	0.47	0.34	0.39
orange	0.87	0.24	0.89	0.95	0.50	0.55	0.45	0.13	0.18	0.34	0.08	0.92	0.76	0.26	0.92	0.34	0.52
black	0.50	0.82	0.68	0.26	0.55	0.82	0.84	0.00	0.42	0.68	0.00	0.05	0.95	0.45	0.61	0.79	0.53
pink	0.58	0.16	0.32	0.66	0.76	0.32	0.34	0.82	0.61	0.11	1.00	0.11	0.95	0.18	0.61	0.97	0.53
yellow	0.16	0.08	0.66	0.68	0.55	0.05	0.55	0.26	0.63	0.71	0.76	0.79	0.79	0.42	0.16	0.39	0.48
red	0.74	0.05	0.84	0.74	0.82	0.61	0.00	0.32	0.84	0.42	0.63	0.61	0.97	0.76	0.45	0.34	0.57

Table 6. Complete experimental results of attribute mining using cross-attention maps in the diffusion model are shown. The table displays the probing accuracy results for various prompt formats: *"a/an <animal> standing in the park," "a <color> car," "a <color><object>"* and involves training on data in the format *"a <color> car"* and testing on data in the format *"a <color><object>."* The results are presented from top to bottom in the corresponding table.

Class	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12	Layer 13	Layer 14	Layer 15	Layer 16	Avg.
green	0.35	0.80	0.03	0.00	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.11
white	0.03	0.00	0.20	0.47	0.00	0.00	0.00	0.00	0.70	0.00	0.00	0.72	0.00	0.30	0.07	0.05	0.16
purple	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00
brown	0.82	0.00	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.15	0.97	0.20
blue	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.07	0.03	0.20	0.00	0.00	0.00	0.02
orange	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
black	0.03	0.03	0.07	0.82	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.60	0.85	0.20	0.88	0.05	0.41
pink	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.03	0.00	0.01
yellow	0.00	0.03	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.01
red	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.93	0.82	0.00	1.00	0.00	0.35	0.00	0.05	0.00	0.26
green	0.67	0.00	0.67	0.00	0.17	0.67	0.33	0.33	0.00	0.00	0.17	0.00	0.17	0.50	0.17	0.00	0.24
white	0.83	0.00	0.33	0.00	0.00	1.00	0.50	0.08	0.83	0.58	0.08	0.00	0.92	0.00	0.83	1.00	0.44
purple	0.75	0.00	0.25	0.00	0.50	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.50	0.25	0.00	0.00	0.15
brown	0.40	1.00	0.40	0.80	0.00	0.00	0.20	0.70	0.20	0.00	0.00	0.90	0.60	0.00	0.70	0.00	0.37
blue	0.44	0.00	0.22	0.00	0.11	0.33	0.00	0.22	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
orange	0.40	0.40	0.60	1.00	0.80	1.00	1.00	0.80	0.80	1.00	1.00	1.00	1.00	0.40	0.80	0.80	0.80
black	1.00	0.00	0.83	0.17	0.67	0.33	0.67	1.00	0.50	0.50	0.00	0.00	0.17	0.00	0.00	0.83	0.42
pink	0.12	0.00	0.38	0.38	0.12	0.12	0.38	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.00	0.20
yellow	0.62	0.00	0.50	0.00	0.12	0.25	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.25	0.00	0.00	0.12
red	0.25	0.00	0.38	0.12	0.62	0.88	0.62	1.00	0.75	0.12	0.00	0.12	0.62	0.38	0.00	0.00	0.37

Table 7. Complete experimental results of attribute mining using cross-attention map in non-edit tokens.

under various denoising steps. The experimental results are presented in Figure 10 and Figure 11.

When only the cross-attention map is replaced during editing, the target image loses the structural information of the original image. As shown in Figure 10, although the leopard in the target image bears a resemblance to the dog in the original image, notable modifications in the background, particularly the grass, are observed. Similarly, when performing a color conversion on a car using only the cross-attention map, the original image's structural information is lost, lead-

Class	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12	Layer 13	Layer 14	Layer 15	Layer 16	Avg.
dog	0.47	0.38	0.53	0.23	0.55	0.60	0.50	0.53	0.78	0.60	0.57	0.53	0.57	0.47	0.38	0.38	0.50
giraffe	0.28	0.47	0.47	0.60	0.68	0.62	0.75	0.60	0.78	0.72	0.62	0.57	0.55	0.68	0.42	0.57	0.59
horse	0.12	0.20	0.50	0.38	0.70	0.70	0.40	0.80	0.82	0.65	0.70	0.68	0.55	0.53	0.30	0.28	0.52
lion	0.07	0.17	0.38	0.38	0.20	0.15	0.35	0.28	0.17	0.23	0.40	0.47	0.42	0.33	0.35	0.25	0.29
rabbit	0.30	0.30	0.25	0.33	0.23	0.20	0.25	0.17	0.20	0.23	0.23	0.20	0.30	0.42	0.28	0.28	0.26
sheep	0.17	0.23	0.53	0.55	0.33	0.45	0.07	0.38	0.25	0.45	0.55	0.62	0.42	0.53	0.38	0.25	0.38
cat	0.07	0.17	0.07	0.17	0.20	0.15	0.12	0.23	0.28	0.12	0.25	0.33	0.25	0.17	0.23	0.15	0.19
monkey	0.33	0.33	0.28	0.38	0.28	0.38	0.07	0.62	0.38	0.45	0.28	0.30	0.33	0.33	0.33	0.33	0.33
leopard	0.72	0.70	0.47	0.62	0.57	0.65	0.38	0.42	0.57	0.60	0.47	0.47	0.42	0.65	0.62	0.60	0.56
tiger	0.38	0.38	0.23	0.42	0.35	0.12	0.23	0.28	0.55	0.20	0.53	0.45	0.35	0.42	0.50	0.53	0.37
green	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.00	0.12	0.05	0.00	0.00	0.00	0.12	0.03
white	0.00	0.33	0.00	0.00	0.53	0.05	0.45	0.68	0.30	0.55	0.07	0.03	0.00	0.15	0.00	0.25	0.21
purple	0.00	0.03	0.00	0.00	0.28	0.60	0.00	0.07	0.05	0.35	0.10	0.07	0.03	0.50	0.00	0.00	0.13
brown	0.00	0.10	0.00	0.00	0.00	0.00	0.05	0.03	0.05	0.00	0.17	0.17	0.05	0.03	0.00	0.10	0.05
blue	0.72	0.07	0.57	0.55	0.10	0.00	0.30	0.05	0.12	0.05	0.28	0.33	0.65	0.00	0.35	0.05	0.26
orange	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
black	0.00	0.35	0.03	0.03	0.07	0.00	0.33	0.05	0.15	0.07	0.17	0.20	0.10	0.25	0.10	0.28	0.14
pink	0.35	0.20	0.65	0.55	0.17	0.03	0.00	0.38	0.47	0.28	0.45	0.50	0.57	0.03	0.80	0.35	0.36
yellow	0.00	0.00	0.00	0.00	0.03	0.42	0.00	0.15	0.07	0.05	0.05	0.00	0.03	0.30	0.15	0.07	0.08
red	0.00	0.00	0.00	0.00	0.00	0.15	0.33	0.03	0.28	0.20	0.05	0.00	0.03	0.20	0.00	0.10	0.08

Table 8. Complete experimental results of attribute mining using self-attention maps in the diffusion model.

Class	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12	Layer 13	Layer 14	Layer 15	Layer 16	Avg.
green	0.81	0.74	0.85	0.81	0.91	0.68	0.79	0.85	0.89	0.87	0.94	0.62	0.96	0.43	0.64	0.40	0.76
white	0.82	0.79	0.91	0.91	1.00	0.82	0.97	0.91	0.91	0.73	0.97	0.94	0.97	0.45	0.61	0.55	0.83
purple	0.98	0.65	1.00	0.86	0.98	0.95	0.79	0.93	0.95	0.93	0.91	0.79	0.98	0.47	0.35	0.72	0.83
brown	0.96	0.76	0.98	0.93	0.98	0.80	0.84	0.98	0.96	0.80	0.89	0.96	0.98	0.73	0.82	0.87	0.89
blue	0.89	0.62	0.95	0.59	0.86	0.95	0.70	1.00	0.97	0.62	0.73	0.89	1.00	0.16	0.59	0.68	0.76
orange	1.00	0.61	0.97	0.95	0.92	1.00	0.79	0.87	0.84	0.71	0.79	0.82	1.00	0.47	1.00	0.87	0.85
black	0.93	0.75	0.95	0.88	1.00	0.88	0.95	0.93	0.93	0.82	0.88	0.82	1.00	0.68	0.85	0.93	0.88
pink	0.90	0.02	0.88	0.98	0.93	0.95	0.88	0.85	0.93	0.98	0.95	0.76	0.98	0.27	0.66	0.66	0.79
yellow	0.89	0.56	0.89	0.78	1.00	0.75	0.92	0.94	0.86	0.69	0.81	0.86	0.92	0.33	0.58	0.47	0.77
red	0.93	0.00	0.90	0.75	0.93	0.95	0.88	0.90	0.97	0.80	0.85	0.75	1.00	0.60	0.53	0.70	0.78
green	0.00	0.00	0.00	0.00	0.36	0.51	0.73	0.42	0.60	0.20	0.44	0.00	0.29	0.00	0.18	0.00	0.23
white	0.00	0.00	0.00	0.00	0.03	0.03	0.12	0.03	0.03	0.03	0.06	0.03	0.03	0.03	0.36	0.00	0.05
purple	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.16	0.07	0.00	0.19	0.00	0.03
brown	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.05	0.11	0.08	0.00	0.00	0.00	0.05	0.00	0.02
blue	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
orange	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.42	0.19	0.00	0.00	0.00	0.08	0.00	0.05
black	0.00	0.00	0.00	0.00	0.03	0.03	0.12	0.07	0.07	0.07	0.10	0.00	0.03	0.00	0.03	0.00	0.03
pink	1.00	1.00	1.00	1.00	0.00	0.00	0.09	0.00	0.00	0.02	0.02	0.85	0.00	0.91	0.04	1.00	0.43
yellow	0.00	0.00	0.00	0.00	0.17	0.22	0.02	0.22	0.20	0.39	0.00	0.00	0.17	0.00	0.17	0.00	0.10
red	0.00	0.00	0.00	0.00	0.68	0.55	0.00	0.40	0.17	0.00	0.17	0.00	0.65	0.00	0.10	0.00	0.17

Table 9. Probing accuracy of attention maps in more complex text templates. Upper: cross-attention map, Lower: self-attention map.

ing to a car that lacks its original structure and takes on a brown appearance.

When both the cross-attention map and the self-attention map are replaced simultaneously, the results depicted in Figures 10 and 11 are obtained by keeping the cross-replace ratio fixed at 0.8 while varying the self-replace ratio. The replacement of the cross-attention map aids in swiftly identifying the target region and reconstructing the structure of the original image. However, it also introduces the original image's feature information, particularly when replacing attention maps in all layers, which significantly includes the original features. As illustrated in Figures 10 and 11, the leopard exhibits attributes similar to a dog, while the car retains its blue color.

When only the self-attention map is replaced with a low self-replace ratio, such as 0.1, the resulting target image closely resembles the one obtained using the target prompt directly. However, when the self-attention map is replaced in all attention layers and for 90% of the denoising steps, a target image that closely matches the original image is generated, as depicted in the top left corner of Figure 10. A more balanced approach involves replacing the self-attention map in layers 4–14 with replacement ratios ranging from 0.4 to 0.8, resulting in more favorable outcomes.

10. Real Image Editing with Null-Text Inversion

Algorithm 3 Editing Method for Real Image Using Null-Text Inversion [18]

Input: P_{src}: a source prompt; P_{dst}: a target prompt; I: real image; S: random seed;
Output: I_{dst}: edited image; I_{res}: reconstructed image;
$\hat{z}_T, \{\varnothing\}_{t=1}^{T} \leftarrow \text{NULLTEXT-INV}(P_{src}, I)$;
$z_T^* \leftarrow \hat{z}_T$;
for $t = T, T-1, ..., 1$ **do**
  $z_{t-1}, M_{self} \leftarrow \text{DM}(\hat{z}_T, \{\varnothing\}_t, P_{src}, t)$;
  $z_{t-1}^* \leftarrow \text{DM}(z_t^*, \{\varnothing\}_t, P_{dst}, t)\{M_{self}^* \leftarrow M_{self}\}$;
end for
Return $(I_{res} \leftarrow Decoder(z_0), I_{dst} \leftarrow Decoder(z_0^*))$;

Algorithm 3 outlines the pseudo-code for editing real

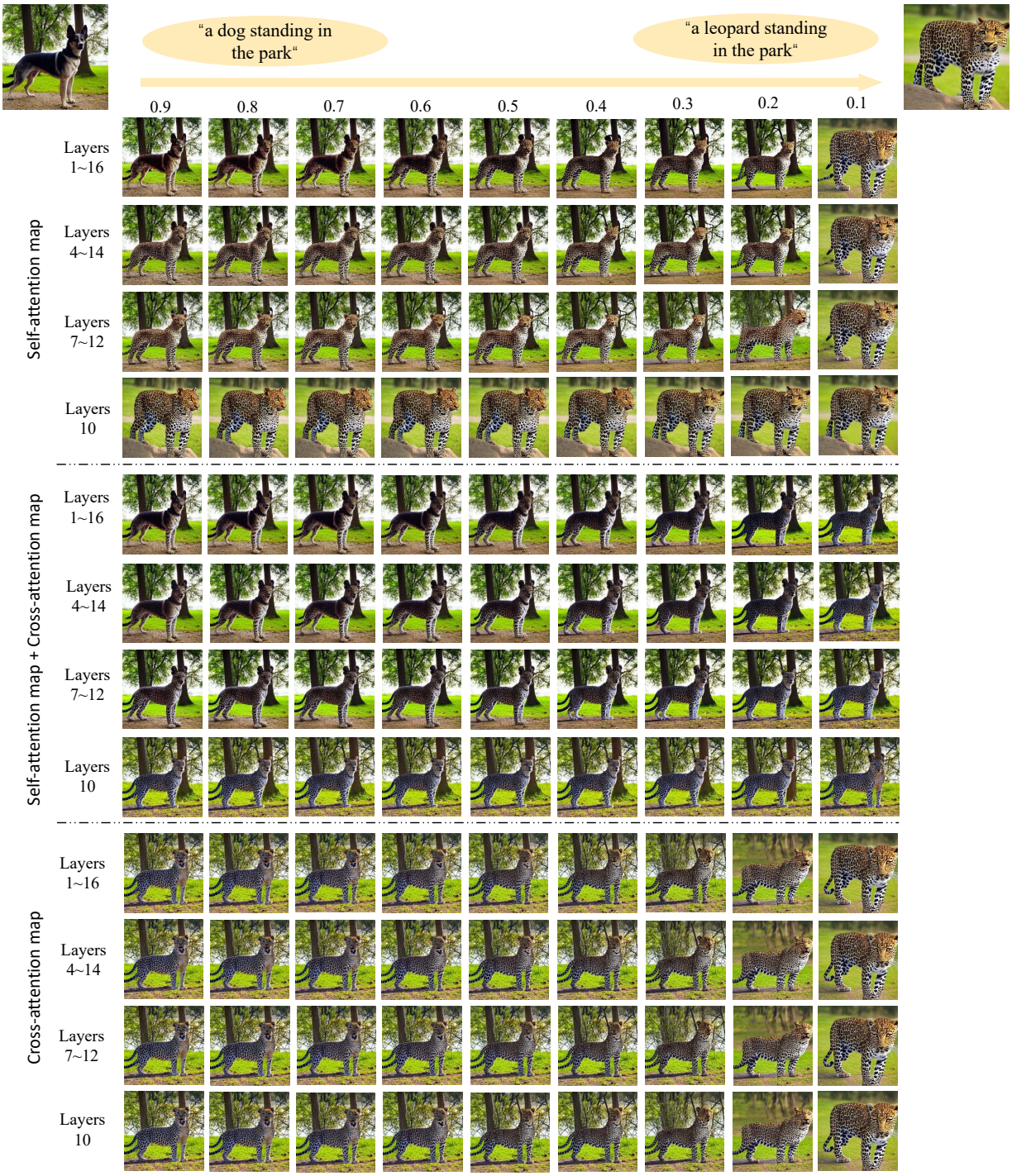

Figure 10. Ablation experiment results showing three types of attention map replacements for noun editing at different replace step ratios. A higher ratio denotes a greater number of replacement steps.

images using the Freeprompt Editing (FPE) combined with Null-Text Inversion [18]. Figure 12 presents the experimental results of editing real images using DDIM Inversion [33] and Null-Text Inversion. Both methods effectively modify the original image based on the target text.

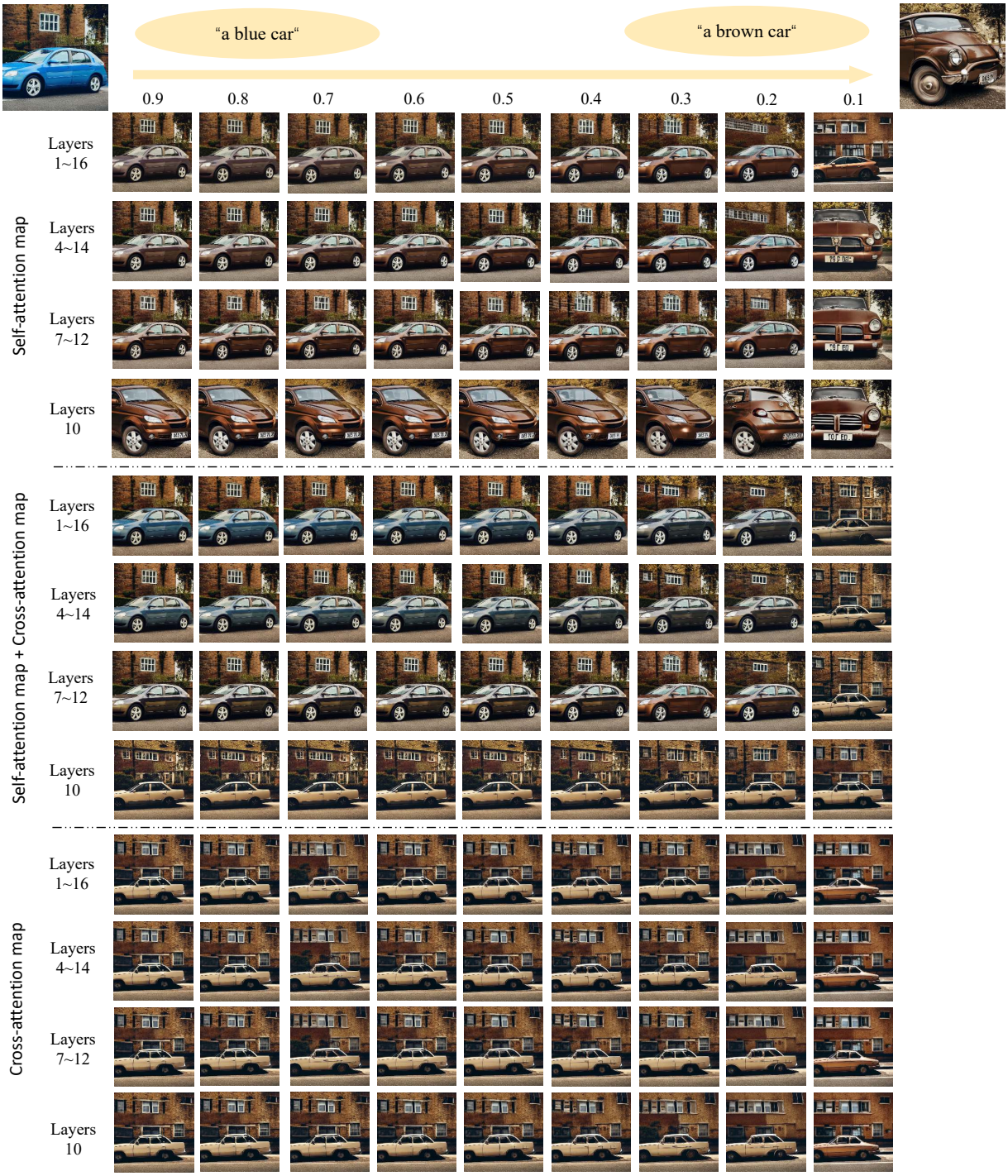

Figure 11. Ablation experiment results showing three types of attention map replacements for color editing at different replace step ratios. A higher ratio denotes a greater number of replacement steps.

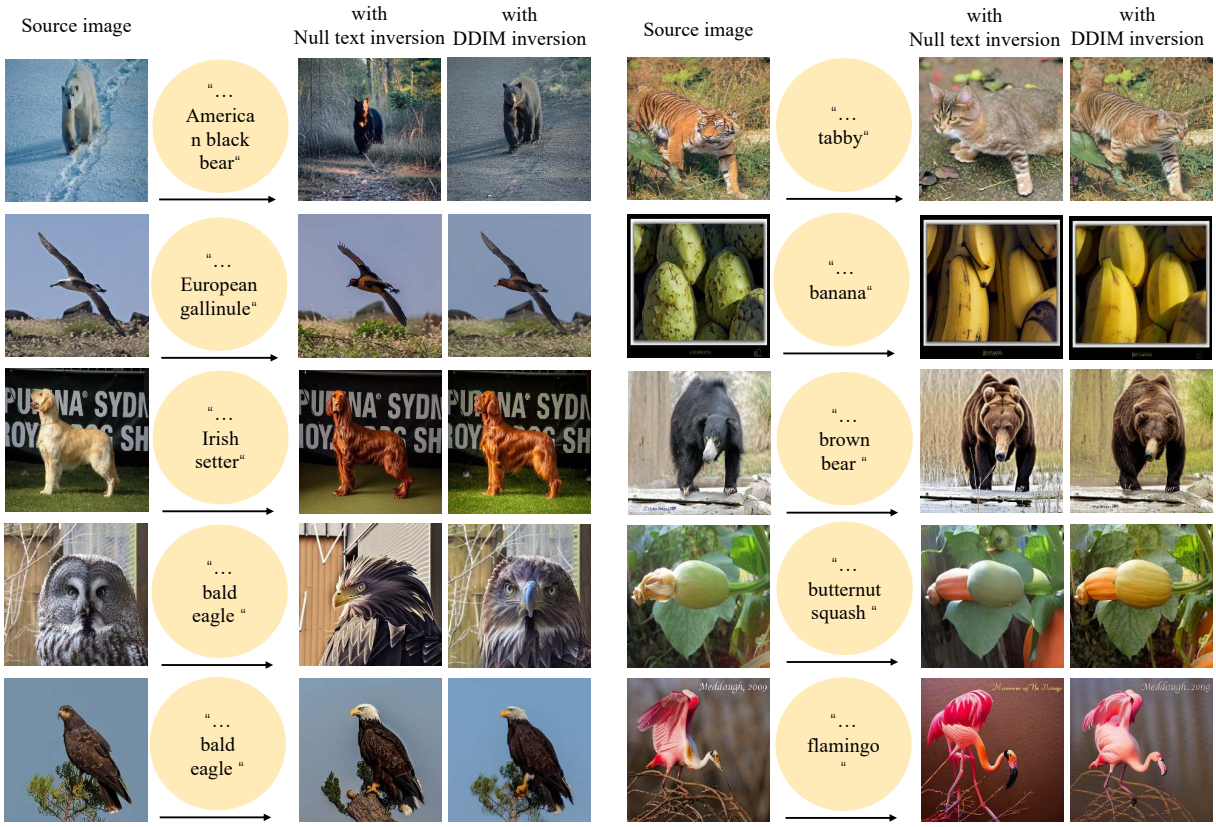

Figure 12. Experiment results of real image editing in ImageNet-Real-Edit with Null-Text Inversion and DDIM Inversion. The source prompt form is *“a photo of a/an <object>”* for Null-Text Inversion.