Appendix

A. Paradigm Comparison

Fig. 2 in the main paper visually represents the differences between our method and the other two types of methods.

In the context of training-based methods, we can categorize them into two types, both of which take a set of 3-5 images representing the same concept as input. Type I methods, such as DreamBooth [33] and Break-A-Scene [3], require training all the weights of the U-Net model. In contrast, type II methods, including Textual Inversion [11], XTI [38], and NeTI [1], typically focus on training a text-embedding that captures the semantic information of the input text, rather than training the entire network.

Furthermore, there is a second category of methods (BLIP Diffusion [20], ELITE [41]) tailored for generalizable customization. These methods typically involve two stages. In the first stage, the model is pre-trained on a large-scale dataset to learn the distribution within a specific domain. In the second stage, the pre-trained model is used in a training-free manner. However, this type of method is time-consuming, as it requires extensive pre-training, and applying it to a new model necessitates retraining.

In contrast to the aforementioned methods, our tuning-free approach differs in several key aspects. Firstly, it does not necessitate training the entire network or learning a text-embedding on a few images. Additionally, our method does not rely on pre-training on a large-scale dataset. Instead, our approach only requires a single image per concept as input. By replacing the self-attention mechanism in select blocks of the U-Net network with MRSA (Multi-Reference Self-Attention), our method enables single-concept customization and multi-concept composition. This means that our method can be easily applied to different models.

In summary, our method stands out by offering a more efficient approach to customization compared to the other two types of methods. It leverages MRSA to enable flexible customization of single or multiple concepts.

B. More Visual Results

B.1. Single-concept Customization

Our approach allows for generating a diverse range of customized images based on a single concept in the input image. Fig. A1 showcases the results obtained using different seeds and target prompts. These results demonstrate the high fidelity and preservation of identity, highlighting the effectiveness and robustness of our approach.

B.2. Empower Other Methods

Our methods have the capability to enhance other methods in terms of identity preservation and visual fidelity. When combined with BLIP diffusion, our methods effectively maintain the identity of the given concept as shown in Fig. A2. Moreover, our method can be seamlessly integrate with ControlNet, as illustrated in Fig. A3.

C. Correspondence Visualization

To illuminate the feature-wise correspondence between the reference concepts and the generated multi-concept composition image, we employ a visualization based on an attention map from layer 10 during the 50th denoising step. As depicted in Fig. A4, pixels exhibiting the highest similarity between the combined and reference concepts are marked with matching colors on the map.

D. Visualization of Multi-attention Map

The visualization of the weighted mask's importance is depicted in Figs. A5 and A6. In the left column, when using the weighted mask, we set $\mathbf{w} = [1, 3, 3, 3]$. The generated image gives higher priority to the reference concepts. This is evident from the attention map, which shows that the image's attention is focused on the reference concepts.

In contrast, the right column represents the scenario without the weighted mask, where $\mathbf{w} = [1, 1, 1, 1]$. Here, the output image prioritizes itself, with the attention being more concentrated on itself rather than the reference concepts. Consequently, this can lead to a weaker preservation of the identity of the reference concepts, such as the scarf and hat.

E. Selective MRSA Replacement

We provide more comprehensive results of selectively replacing the original self-attention in the basic block with MRSA as shown in Fig. A7 and Fig. A8. Fig. A7 shows a more fine-grained replacement strategy, Fig. A8 presents the ablation results of the selective replacement strategy when using different combinations of reference concepts as input.

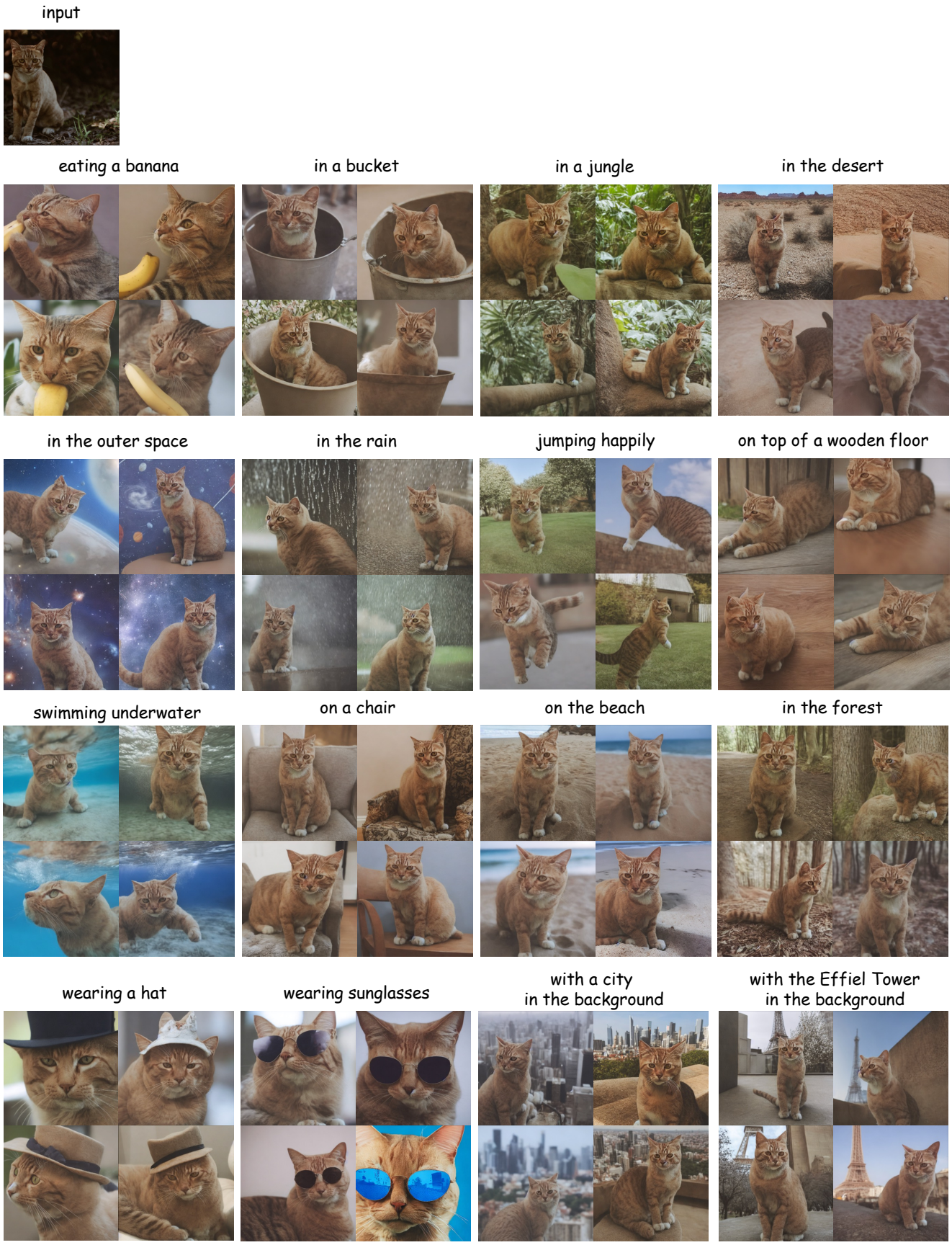

Figure A1. **Single concept qualitative results.** Our method enables extensive customization of a single concept by inputting a single image.

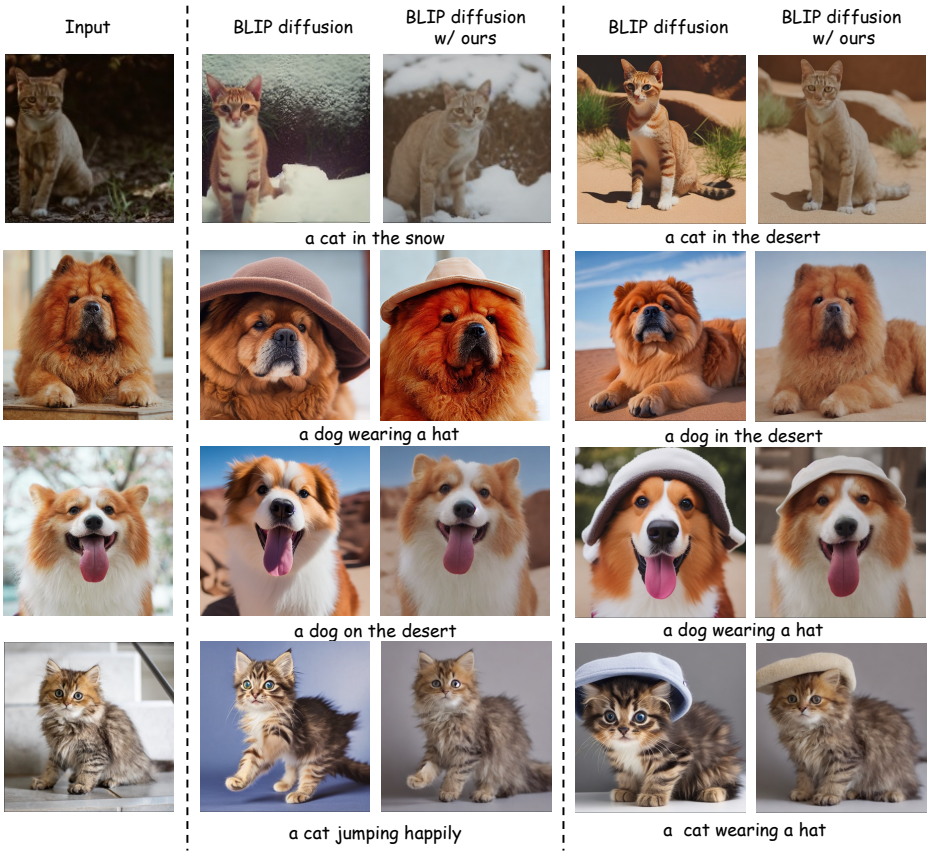

Figure A2. **BLIP diffusion vs. BLIP diffusion with FreeCustom.** The images generated by BLIP diffusion are visually appealing, but they may not achieve perfect identity recovery. However, when our approach is combined with BLIP diffusion, the generated results effectively maintain the identity of the given concept.

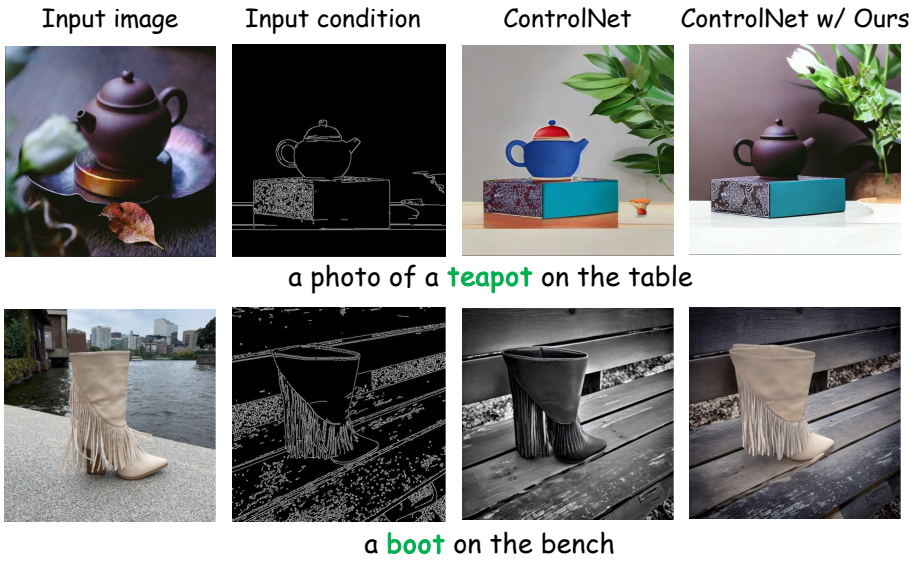

Figure A3. **ControlNet with FreeCustom.** Our method can be seamlessly integrate with ControlNet. From left to right, the images are arranged as follows: the input image, the input conditions, the image output by vanilla ControlNet, and the image output by ControlNet enhanced by FreeCustom.

Figure A4. **Correspondence visualization.** We visualize the correspondence between each feature in the reference concepts and each feature in the generated multi-concept composition image using an attention map. In this figure, the features with the highest similarity between the combined concept and the reference concepts are marked with the same color to indicate the correspondence between them. The results indicate that we have achieved relatively good consistency across the features. For instance, the hat and scarf in the reference concept exhibit a strong match with the hat and scarf in the generated image, underscoring the effectiveness of our approach.

Figure A5. **Visualization of multi-attention map.** Each column of pictures in the figure represents, from left to right: the image output by our method (1st column), and the multi-attention map $A = \text{Softmax}(\frac{\mathbf{M}_w \odot (\mathbf{Q}\mathbf{K}'^T)}{\sqrt{d}})$, $A \in \mathbb{R}^{H \times ((N+1)W)}$, $N = 3$ here (the following 4 columns). The red box indicates the feature that queries other keys. When using the weighted mask, the generated image will prioritize the reference concepts, while without the weighted mask, the output image will prioritize itself, resulting in a weaker preservation of the identity of the reference concepts.

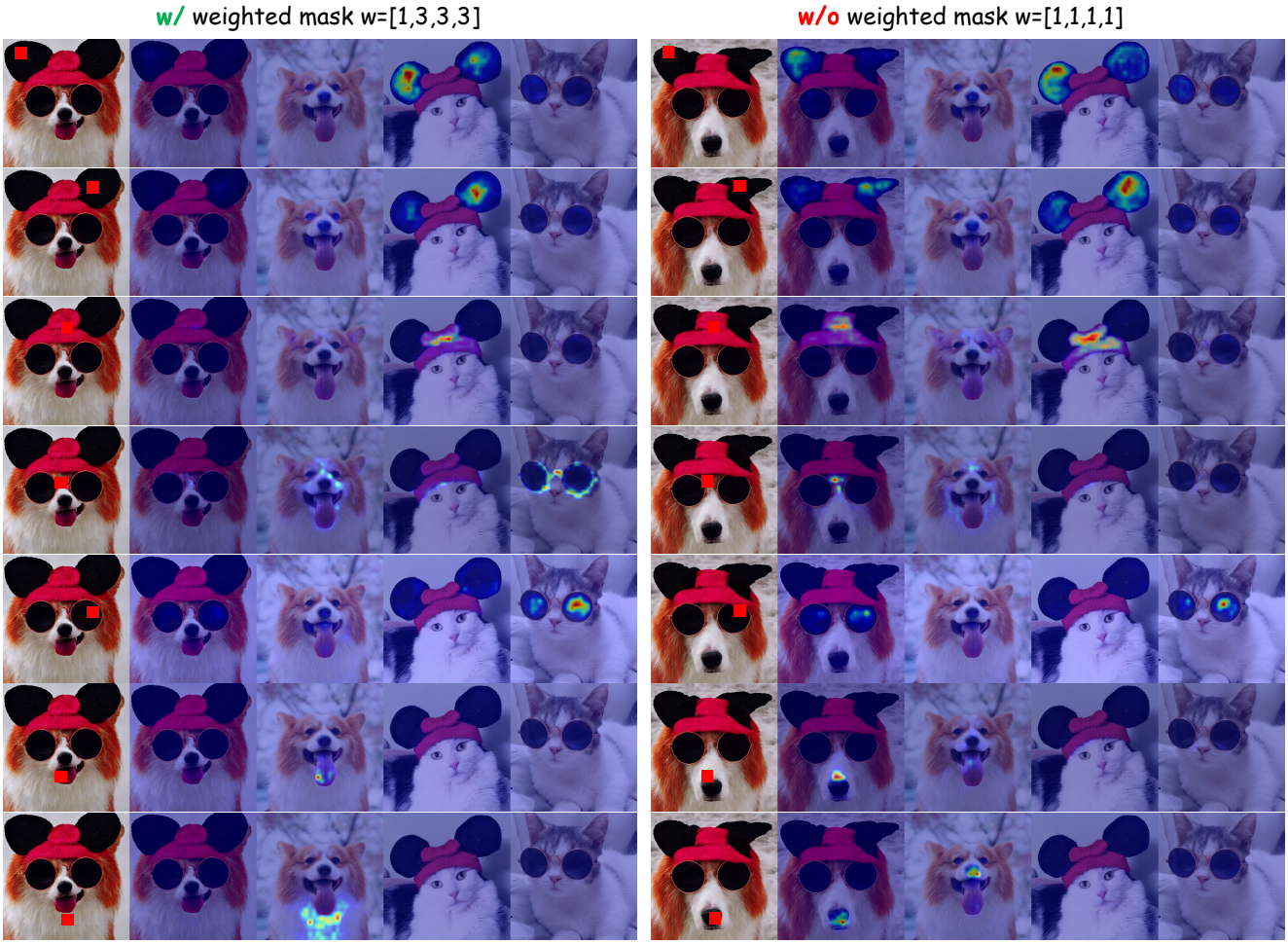

Figure A6. **Visualization of multi-attention map.** Each column of pictures in the figure represents, from left to right: the image output by our method (1st column), and the multi-attention map $A = \text{Softmax}(\frac{\mathbf{M}_w \odot (\mathbf{Q}\mathbf{K}'^T)}{\sqrt{d}})$, $A \in \mathbb{R}^{H \times ((N+1)W)}$, $N = 3$ here (the following 4 columns). The red box indicates the feature that queries other keys. When using the weighted mask, the generated image will prioritize the reference concepts, while without the weighted mask, the output image will prioritize itself, resulting in a weaker preservation of the identity of the reference concepts.

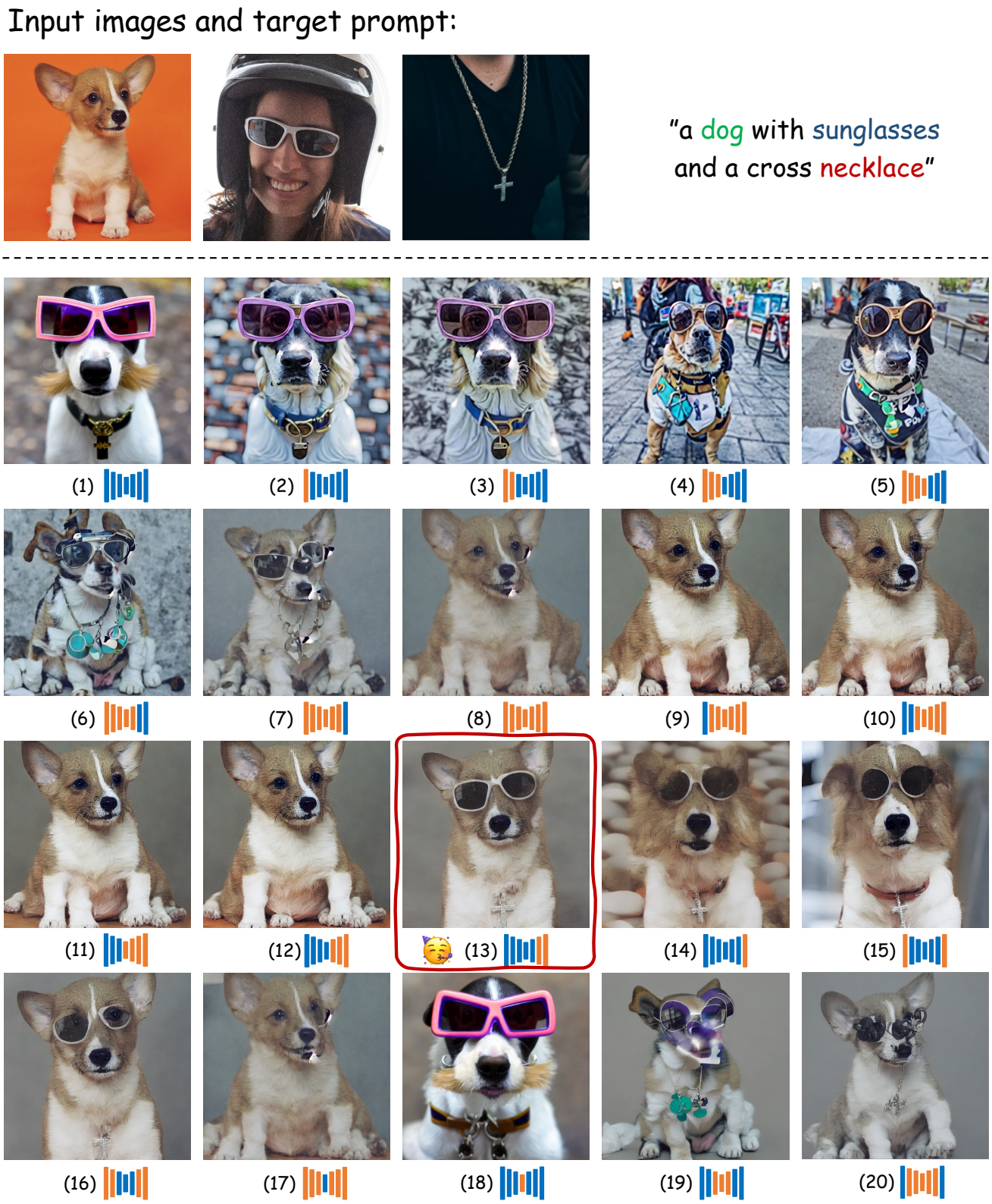

Figure A7. **Selective applying MRSA to basic blocks.** The blue color represents the original basic block and the yellow color indicates the basic block whose self-attention is replaced by MRSA.

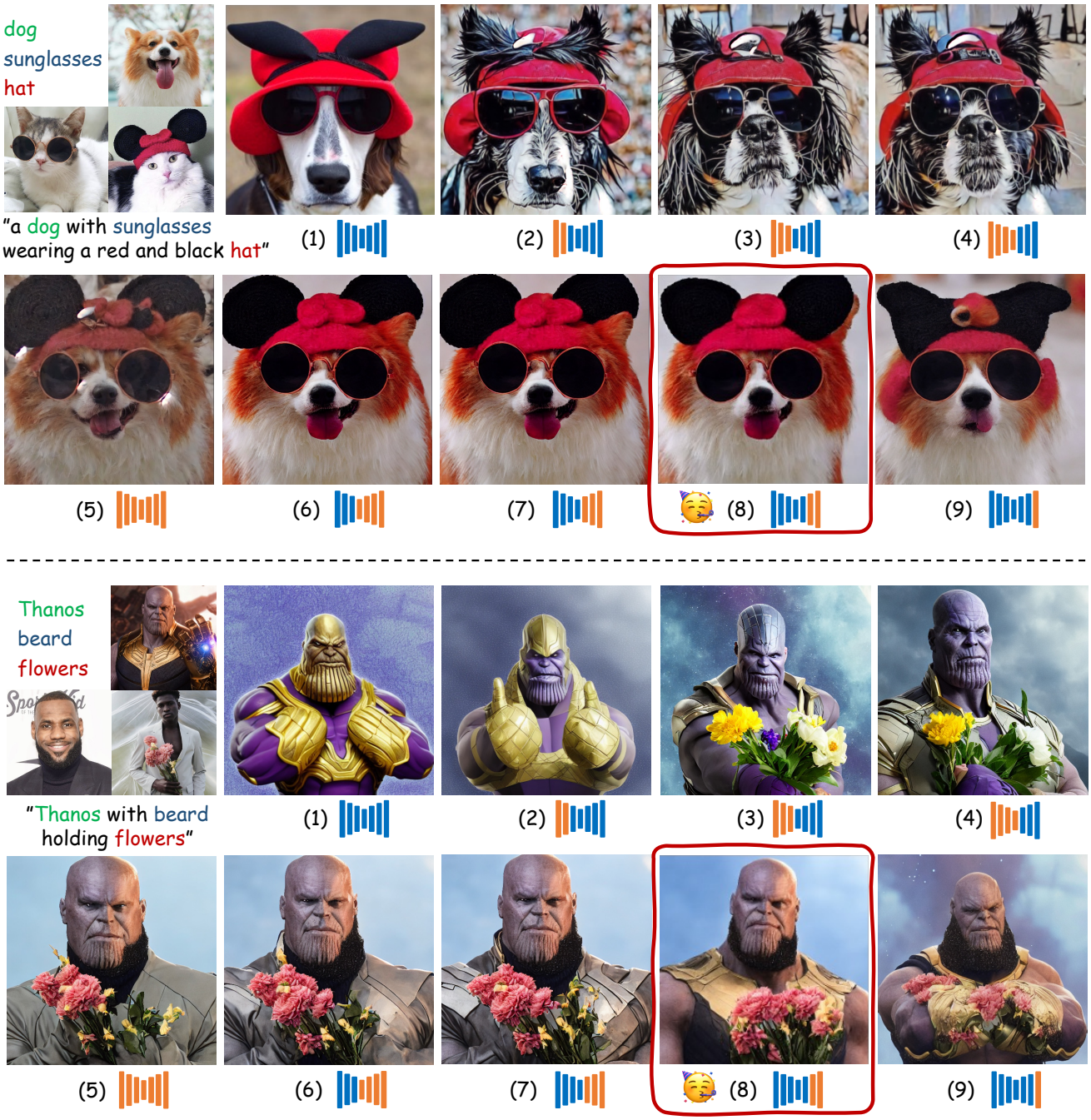

Figure A8. **Selective applying MRSA to basic blocks.** The blue color represents the original basic block and the yellow color indicates the basic block whose self-attention is replaced by MRSA.