Prompt-to-Prompt Image Editing with Cross Attention Control

Jun Hyung Lee

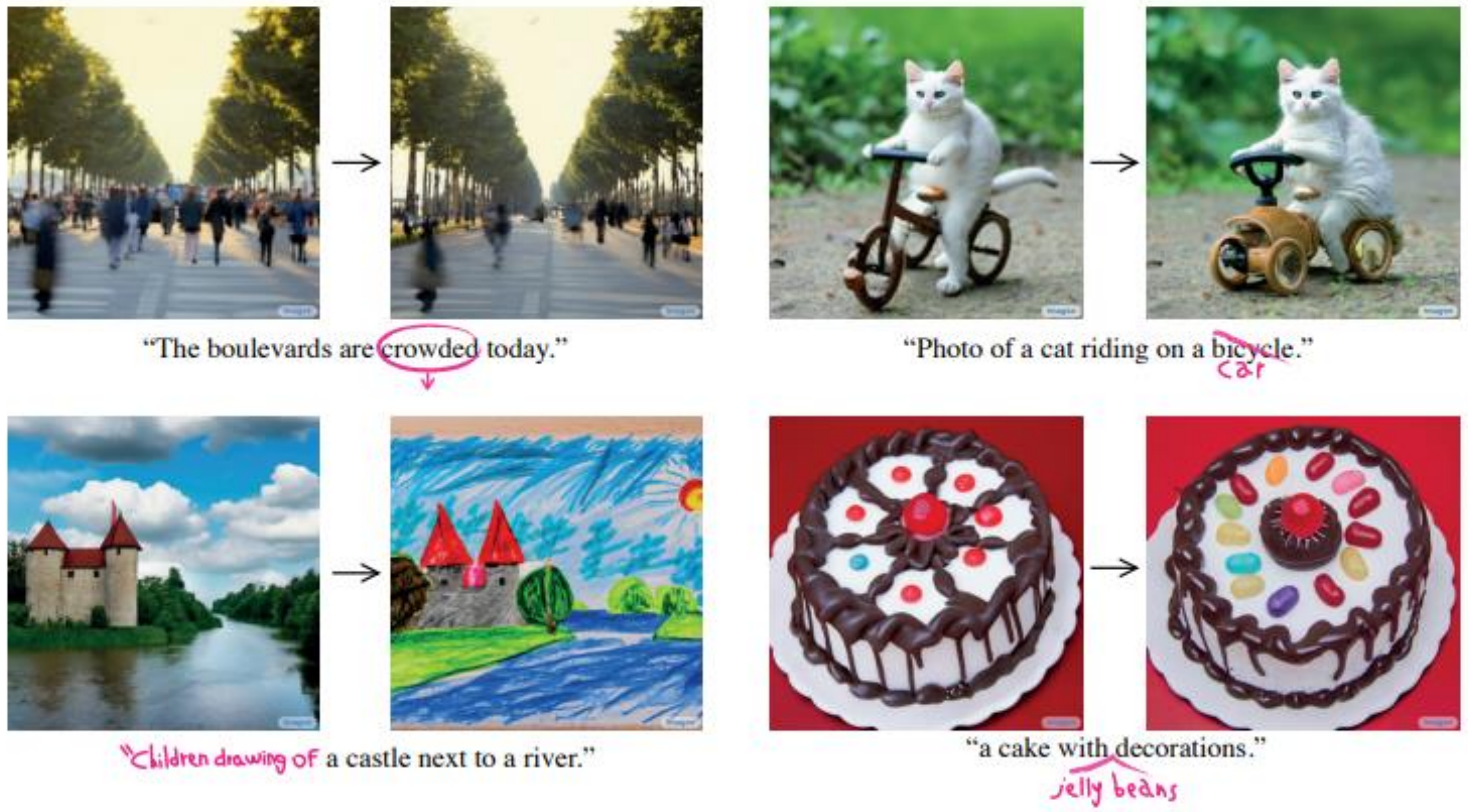
References

- **Denoising Diffusion Implicit Models**, Jiaming Song (2021)
- **Denoising Diffusion Probabilistic Models**, Jonathan Ho (2020)
- **Diffusion Models beat GANs on image synthesis**, Prafulla Dhariwal (2021)
- **Classifier-free diffusion guidance**, Jonathan Ho (2021)

Overview

- Prompt2Prompt is a technique for image editing without masking and fine-tuning.
- The manipulations are infiltrated through the cross-attention mechanism of the diffusion model without the need for any specifications over the image pixel space.
- While most works that require only text are limited to global editing. Previous localized editing paper had limitation of just changing textures, but not modifying complex structures, such as changing a bicycle to a car.

Edits are controlled by text only



"The boulevards are crowded today."

"Photo of a cat riding on a ~~bicycle~~ car."

"Children drawing of a castle next to a river."

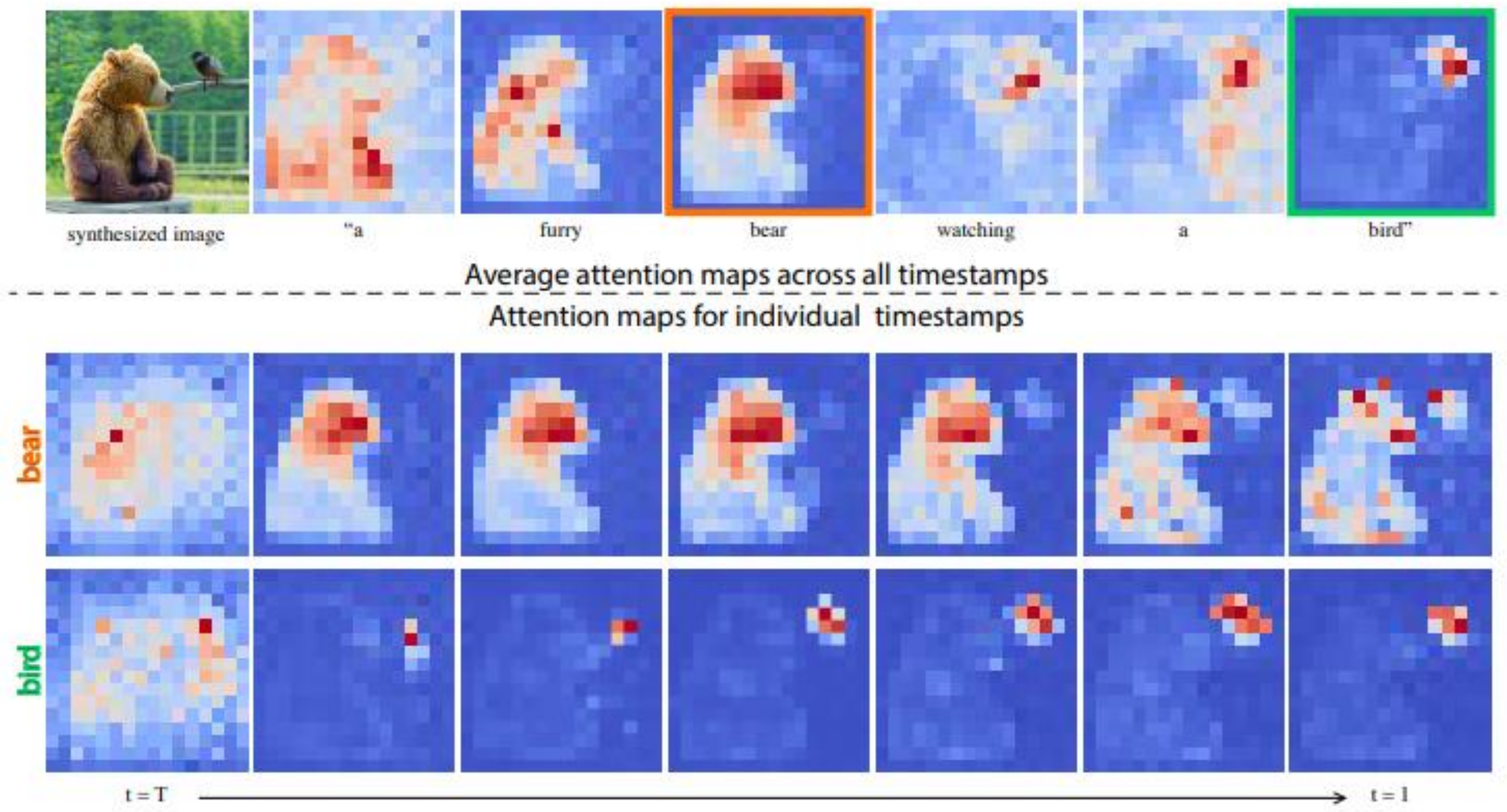
"a cake with decorations." jelly beans

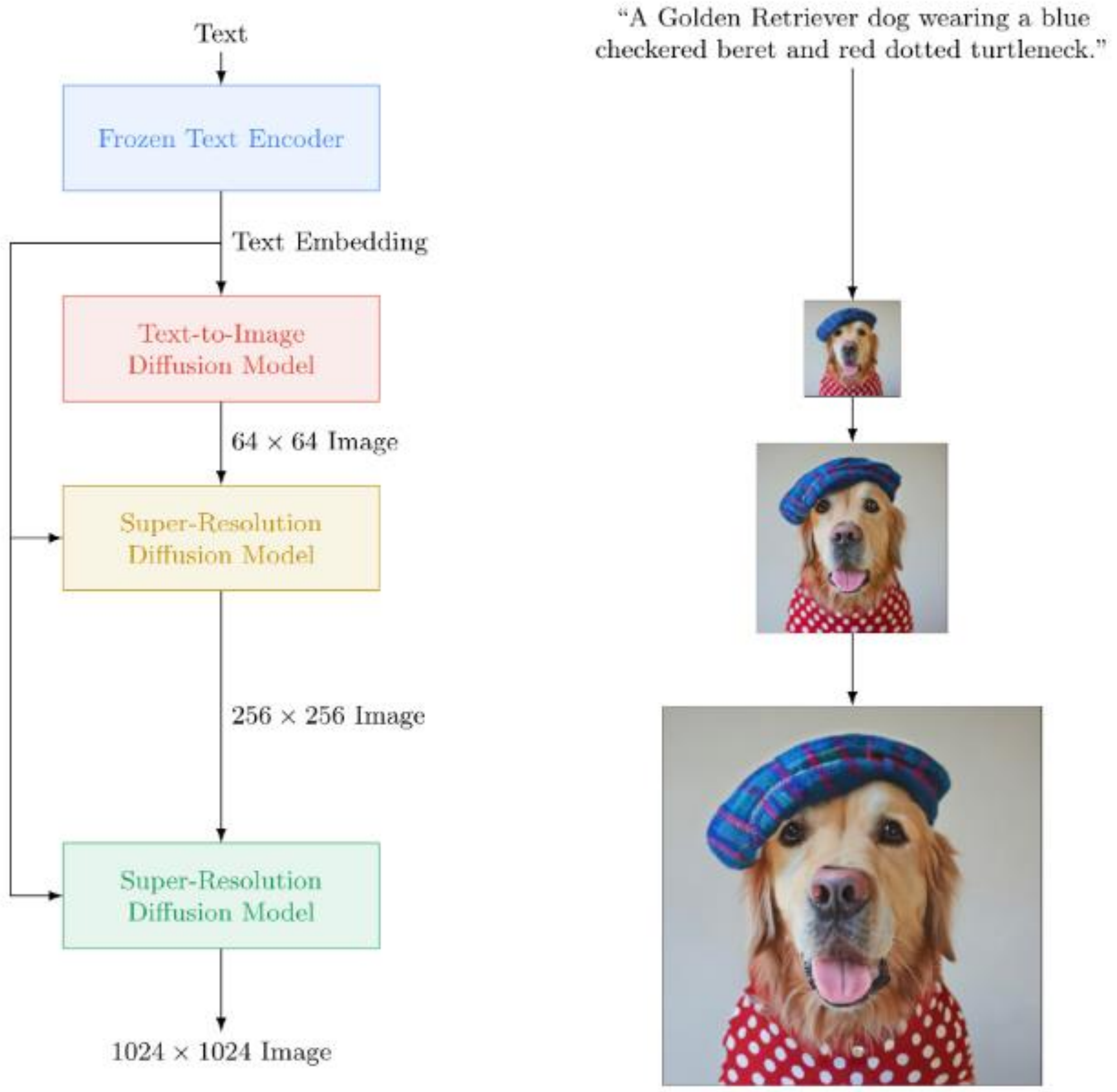
Cross-attention in text-conditioned Diffusion Models

- A cross attention (Q: pixels, K; texts) determines the structure of generated images during backward diffusion steps.
- Image editing by explicit control of attention map

<u>Structure of the image is already determined in the early steps of the diffusion process</u>

$$M = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

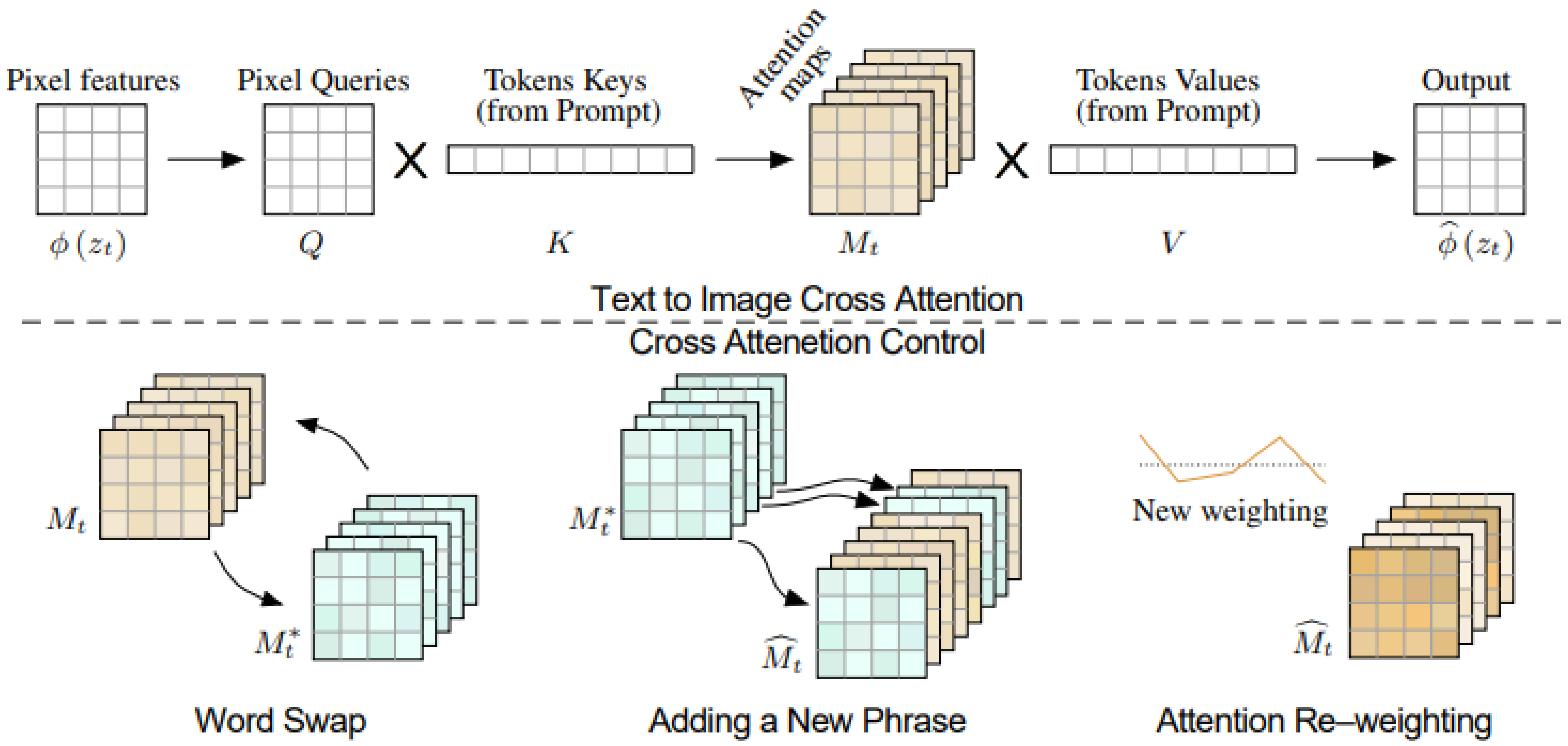




Imagen

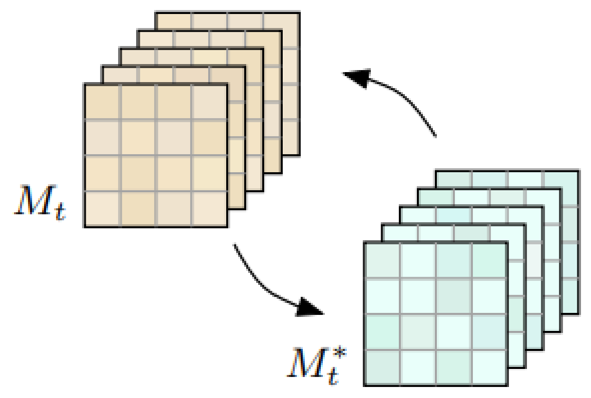
Method

- ➢ Cross-attention maps are high-dimensional tensors that bind pixels and tokens extracted from the prompt text.
- ➢ These maps contains rich semantic relations which critically affect the generated image.
- ➢ A cross attention (Q: pixels, K; texts) determines the structure of generated images during backward diffusion steps.
- ➢ Cross attention maps can be manually replaced or revised(Word Swap, Adding a New Phrase, Attention Re-weighting) during inference for image generation.



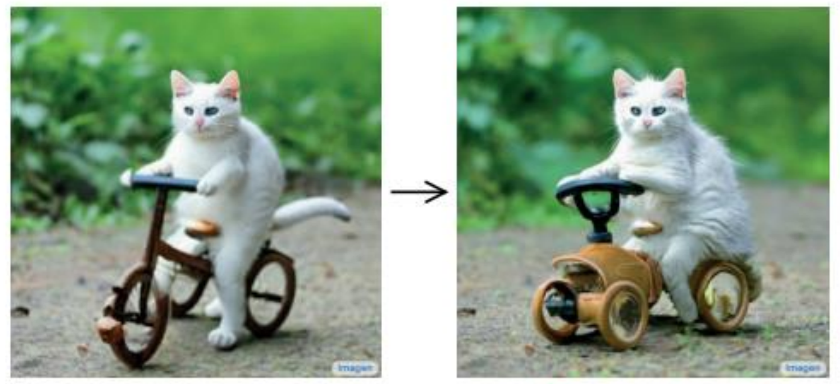
Method

①



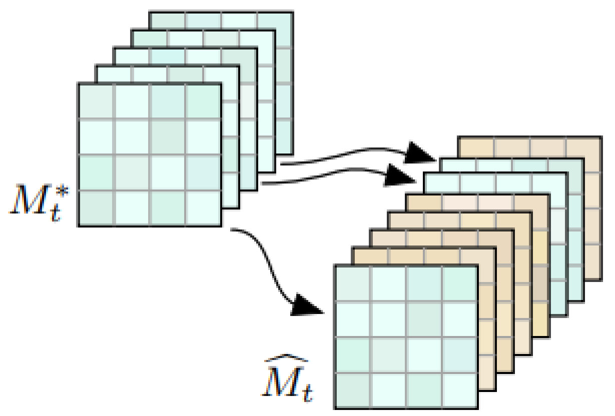
Word Swap

Change single token’s value in the prompt(e.g., “bicycle” to “car”) fixing the cross-attention maps, to preserve the scene composition.



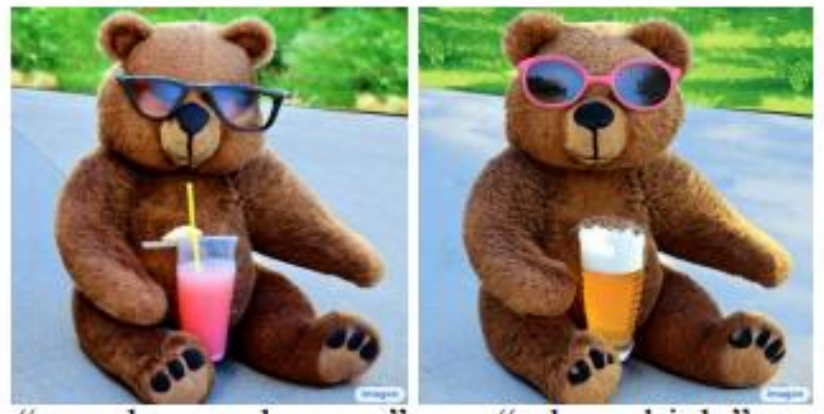
“Photo of a cat riding on a ~~bicycle~~.” car

②



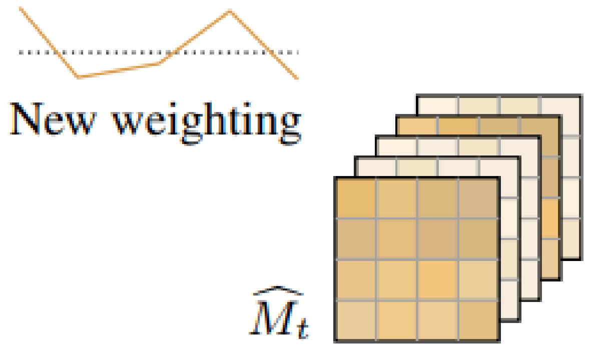
Adding a New Phrase

Globally edit an image (Change the style by adding new words to the prompt and freezing the attention on previous tokens, while allowing new attention to flow to the new tokens.



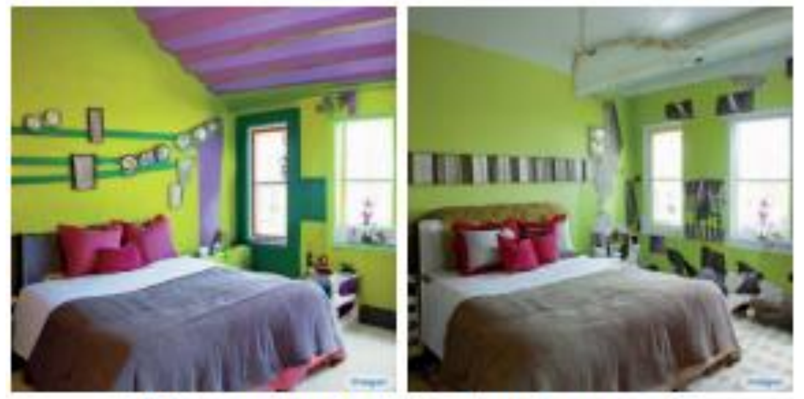
“...geeky sunglasses...” “...beer drink.”

③



Attention Re–weighting

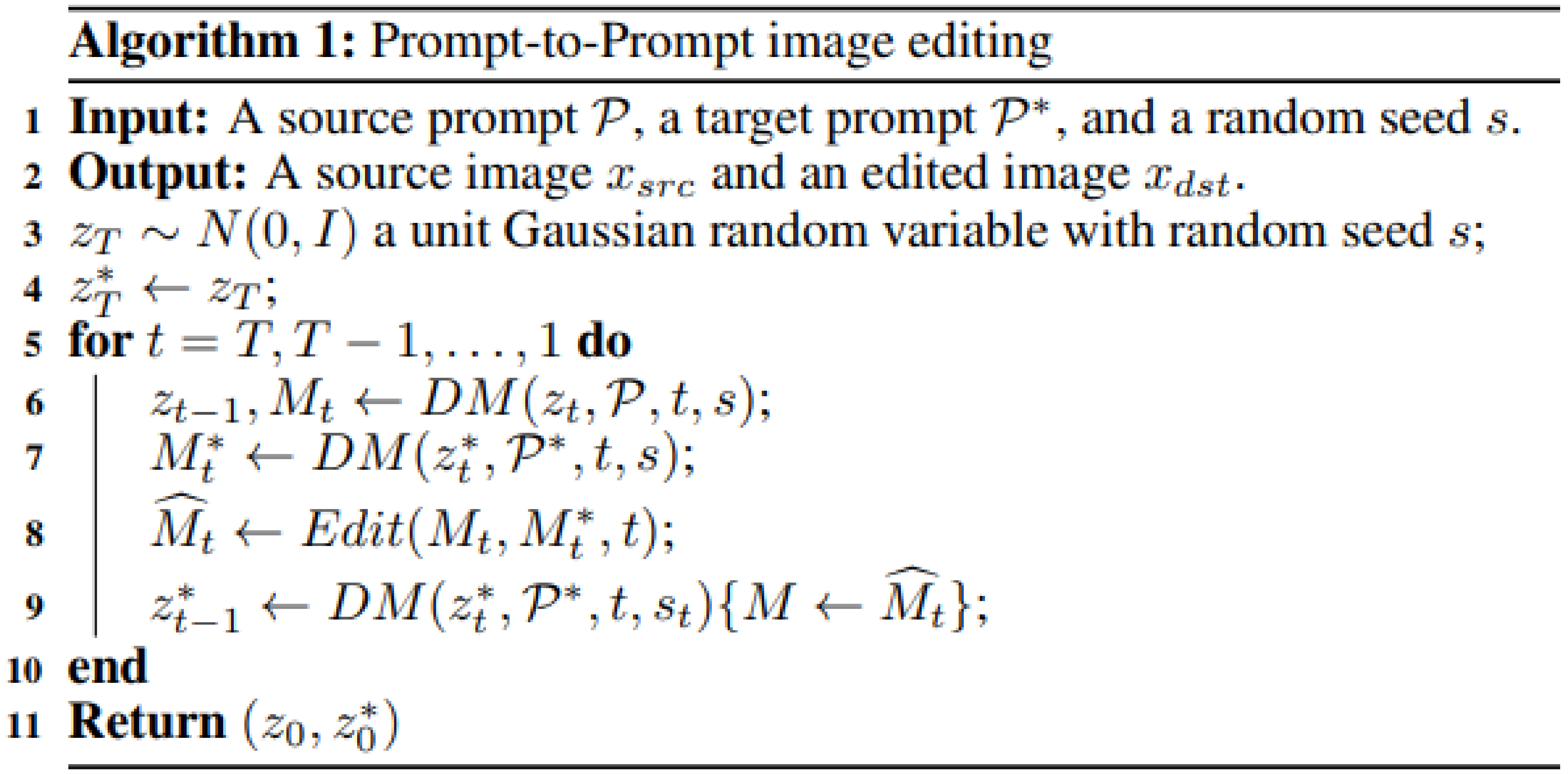
Amplify or attenuate the semantic effect of a word in the generated image.



“My colorful(↓) bedroom.”

Algorithm

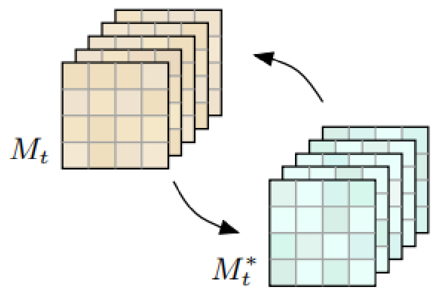
- Algorithm for controlled image generation consists of performing the iterative diffusion process for both prompts simultaneously, where an attention-based manipulation is applied in each step according to the desired editing task. (fix random seed because the same prompt could result in different outputs due to the nature of diffusion models)



Algorithm 1: Prompt-to-Prompt image editing

```
Input: A source prompt $\mathcal{P}$, a target prompt $\mathcal{P}^*$, and a random seed $s$.
Output: A source image $x_{src}$ and an edited image $x_{dst}$.
$z_T \sim N(0, I)$ a unit Gaussian random variable with random seed $s$;
$z_T^* \leftarrow z_T$;
for $t = T, T-1, \ldots, 1$ do
    $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, s)$;
    $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)$;
    $\widehat{M}_t \leftarrow Edit(M_t, M_t^*, t)$;
    $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s_t)\{M \leftarrow \widehat{M}_t\}$;
end
Return $(z_0, z_0^*)$
```

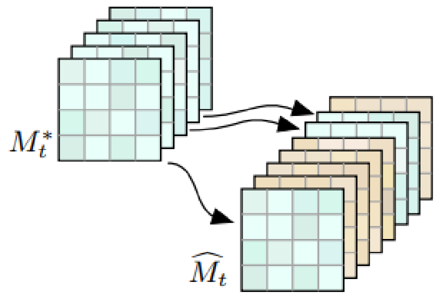
①



Word Swap

$$Edit(M_t, M_t^*, t) := \begin{cases} M_t^* & \text{if } t < \tau \\ M_t & \text{otherwise.} \end{cases}$$

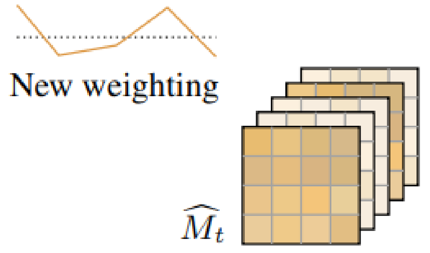
②



Adding a New Phrase

$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} (M_t^*)_{i,j} & \text{if } A(j) = None \\ (M_t)_{i,A(j)} & \text{otherwise.} \end{cases}$$

③

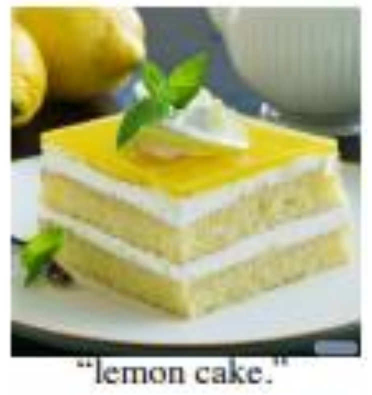


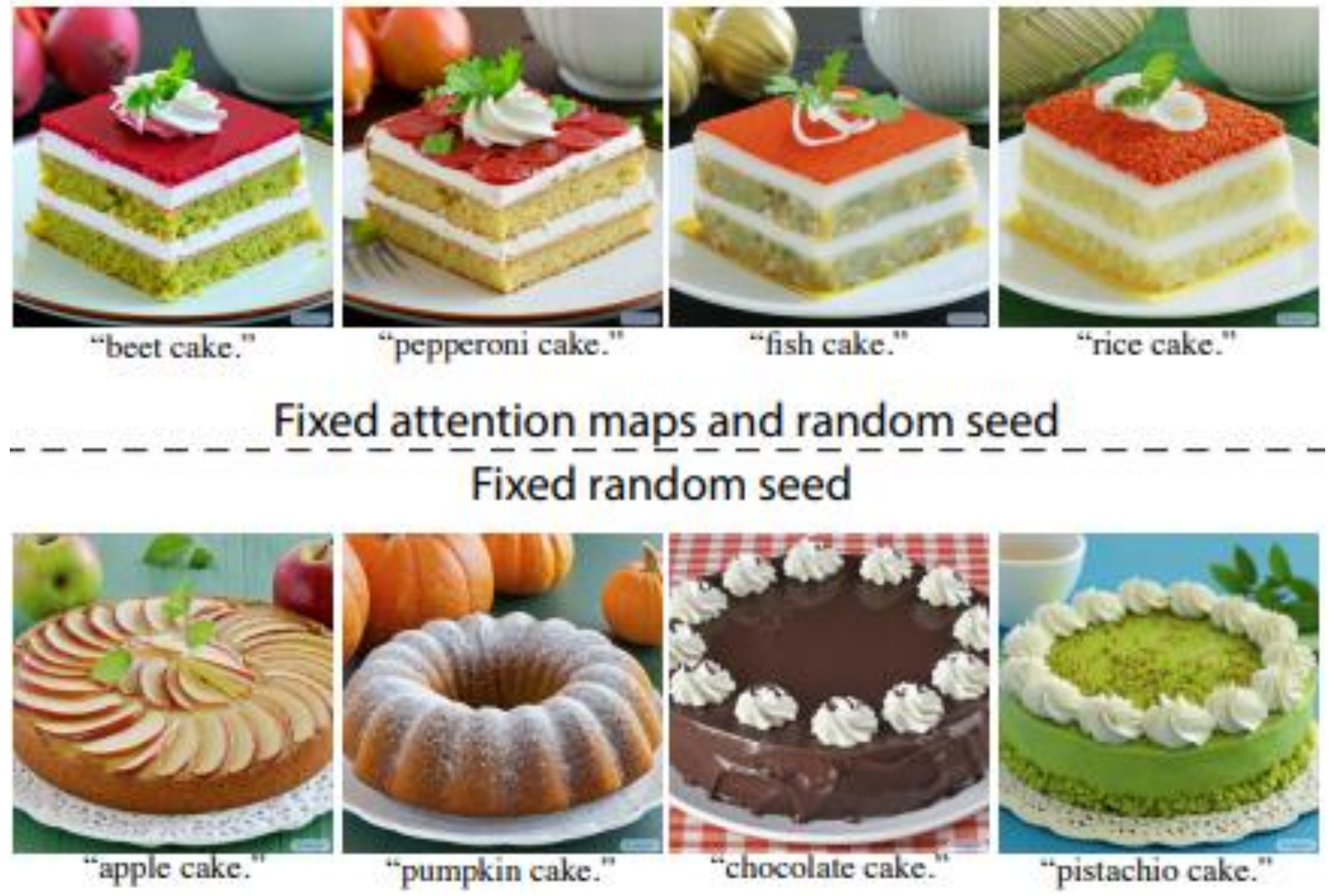
Attention Re–weighting

$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} c \cdot (M_t)_{i,j} & \text{if } j = j^* \\ (M_t)_{i,j} & \text{otherwise.} \end{cases}$$

Content modification through attention injection

- Attempt to fix the internal randomness and regenerate using the edited text prompt results in a completely different image with a different structure and composition.
- Key observation is that the structure and appearances of the generated image depend not only on the random seed, but also on the interaction between the pixels to the text embedding through the diffusion process.



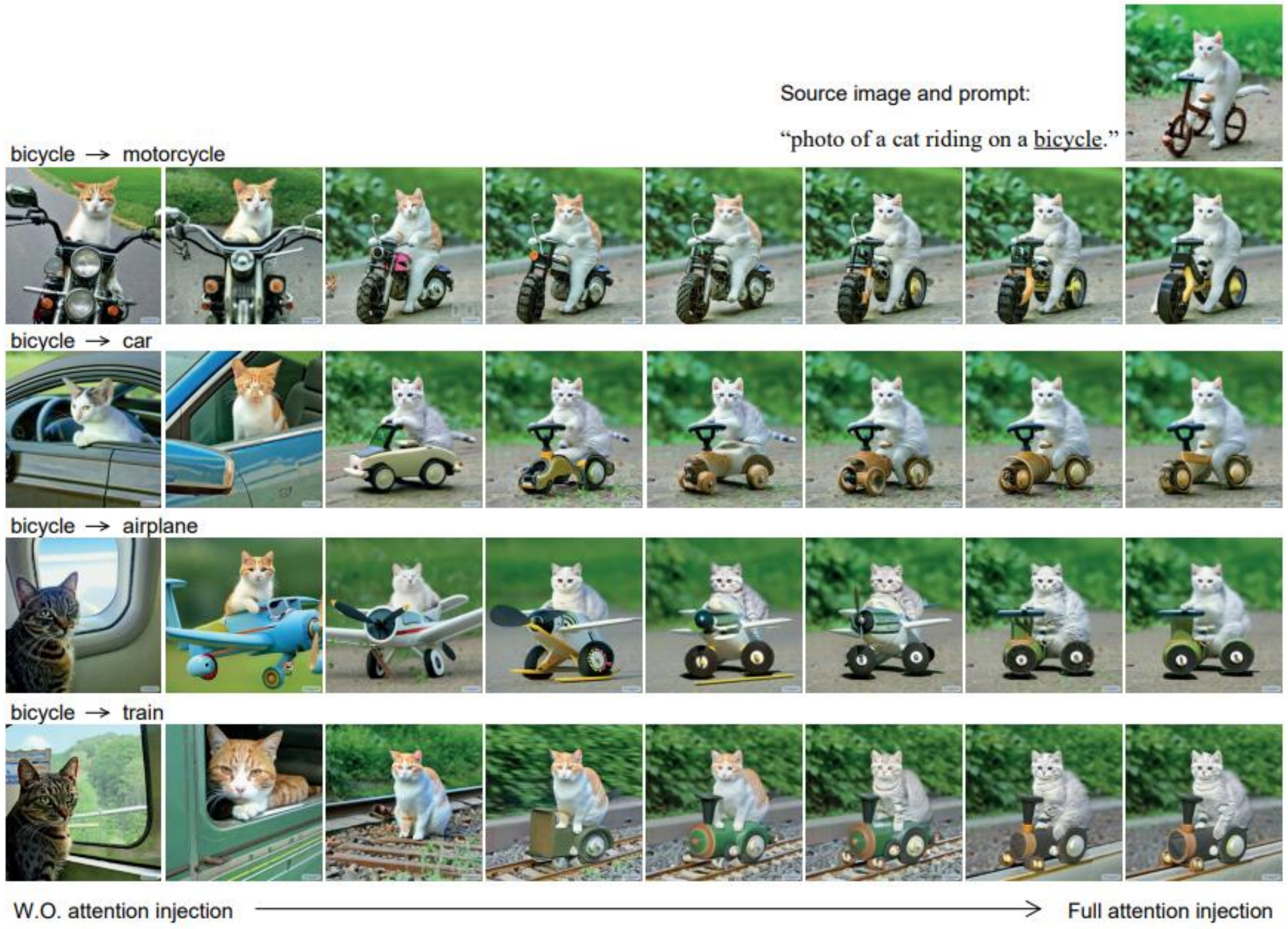


Fixed attention maps and random seed: Inject the attention weights of the original image during the diffusion process.

Fixed random seed: Using the same random seeds as the original image, without injecting the attention weights.

Attention injection through a varied number of diffusion steps

Global Changes: without cross attention injection which leads to an entirely different outcome.



Local Changes: injecting cross attention to an increasing number of diffusion steps.

Examples

①

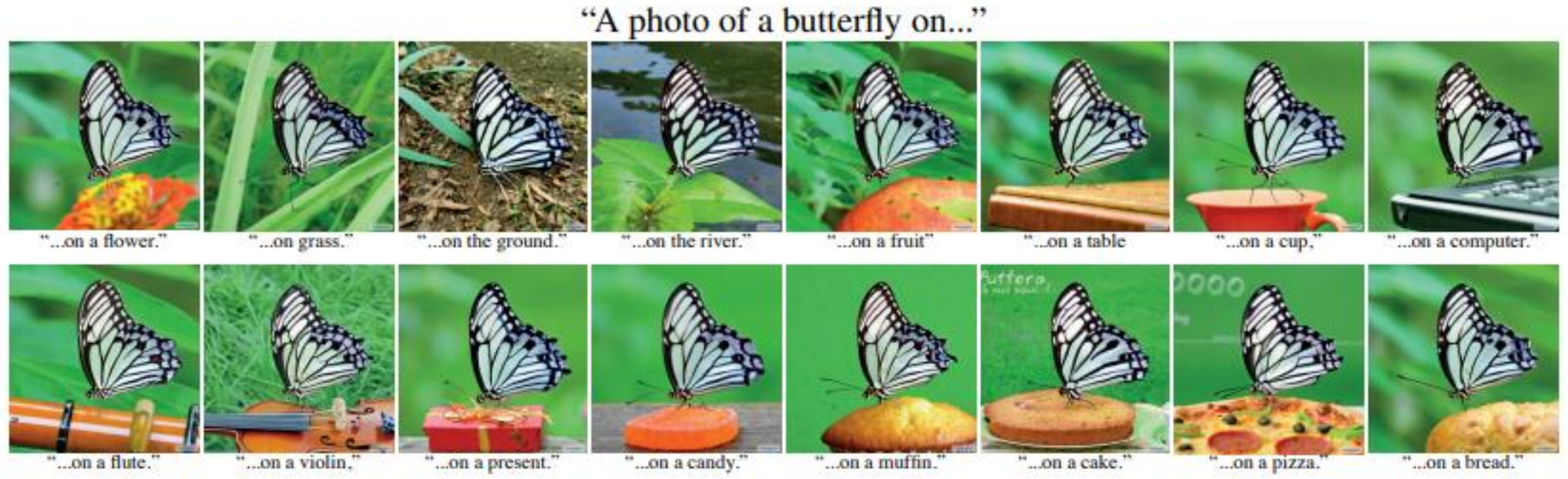


Figure 5: Object preservation. By injecting only the attention weights of the word “butterfly”, taken from the top-left image, we can preserve the structure and appearance of a single item while replacing its context. Note how the butterfly sits on top of all objects in a very plausible manner.

②

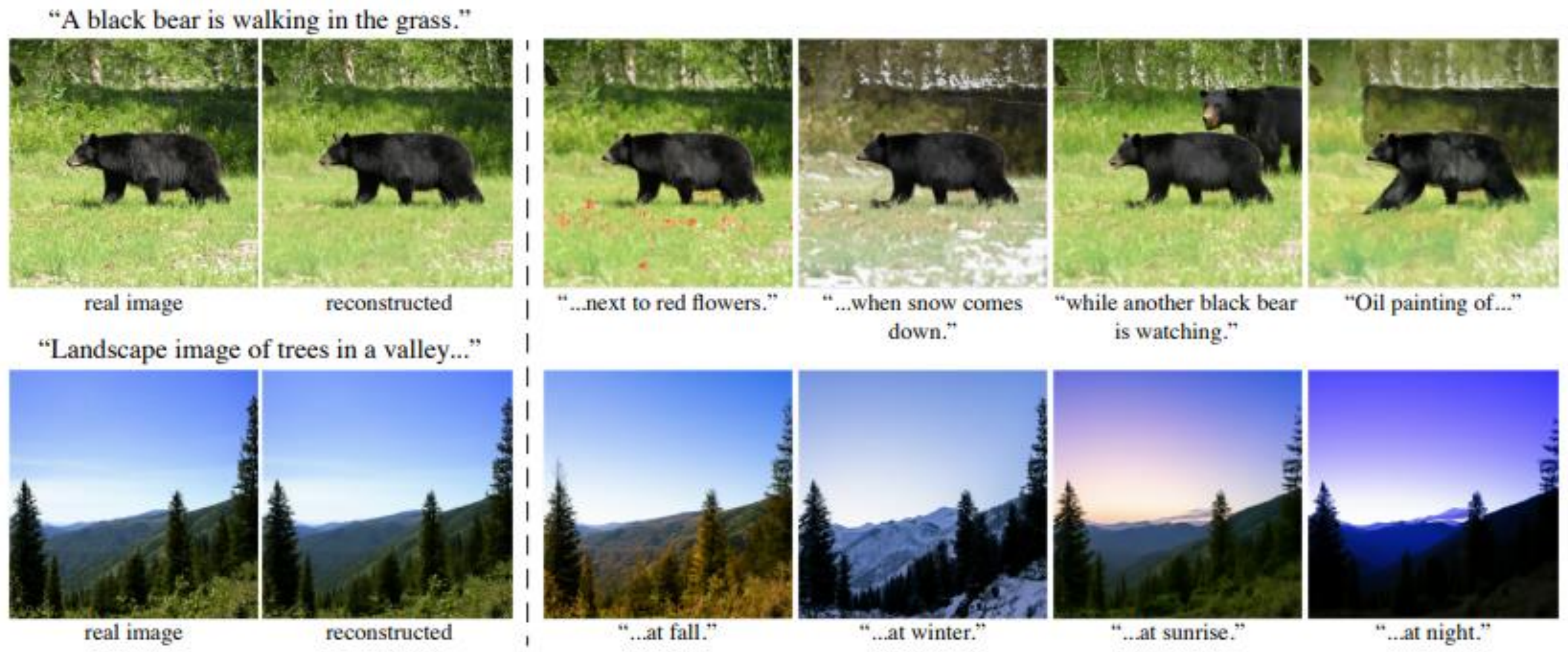

“A black bear is walking in the grass.”
real image
reconstructed
“...next to red flowers.”
“...when snow comes down.”
“while another black bear is watching.”
“Oil painting of...”
“Landscape image of trees in a valley...”
real image
reconstructed
“...at fall.”
“...at winter.”
“...at sunrise.”
“...at night.”

Figure 10: Editing of real images. On the left, inversion results using DDIM [40] sampling. We reverse the diffusion process initialized on a given real image and text prompt. This results in a latent noise that produces an approximation to the input image when fed to the diffusion process. Afterward, on the right, we apply our Prompt-to-Prompt technique to edit the images.

③

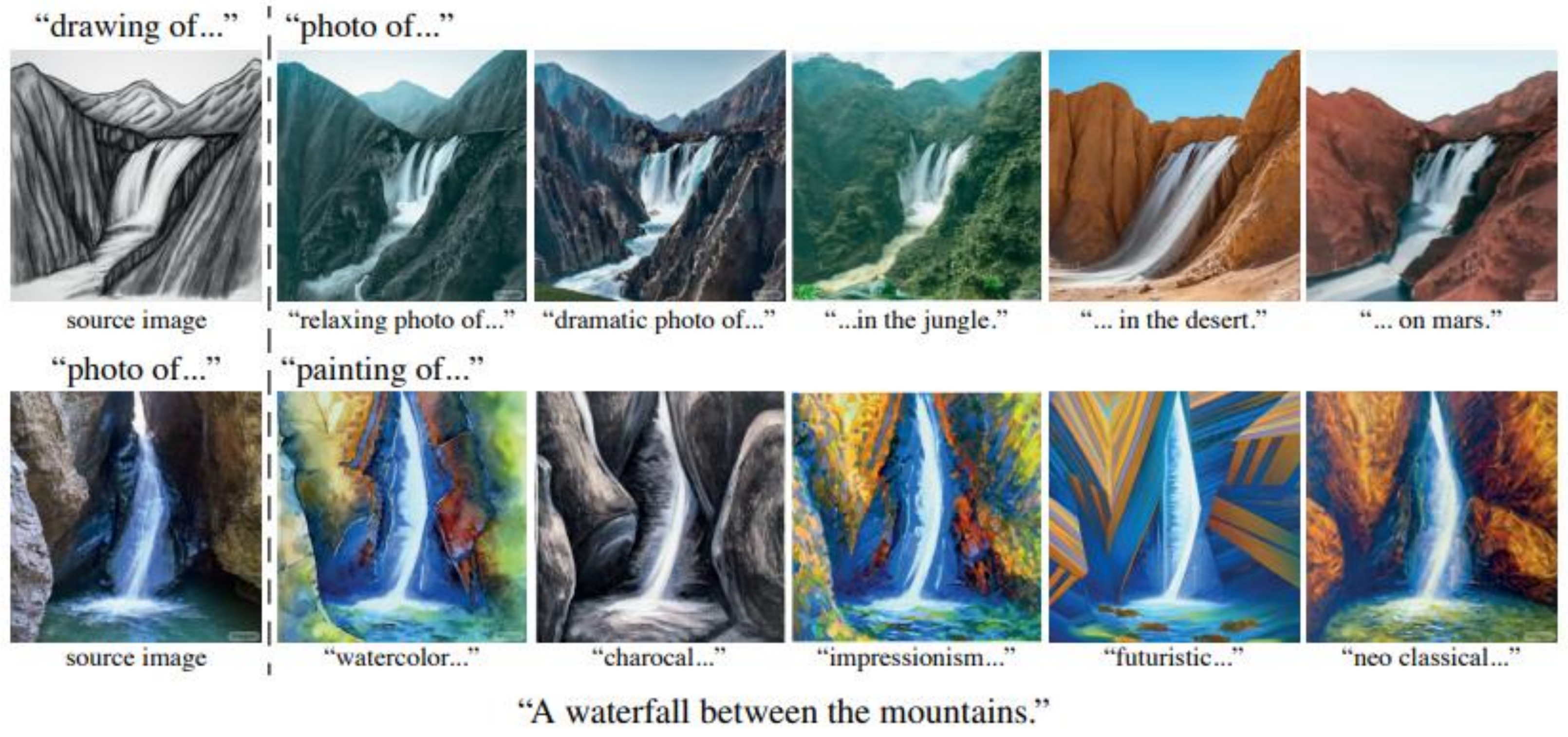


Figure 8: Image stylization. By adding a style description to the prompt while injecting the source attention maps, we can create various images in the new desired styles that preserve the structure of the original image.

④

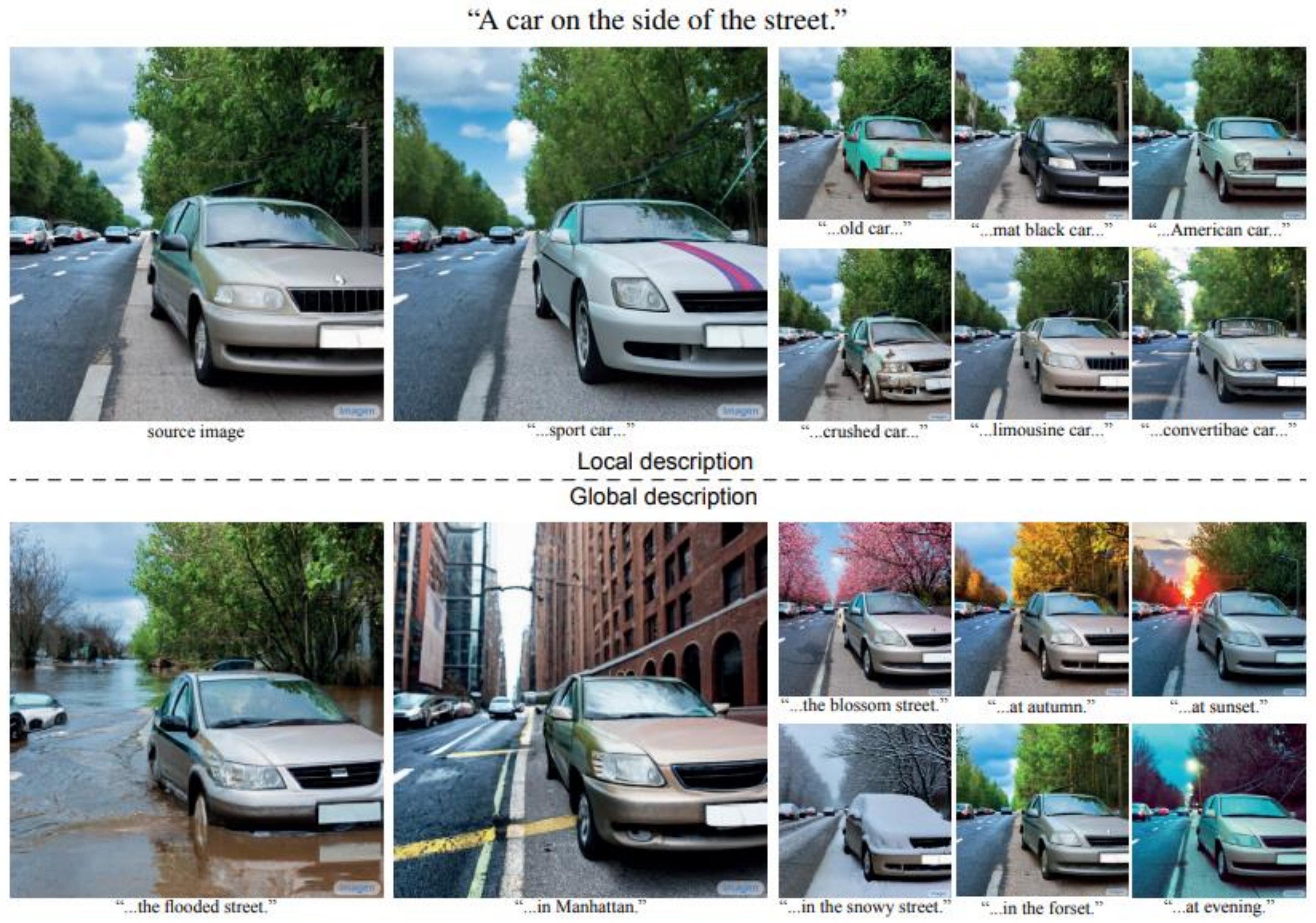


Figure 7: Editing by prompt refinement. By extending the description of the initial prompt, we can make local edits to the car (top rows) or global modifications (bottom rows).