Diffusion-Based Conditional Image Editing through Optimized Inference with Guidance

Supplementary Document

Hyunsoo Lee[1] Minsoo Kang[1] Bohyung Han[1,2]
[1]ECE & [2]IPAI, Seoul National University
{philip21, kminsoo, bhhan}@snu.ac.kr

In the Appendix, we analyze the runtime and computational complexity of our method and compare it with state-of-the-art methods [2, 3, 6, 7, 12] in Section A. Section B introduces additional component, referred to as coherence guidance, which can be combined with the proposed method to enhance the quality of the target image. In section C, we visualize additional ablation study results. Section D demonstrates the qualitative results of our method combined with other pretrained diffusion models other than Stable Diffusion [8] to highlight the generalizability of the proposed method. Additional qualitative results are provided in Section E. Finally, we discuss the limitations and potential social impacts of the proposed method in Section F and G, respectively.

A. Analysis on computational complexity

In Table 3, we report the runtime and computational complexity of the proposed method and state-of-the-art methods [2,3,6,7,12] analyzed on a single NVIDIA A100 GPU. As shown in the table, our method shows comparable computational cost to prior works. Since the proposed method shows superior performance compared to prior works, this demonstrates that our method is a simple but effective approach.

B. Discussion on coherence guidance

B.1. Revised target generation

Different from previous frameworks [3,4,12] that revise the backward process only, the revised method alternates the estimation of the source and target latents in the order of $\{\bar{\mathbf{x}}^{\text{src}}_{T-t}, \bar{\mathbf{x}}^{\text{tgt}}_{t}\}_{t=T-1:0}$, where $\bar{\mathbf{x}}^{\text{src}}_{T-t}$ and $\bar{\mathbf{x}}^{\text{tgt}}_{t}$ are source and target latents obtained from our modified forward and backward processes as shown in Figure 7. Note that, in the case of $t = T$, $\bar{\mathbf{x}}^{\text{src}}_{0}$ is equal to $\mathbf{x}^{\text{src}}_{0}$, which is the source image, while $\bar{\mathbf{x}}^{\text{tgt}}_{T}$ is set to $\mathbf{x}^{\text{src}}_{T}$, which is given by recursively performing the deterministic DDIM inversion from the source image. The two modified processes are denoted by *forward with guidance* and *backward with guidance*. We refer our revised method to Optimized Inference with Guidance^{+} (OIG^{+}). Algorithm 2 summarizes the detailed procedures of the proposed guidances.

B.1.1 Forward with guidance

We revise the forward process in Eq. (1) by additionally optimizing the proposed efficient version of the cycle-consistency objective $\mathcal{L}^{\text{cycle, eff}}$ as described in Section B.3 with respect to the source latent $\bar{\mathbf{x}}^{\text{src}}_{T-t}$ as follows:

$$\begin{aligned}\bar{\mathbf{x}}^{\text{src}}_{T-t+1} =& \bar{f}^{\text{fwd}}_{T-t}(\bar{\mathbf{x}}^{\text{src}}_{T-t}) \\ =& \sqrt{\frac{\alpha_{T-t+1}}{\alpha_{T-t}}}\bar{\mathbf{x}}^{\text{src}}_{T-t} \\ & - \sqrt{1-\alpha_{T-t}}\gamma'_{T-t}\epsilon_\theta(\bar{\mathbf{x}}^{\text{src}}_{T-t}, T-t, \mathbf{y}^{\text{src}}) \\ & - \nabla_{\bar{\mathbf{x}}^{\text{src}}_{T-t}}\lambda_3\mathcal{L}^{\text{cycle, eff}},\end{aligned} \tag{7}$$

where $\bar{f}^{\text{fwd}}_{T-t}(\cdot)$ denotes the modified forward process at time step $T-t$, γ'_{T-t} is equal to $\sqrt{\frac{\alpha_{T-t+1}}{\alpha_{T-t}}} - \sqrt{\frac{1-\alpha_{T-t+1}}{1-\alpha_{T-t}}}$, and λ_3 is a hyperparameter.

B.1.2 Backward with guidance

In the backward with guidance process, we improve the backward process in Eq. (2) by optimizing both the distance term $\mathcal{L}^{\text{dist}}_t$ and the cycle-consistency objective $\mathcal{L}^{\text{cycle}}$, where the modified backward process is given by

$$\begin{aligned}\bar{\mathbf{x}}^{\text{tgt}}_{t-1} =& \bar{f}^{\text{bwd}}_{t}(\bar{\mathbf{x}}^{\text{tgt}}_{t}) \\ =& \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\bar{\mathbf{x}}^{\text{tgt}}_{t} - \sqrt{1-\alpha_t}\gamma_t\epsilon_\theta(\bar{\mathbf{x}}^{\text{tgt}}_{t}, t, \mathbf{y}^{\text{tgt}}) \\ & - \nabla_{\bar{\mathbf{x}}^{\text{tgt}}_{t}}(\mathcal{L}^{\text{dist}}_t + \lambda_4\mathcal{L}^{\text{cycle}}),\end{aligned} \tag{8}$$

where $\bar{f}^{\text{bwd}}_{t}(\cdot)$ is defined by our modified backward process at time step t, γ_t is equal to $\sqrt{\frac{\alpha_{t-1}}{\alpha_t}} - \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}$, and λ_4 is

Table 3. Runtime and computational complexity analysis of DiffEdit [3], Plug-and-Play [12] Pix2Pix-Zero [7], Null-text inversion [6], MasaCtrl [2] and the proposed method. Each algorithm is tested on a single NVIDIA A100 GPU. The proposed method achieves comparable runtime and memory consumption compared to prior works, while outperforming prior works.

Method	DiffEdit	Plug-and-Play	Pix2Pix-Zero	Null-text inversion	MasaCtrl	OIG (Ours)
time/image (s)	18.24	19.78	30.53	105.80	16.08	37.89
GPU Memory (GB)	4.188	4.103	11.975	10.110	10.68	8.363

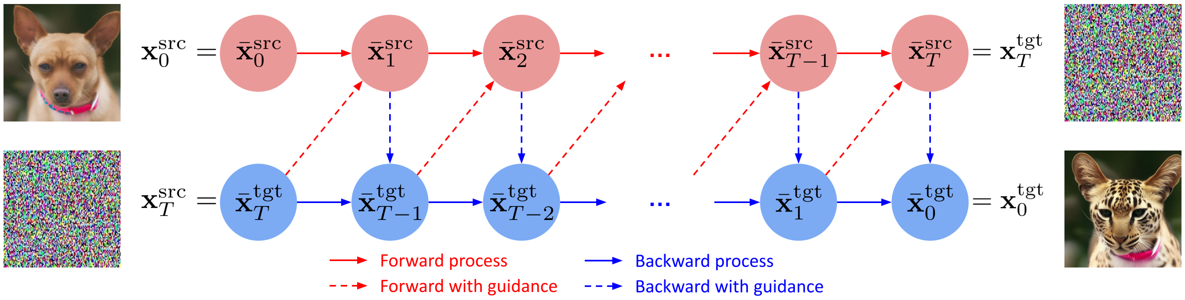

$\mathbf{x}_0^{\text{src}} = \bar{\mathbf{x}}_0^{\text{src}}$
$\bar{\mathbf{x}}_1^{\text{src}}$
$\bar{\mathbf{x}}_2^{\text{src}}$
...
$\bar{\mathbf{x}}_{T-1}^{\text{src}}$
$\bar{\mathbf{x}}_T^{\text{src}} = \mathbf{x}_T^{\text{tgt}}$
$\mathbf{x}_T^{\text{src}} = \bar{\mathbf{x}}_T^{\text{tgt}}$
$\bar{\mathbf{x}}_{T-1}^{\text{tgt}}$
$\bar{\mathbf{x}}_{T-2}^{\text{tgt}}$
...
$\bar{\mathbf{x}}_1^{\text{tgt}}$
$\bar{\mathbf{x}}_0^{\text{tgt}} = \mathbf{x}_0^{\text{tgt}}$
Forward process
Forward with guidance
Backward process
Backward with guidance

Figure 7. Overview of the revised method about the forward guidance and backward guidance.

Algorithm 2 Text-Driven Image Editing based on Forward and Backward Guidances

```
Inputs: A source image $\mathbf{x}_0^{\text{src}}$, a source prompt $\mathbf{y}^{\text{src}}$, a target prompt $\mathbf{y}^{\text{tgt}}$
for $t \leftarrow 0, \cdots, T-1$ do
    Compute $\mathbf{x}_{t+1}^{\text{src}}$ using Eq. (1)
end for
$\bar{\mathbf{x}}_T^{\text{tgt}} \leftarrow \mathbf{x}_T^{\text{src}}$ and $\bar{\mathbf{x}}_0^{\text{src}} \leftarrow \mathbf{x}_0^{\text{src}}$
for $t \leftarrow T, \cdots, 1$ do
    Calculate $\mathcal{L}^{\text{cycle, eff}}$ using Eq. (13)
    Compute $\gamma'_{T-t} \leftarrow \sqrt{\frac{\alpha_{T-t+1}}{\alpha_{T-t}}} - \sqrt{\frac{1-\alpha_{T-t+1}}{1-\alpha_{T-t}}}$
    Calculate $\bar{\mathbf{x}}_{T-t+1}^{\text{src}}$ using Eq. (7)
                                        ▷ Forward with guidance

    Calculate $\hat{\mathbf{x}}_0(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}})$ using Eq. (5)
    Calculate $\mathcal{L}_t^{\text{dist}}$ using Eq. (6)
    Compute $\mathcal{L}^{\text{cycle}}$ using Eq. (9)
    Compute $\gamma_t \leftarrow \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} - \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}$
    Compute $\bar{\mathbf{x}}_{t-1}^{\text{tgt}}$ using Eq. (8)
                                        ▷ Backward with guidance
end for
$\mathbf{x}_0^{\text{tgt}} \leftarrow \bar{\mathbf{x}}_0^{\text{tgt}}$
Output: A target image $\mathbf{x}_0^{\text{tgt}}$
```

a hyperparameter. $\mathcal{L}_t^{\text{dist}}$ is the distance objective defined in Eq. (6) of Section 3.4. We will discuss $\mathcal{L}^{\text{cycle}}$ in Section B.2.

B.2. Coherence guidance via cycle-consistency

The *simple DDIM translation*, recursively using the backward process defined in Eq. (2) from the final target latent $\mathbf{x}_T^{\text{tgt}}$, guarantees the cycle-consistency property as verified by [11]. In other words, after converting the source domain image $\mathbf{x}_0^{\text{src}}$ into $\mathbf{x}_0^{\text{tgt}}$ in the target domain and then transforming $\mathbf{x}_0^{\text{tgt}}$ back to the source domain image denoted by $\hat{\mathbf{x}}_0^{\text{src}}$, the equality $\mathbf{x}_0^{\text{src}} = \hat{\mathbf{x}}_0^{\text{src}}$ holds.

Although the simple DDIM translation guarantees the cycle-consistency, the property fails to hold in OIG because the generation process is modified by incorporating the representation guidance. Hence, we add an objective to enforce the cycle-consistency to further enhance translation results. As described in CycleGAN [13], the cycle-consistency term is defined as $\|\mathbf{x}_0^{\text{src}} - h(g(\mathbf{x}_0^{\text{src}}))\|_{2,2}$ in principle, where $g(\cdot)$ is the image-to-image translation operation from the source domain to the target domain, and vise versa for $h(\cdot)$. However, with the assumption that $g(\cdot)$ and $h(\cdot)$ are invertible, we alternatively optimize the following cycle-consistency objective, which is given by

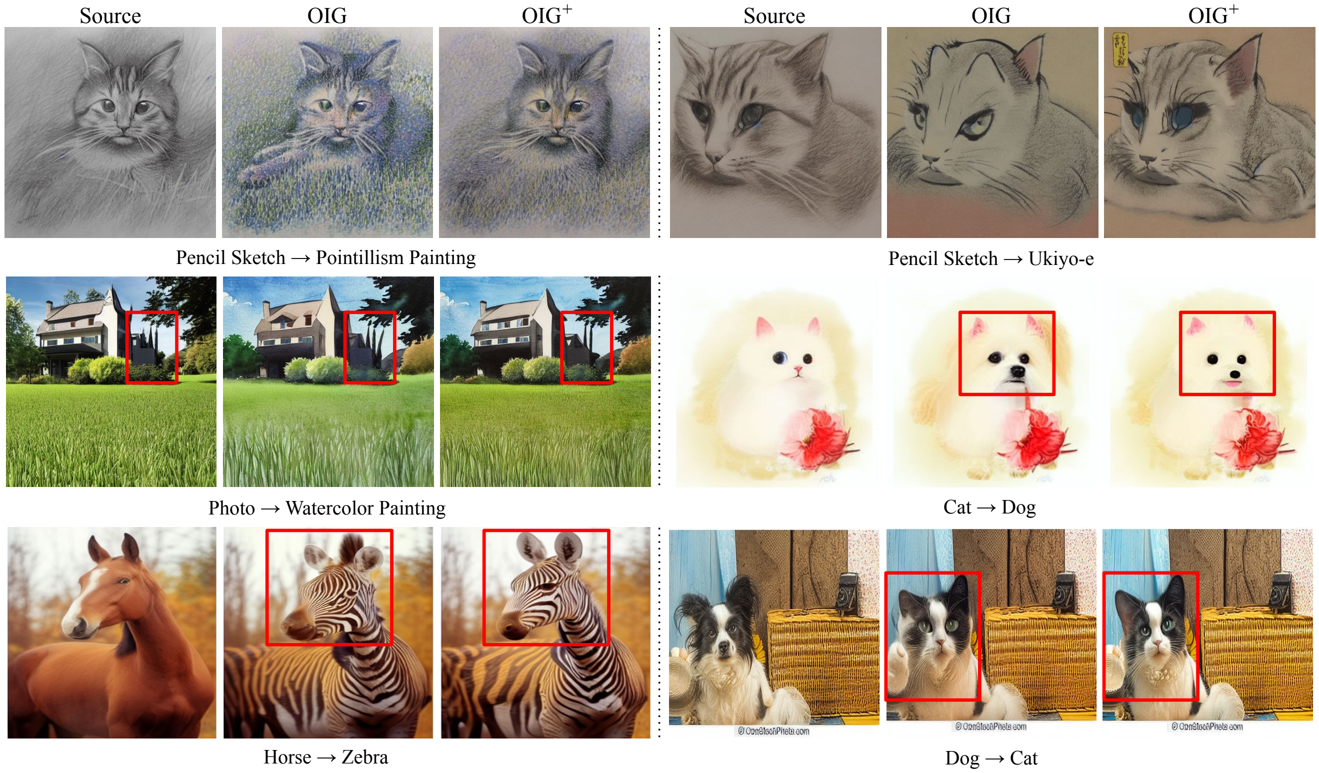
$$\mathcal{L}^{\text{cycle}} := \|\bar{\mathbf{x}}_{0,f}^{\text{tgt}} - \bar{\mathbf{x}}_{0,b}^{\text{tgt}}\|_{2,2}, \tag{9}$$

where we denote $\bar{\mathbf{x}}_{0,f}^{\text{tgt}}$ by $h^{-1}(\mathbf{x}_0^{\text{src}})$ and $\bar{\mathbf{x}}_{0,b}^{\text{tgt}}$ by $g(\mathbf{x}_0^{\text{src}})$. The definition of $h^{-1}(\cdot)$ and $g(\cdot)$ are given by

$$\begin{aligned}\bar{\mathbf{x}}_{0,f}^{\text{tgt}} &= h^{-1}(\mathbf{x}_0^{\text{src}}) = F^{\text{bwd}}(\bar{F}^{\text{fwd}}(\mathbf{x}_0^{\text{src}})) \\ \bar{\mathbf{x}}_{0,b}^{\text{tgt}} &= g(\mathbf{x}_0^{\text{src}}) = \bar{F}^{\text{bwd}}(F^{\text{fwd}}(\mathbf{x}_0^{\text{src}})),\end{aligned} \tag{10}$$

where the auxiliary functions are defined as

$$\begin{aligned}F^{\text{fwd}}(\cdot) &= f_{T-1}^{\text{fwd}} \circ f_{T-2}^{\text{fwd}} \cdots \circ f_0^{\text{fwd}}(\cdot), \\ F^{\text{bwd}}(\cdot) &= f_1^{\text{bwd}} \circ f_2^{\text{bwd}} \cdots \circ f_T^{\text{bwd}}(\cdot) \\ \bar{F}^{\text{fwd}}(\cdot) &= \bar{f}_{T-1}^{\text{fwd}} \circ \bar{f}_{T-2}^{\text{fwd}} \cdots \circ \bar{f}_0^{\text{fwd}}(\cdot) \\ \bar{F}^{\text{bwd}}(\cdot) &= \bar{f}_1^{\text{bwd}} \circ \bar{f}_2^{\text{bwd}} \cdots \circ \bar{f}_T^{\text{bwd}}(\cdot).\end{aligned}$$



Source OIG OIG$^+$

Pencil Sketch → Pointillism Painting

Pencil Sketch → Ukiyo-e

Photo → Watercolor Painting

Cat → Dog

Horse → Zebra

Dog → Cat

Figure 8. Qualitative results of coherence guidance on the data from LAION-5B dataset [9] and synthetic images using the pretrained Stable Diffusion [8]. Coherence guidance effectively modifies the details of the target image when combined with OIG.

Estimation of $\bar{\mathbf{x}}_{0,f}^{\text{tgt}}$ Using the equivalent ordinary differential equation of the simple DDIM forward process in Eq (1), we first approximate $\bar{\mathbf{x}}_{T,f}^{\text{tgt}}$ which is equal to $\bar{\mathbf{x}}_{T}^{\text{src}}$ as

$$\bar{\mathbf{x}}_{T,f}^{\text{tgt}} = \bar{\mathbf{x}}_{T}^{\text{src}} \approx \sqrt{\frac{\alpha_T}{\alpha_{t_0}}}\bar{\mathbf{x}}_{t_0}^{\text{src}} + \left(\sqrt{1-\alpha_T} - \sqrt{\frac{\alpha_T(1-\alpha_{t_0})}{\alpha_{t_0}}}\right)\epsilon_\theta(\bar{\mathbf{x}}_{t_0}^{\text{src}}, t_0, \mathbf{y}^{\text{src}}), \quad (11)$$

where t_0 is an intermediate time step. Although the approximation incurs discretization errors due to the one-step estimation of $\bar{\mathbf{x}}_{T}^{\text{src}}$ from $\bar{\mathbf{x}}_{t_0}^{\text{src}}$, we empirically observe that the proposed method achieves remarkable performance as demonstrated in Section B.4. When we estimate $\bar{\mathbf{x}}_{t-1}^{\text{tgt}}$ from $\bar{\mathbf{x}}_{t}^{\text{tgt}}$ during the backward with guidance, $\bar{\mathbf{x}}_{T,f}^{\text{tgt}}$ is derived from Eq. (11) by plugging $T-t+1$ into t_0. Finally, $\bar{\mathbf{x}}_{0,f}^{\text{tgt}}$ is obtained by $F^{\text{bwd}}(\bar{\mathbf{x}}_{T,f}^{\text{tgt}})$, where $F^{\text{bwd}}(\cdot)$ performs T steps of the recursive backward process in Eq. (2).

Estimation of $\bar{\mathbf{x}}_{0,b}^{\text{tgt}}$ To reduce the computational costs for the estimation of $\bar{\mathbf{x}}_{0,b}^{\text{tgt}}$ from $\bar{\mathbf{x}}_{t}^{\text{tgt}}$, we approximate $\bar{\mathbf{x}}_{0,b}^{\text{tgt}}$ using the Tweedie's formula [10] in Eq. (5) as

$$\bar{\mathbf{x}}_{0,b}^{\text{tgt}} \approx \hat{\mathbf{x}}_0(\bar{\mathbf{x}}_{t}^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}}). \quad (12)$$

We eventually calculate $\mathcal{L}^{\text{cycle}}$ by plugging $\bar{\mathbf{x}}_{0,f}^{\text{tgt}}$ and $\bar{\mathbf{x}}_{0,b}^{\text{tgt}}$ into Eq. (9).

B.3. Efficient coherence guidance

In the case of the forward with guidance, computing the gradient of $\mathcal{L}^{\text{cycle}}$ in Eq. (9) with respect to the source latent $\bar{\mathbf{x}}_{T-t}^{\text{src}}$, denoted as $\nabla_{\bar{\mathbf{x}}_{T-t}^{\text{src}}}\mathcal{L}^{\text{cycle}}$, is memory-intensive and time-consuming since it involves multiple times of backpropagation through the noise prediction network. To tackle this issue, we alternatively derive the following efficient version of the cycle-consistency objective that matches the final target latents instead of the target images as

$$\mathcal{L}^{\text{cycle, eff}} := \|\bar{\mathbf{x}}_{T,f}^{\text{tgt}} - \bar{\mathbf{x}}_{T,b}^{\text{tgt}}\|_{2,2}. \quad (13)$$

In the above equation, $\bar{\mathbf{x}}_{T,f}^{\text{tgt}}$ is obtained based on a single forward propagation of the noise prediction network from Eq. (11) by setting $t_0 = T-t$. We obtain $\bar{\mathbf{x}}_{T,b}^{\text{tgt}}$ from $F^{\text{fwd}}(\bar{\mathbf{x}}_{0,b}^{\text{tgt}})$, where $\bar{\mathbf{x}}_{0,b}^{\text{tgt}}$ is estimated using Eq. (12) and $F^{\text{fwd}}(\cdot)$ recursively applies the DDIM inversion process in Eq. (1) for T times.

Therefore, we can compute the gradient of $\mathcal{L}^{\text{cycle, eff}}$ with respect to $\bar{\mathbf{x}}_{T-t}^{\text{src}}$ just by performing a single backpropagation through the noise prediction network.

B.4. Qualitative results of OIG$^+$

We visualize the effect of coherence guidance in Figure 8. As shown in the Figure, OIG$^+$ enhances the fine details of the target image generated by OIG, such as reducing high-frequency noise or facilitating the alignment of small structural elements.

C. Additional ablation study

We report additional ablation study results of the proposed method in Figure 9. We emphasize that the triplet-based distance term in Eq. (6) enhances the fidelity of the target image and preserves the overall structure well compared to the naïve distance objective in Eq. (4).

D. Additional results using other pretrained diffusion models

To demonstrate that the proposed method generalizes well to other pretrained models, we generated target images using our method with pretrained Distilled Stable Diffusion[1] and Latent Diffusion Model (LDM) [8]. Note that Distilled Stable Diffusion is a lightweight model that has been trained by reducing the parameters of the denoising U-Net. Also, the pipeline of LDM is similar to Stable Diffusion, however the resolution of training data for LDM differ from those of Stable Diffusion.

As visualized in Figure 10 and 11, the proposed method shows superior performance when combined with Distilled Stable Diffusion and LDM, which demonstrates that our method can generalize well.

E. Additional qualitative results

We present additional qualitative results of OIG in Figure 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24 to compare with state-of-the-art methods [1–3, 6, 7, 12] on real images sampled from the LAION-5B dataset [9] using the pretrained Stable Diffusion [8]. As visualized in the figures, OIG achieves outstanding results while the previous methods often fail to preserve the structure or background of the source images. Figure 25 also verify the effectiveness of OIG on the synthesized images given by the pretrained Stable Diffusion.

F. Limitations

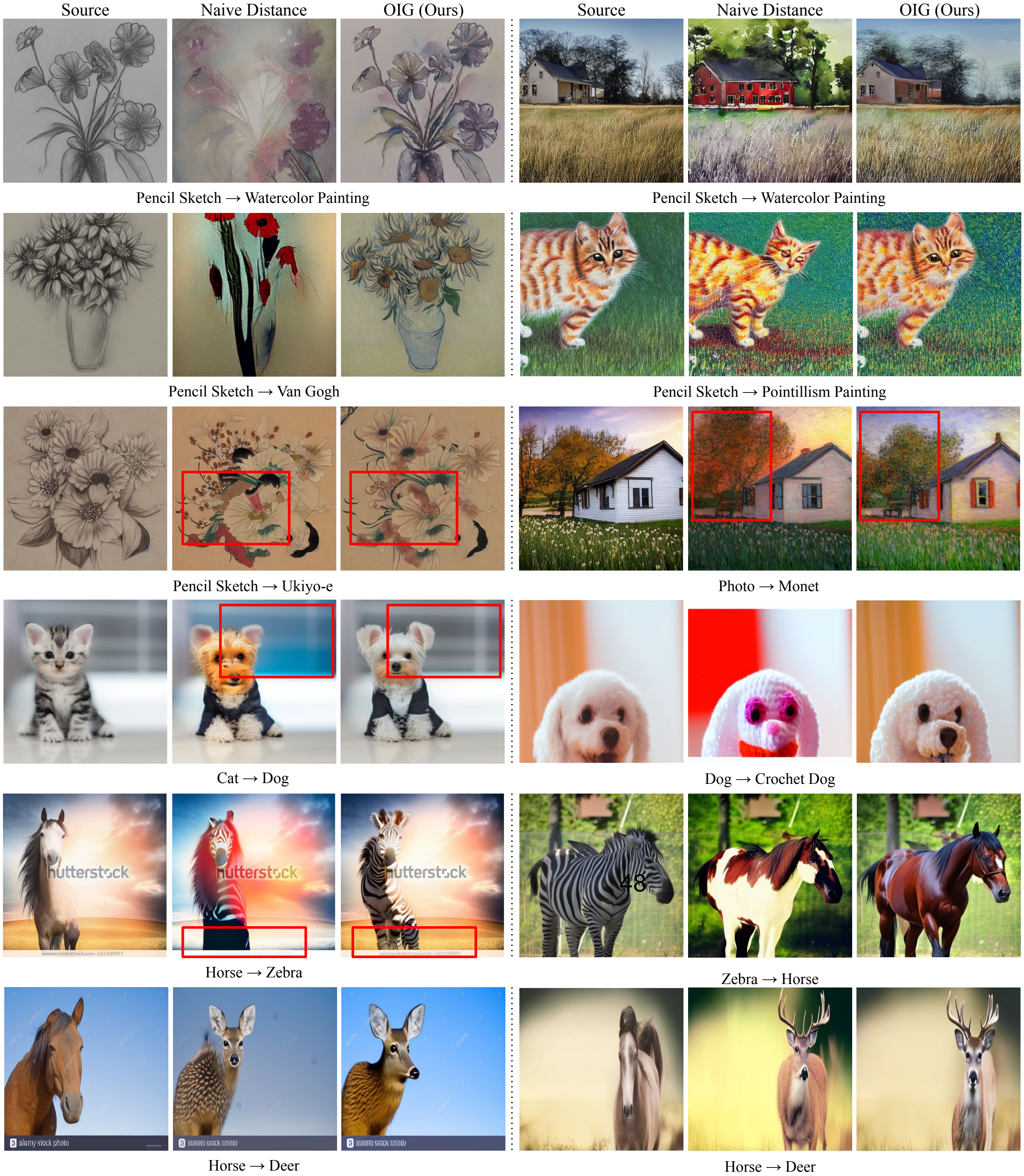
We visualize the failure cases of our method in Figure 12. These failure cases can be addressed by using OIG$^+$, which combines representation guidance and coherence guidance. OIG$^+$ effectively removes the artifacts and resolves inconsistencies in the target image, thereby improving the editing performance of the proposed method.

In addition, since the DDIM inference process sometimes does not completely reconstruct the original image, our method can struggle to preserve the information about the source image and result in suboptimal image-to-image translation results. Furthermore, the performance of the proposed method is reliant on pretrained text-to-image diffusion models, which may limit its ability to generate target images for complex tasks effectively.

G. Social impacts

The proposed method may synthesize undesirable or inappropriate images depending on the pretrained text-to-image generation model [8]. For example, the incompleteness of the pretrained diffusion model can lead to the generation of images that violate ethical regulations.

[1] https://huggingface.co/docs/diffusers/en/using-diffusers/distilled_sd



Pencil Sketch → Watercolor Painting

Pencil Sketch → Watercolor Painting

Pencil Sketch → Van Gogh

Pencil Sketch → Pointillism Painting

Pencil Sketch → Ukiyo-e

Photo → Monet

Cat → Dog

Dog → Crochet Dog

Horse → Zebra

Zebra → Horse

Horse → Deer

Horse → Deer

Figure 9. Qualitative results of our component on the data from LAION-5B dataset [9] and synthetic images using the pretrained Stable Diffusion [8]. Representation guidance significantly improves the details of the target image, such as correcting the structural inconsistencies between source and target images, preserving the structure of foreground region, and enhancing the fidelity.

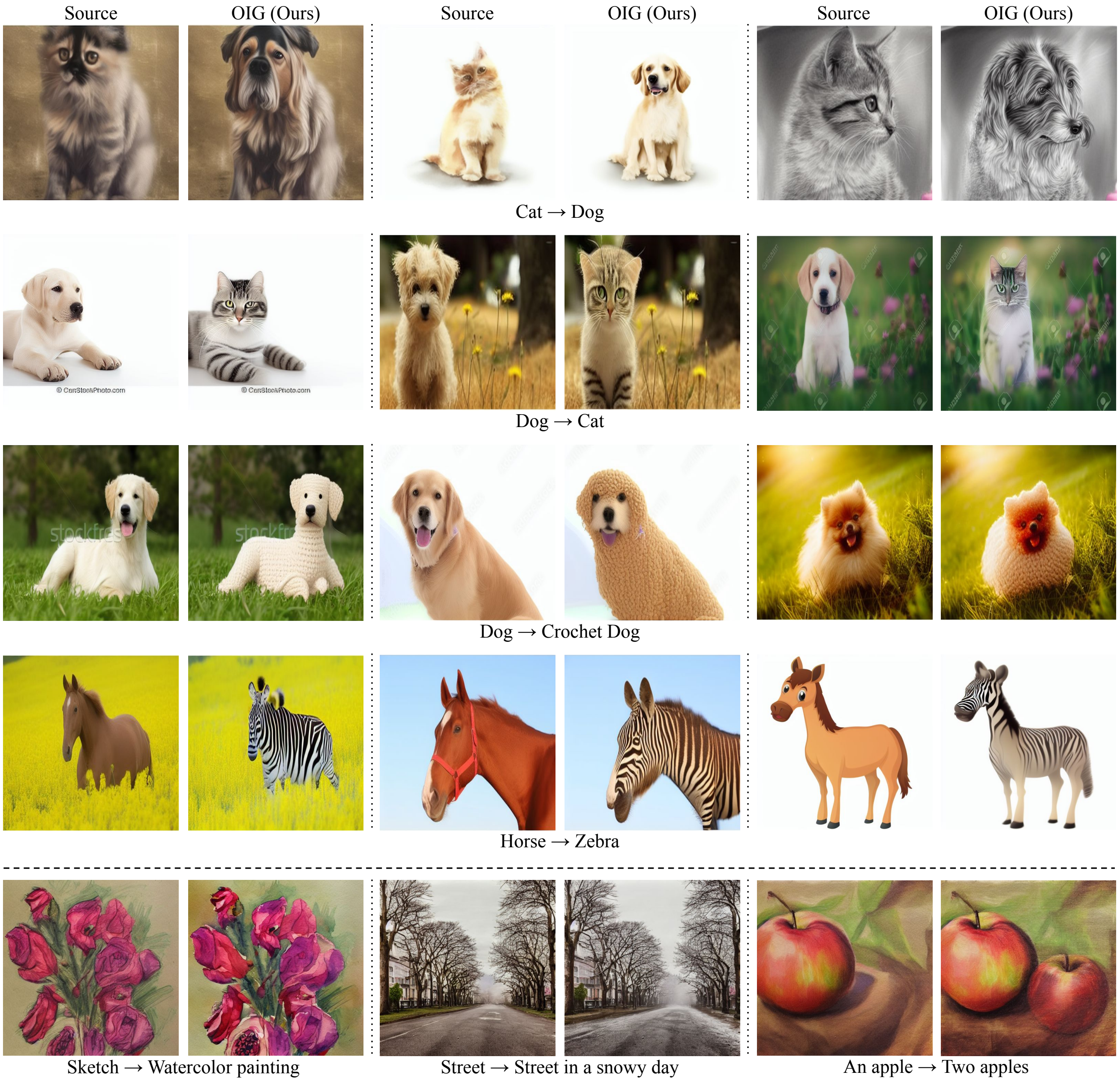


Figure 10. Qualitative results of the proposed method combined with Distilled Stable Diffusion on real images (1st - 4th row) sampled from the LAION-5B dataset [9] and synthetic images (5th row) given by the pretrained Distilled Stable Diffusion.

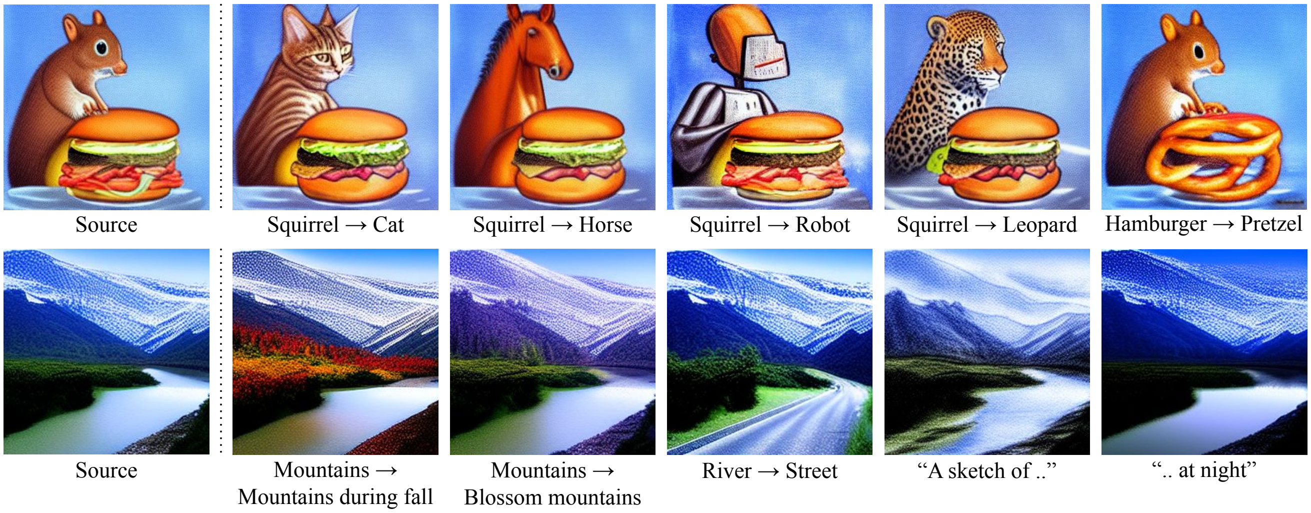


Figure 11. Qualitative results of the proposed method combined with LDM [8] on synthetic images given by the pretrained LDM.

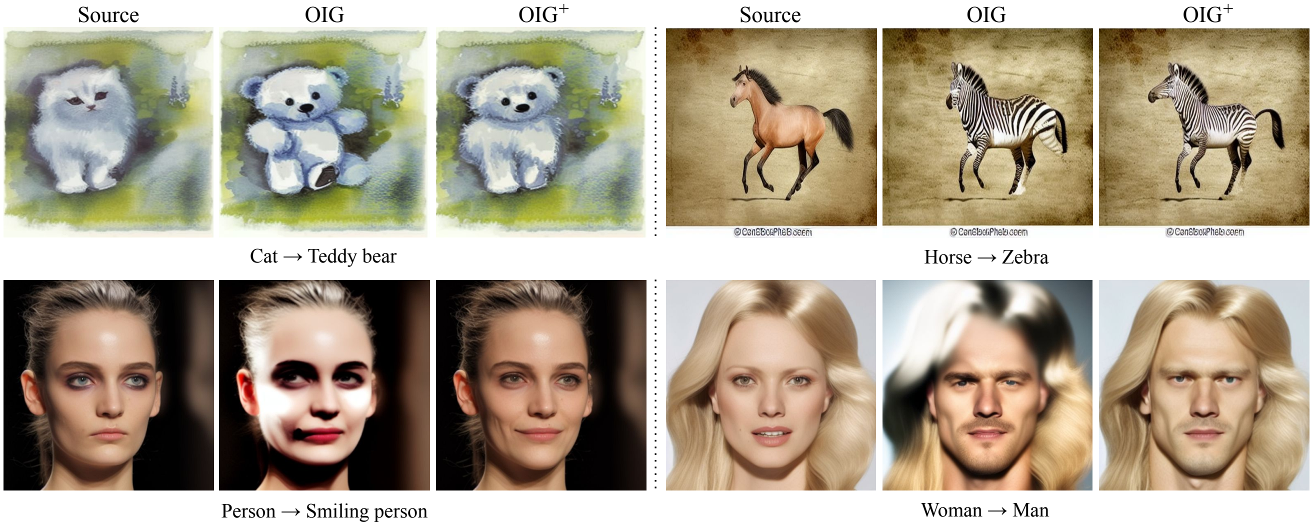


Figure 12. Failure cases of the proposed method on real images sampled from the LAION 5B dataset [9] (1st row) and CelebA-HQ dataset [5] (2nd row). Provided failure samples can be addressed by utilizing OIG$^+$.

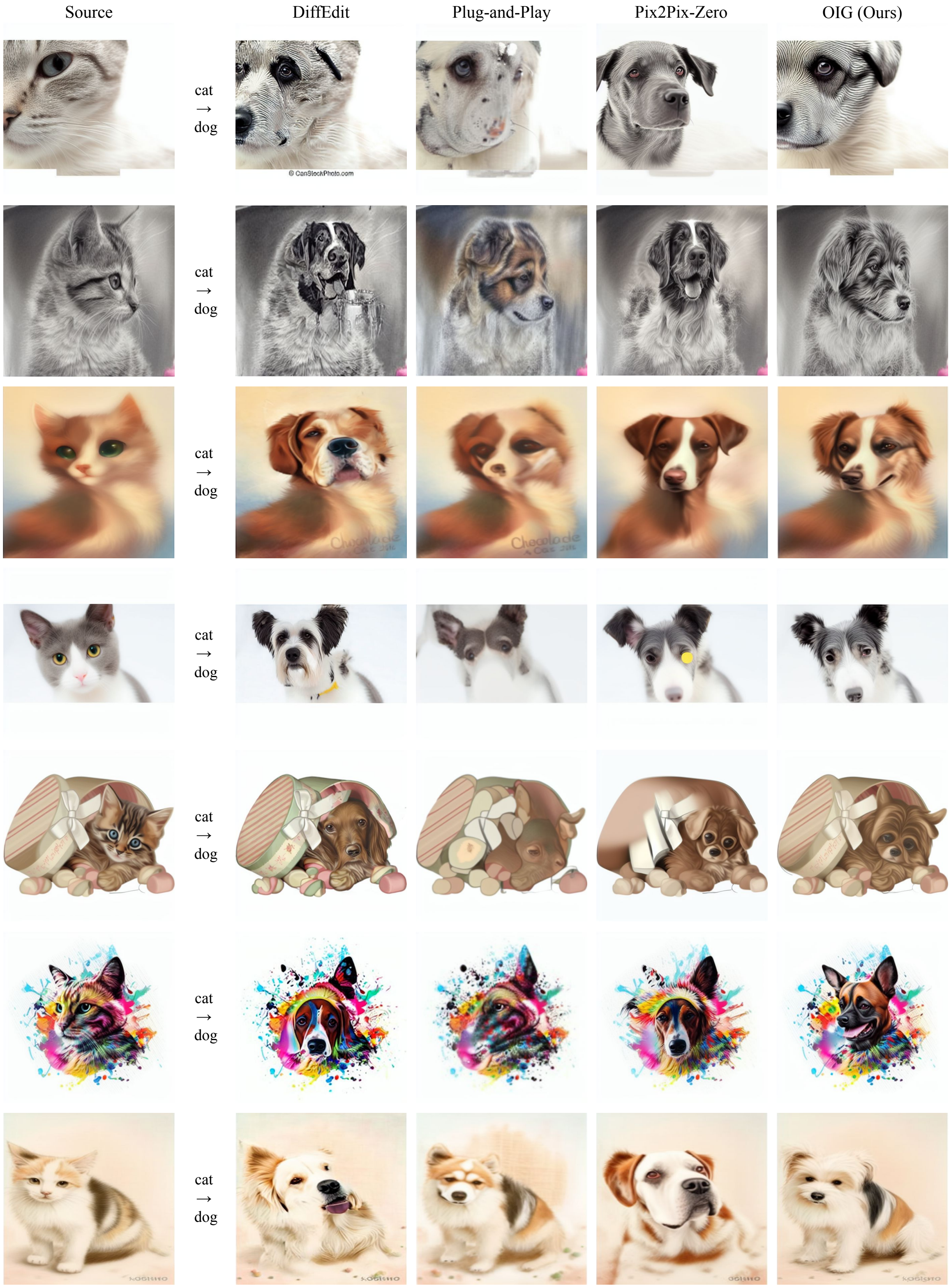


Figure 13. Additional qualitative results of the proposed method, DiffEdit [3], Plug-and-Play [12], and Pix2Pix-Zero [7] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the cat → dog task.

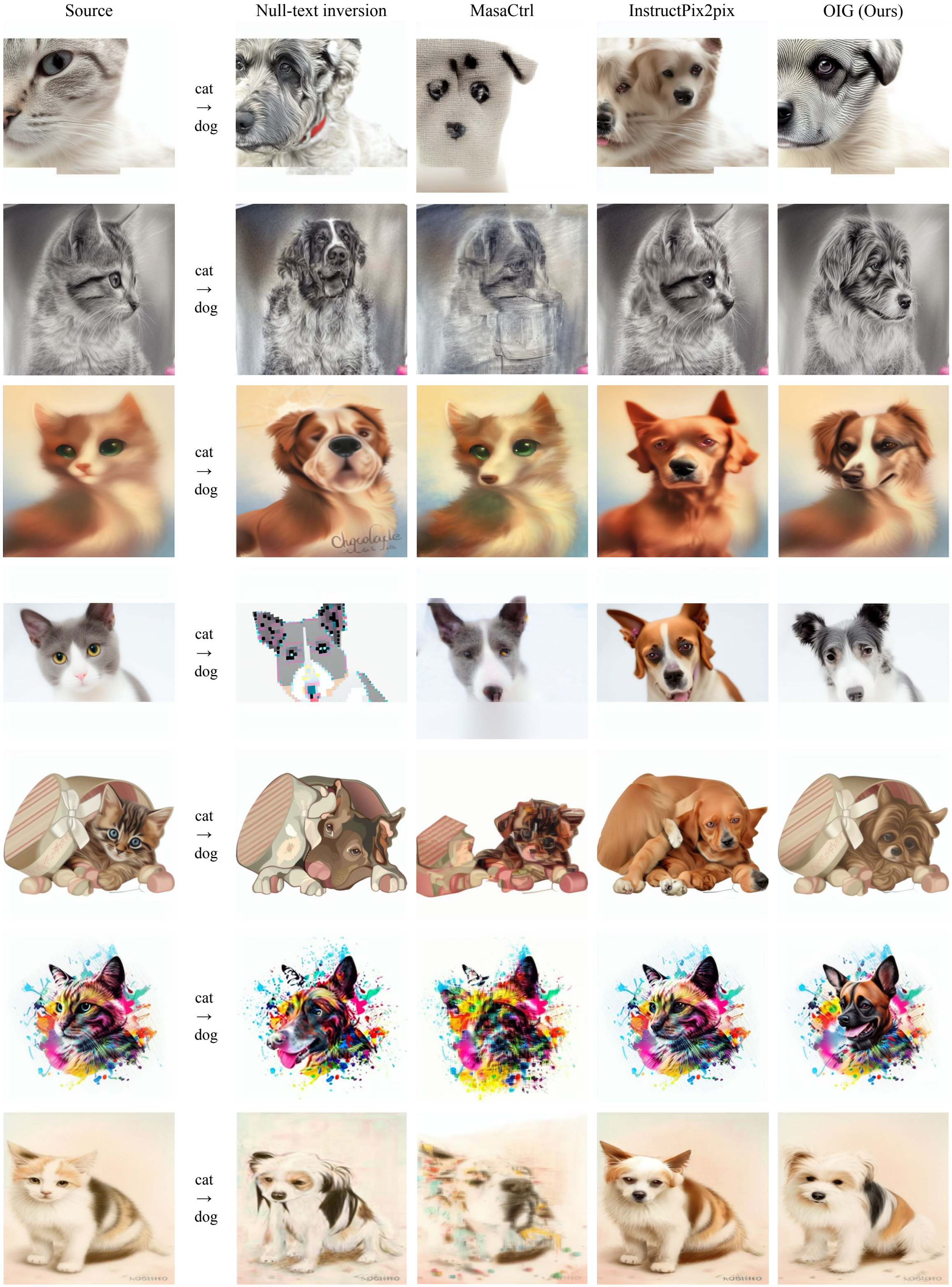


Figure 14. Additional qualitative results of the proposed method, Null-text inversion [6], MasaCtrl [2], and InstructPix2Pix [1] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the cat → dog task.

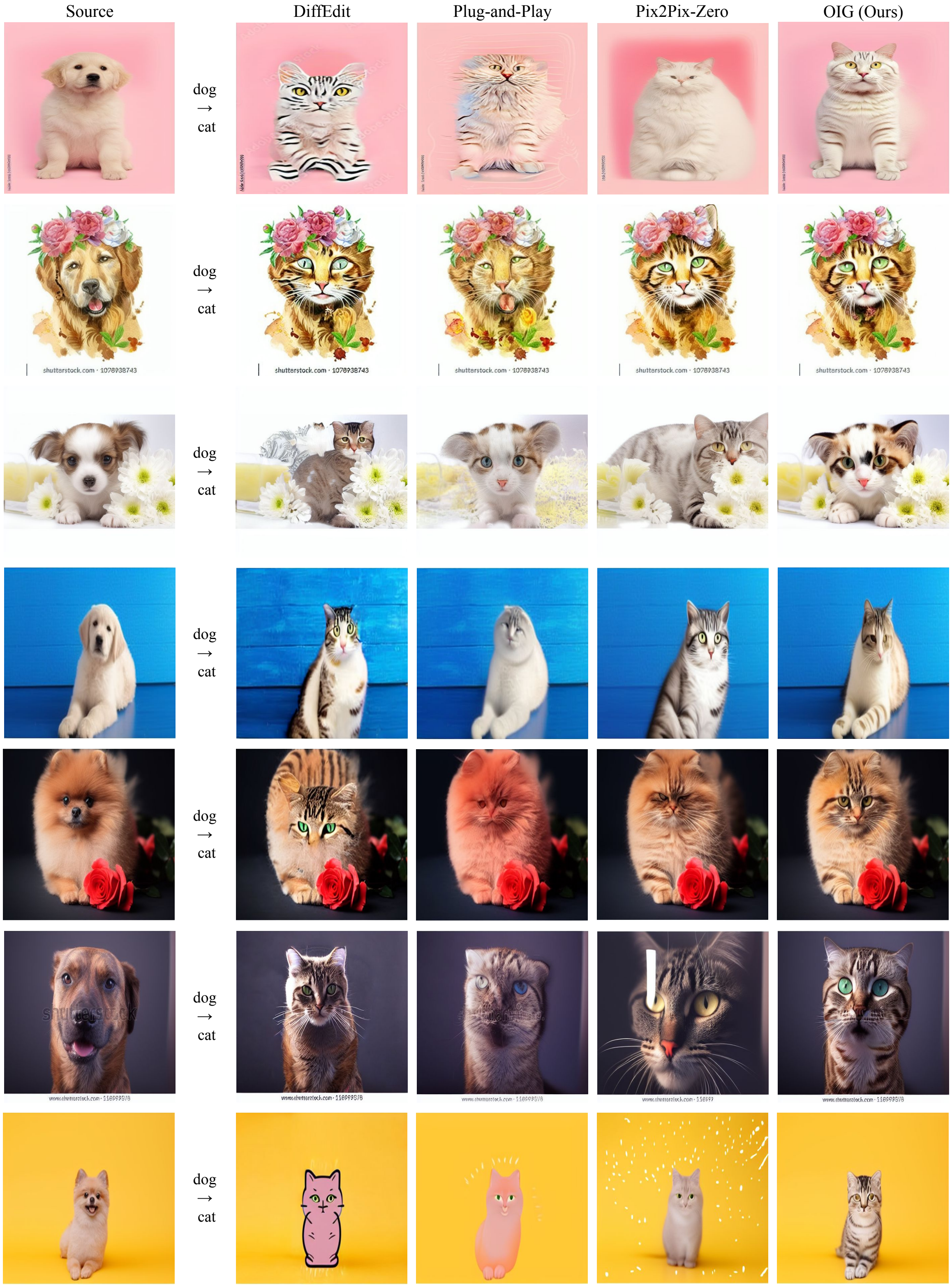


Figure 15. Additional qualitative results of the proposed method, DiffEdit [3], Plug-and-Play [12], and Pix2Pix-Zero [7] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the dog → cat task.

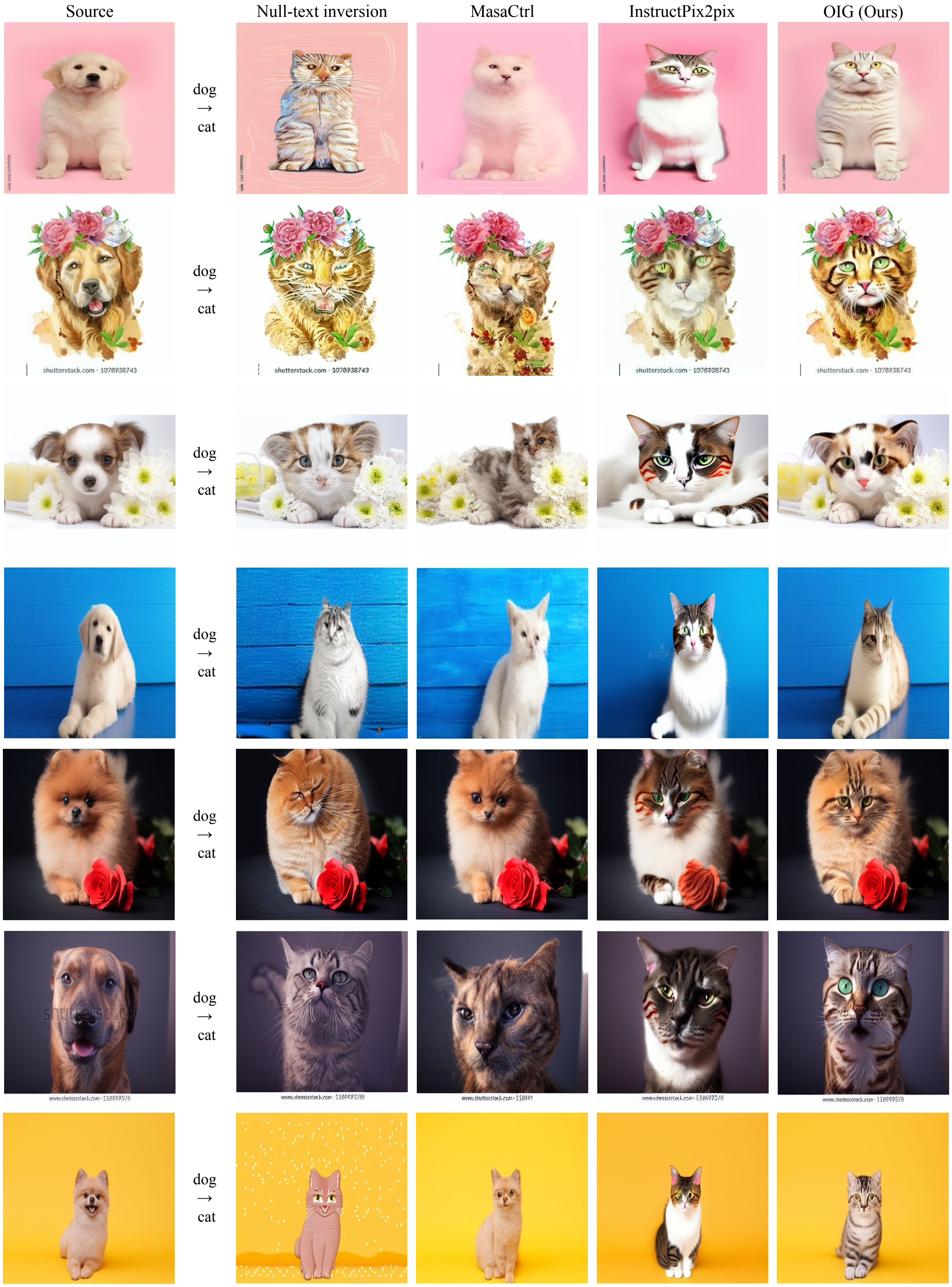


Figure 16. Additional qualitative results of the proposed method, Null-text inversion [6], MasaCtrl [2], and InstructPix2Pix [1] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the dog → cat task.

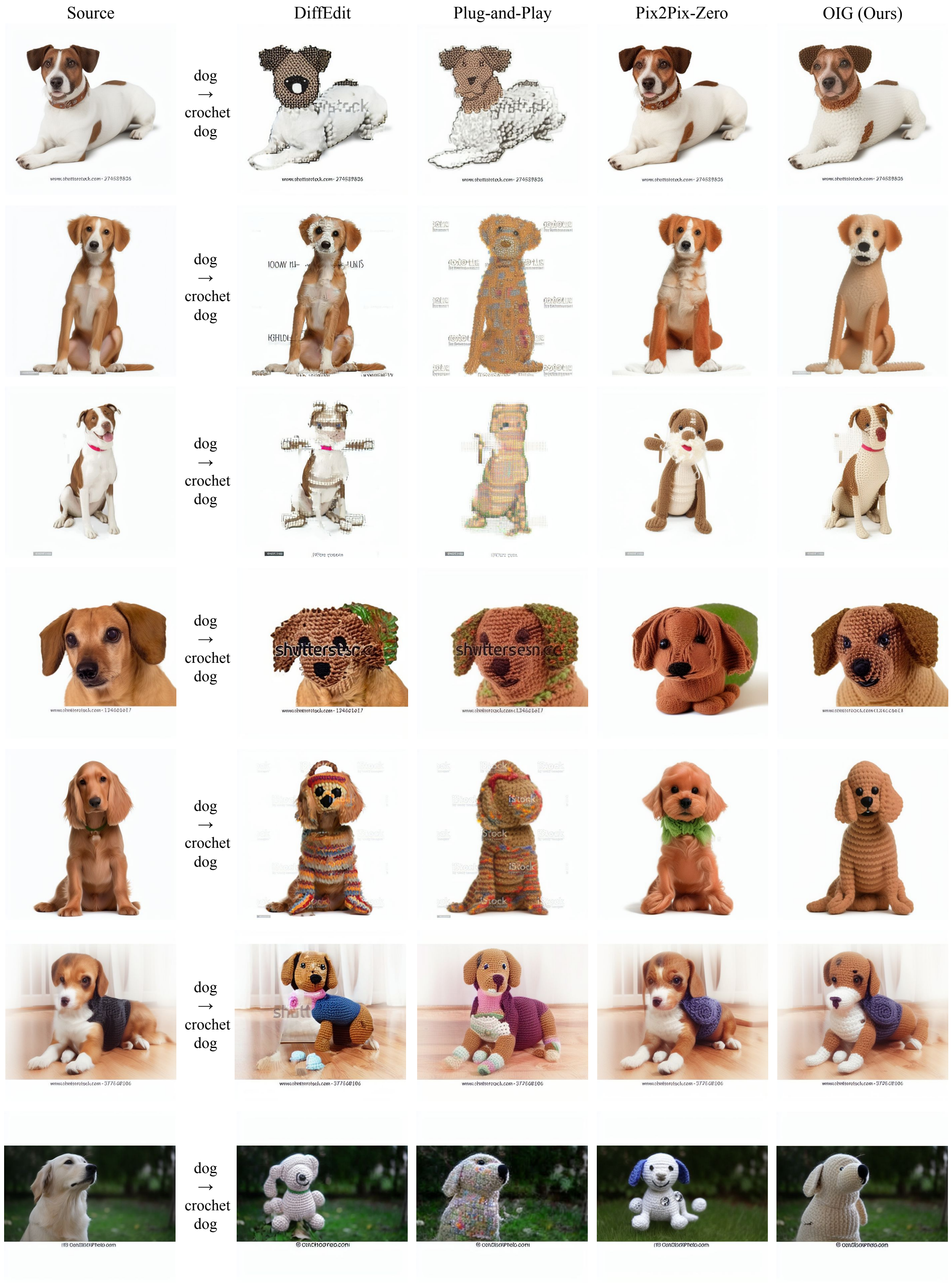


Figure 17. Additional qualitative results of the proposed method, DiffEdit [3], Plug-and-Play [12], and Pix2Pix-Zero [7] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the dog → crochet dog task.

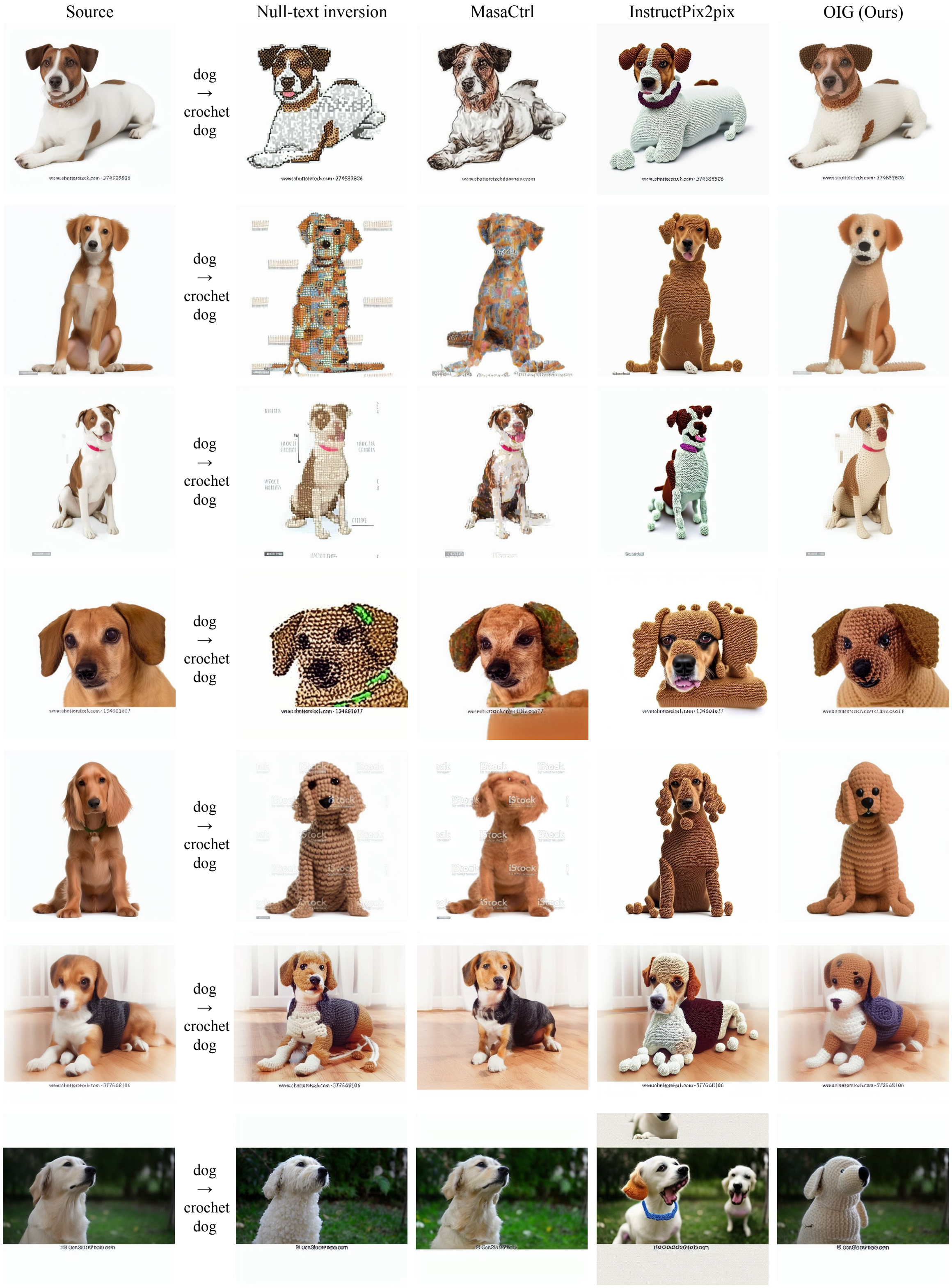


Figure 18. Additional qualitative results of the proposed method, Null-text inversion [6], MasaCtrl [2], and InstructPix2Pix [1] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the dog → crochet dog task.

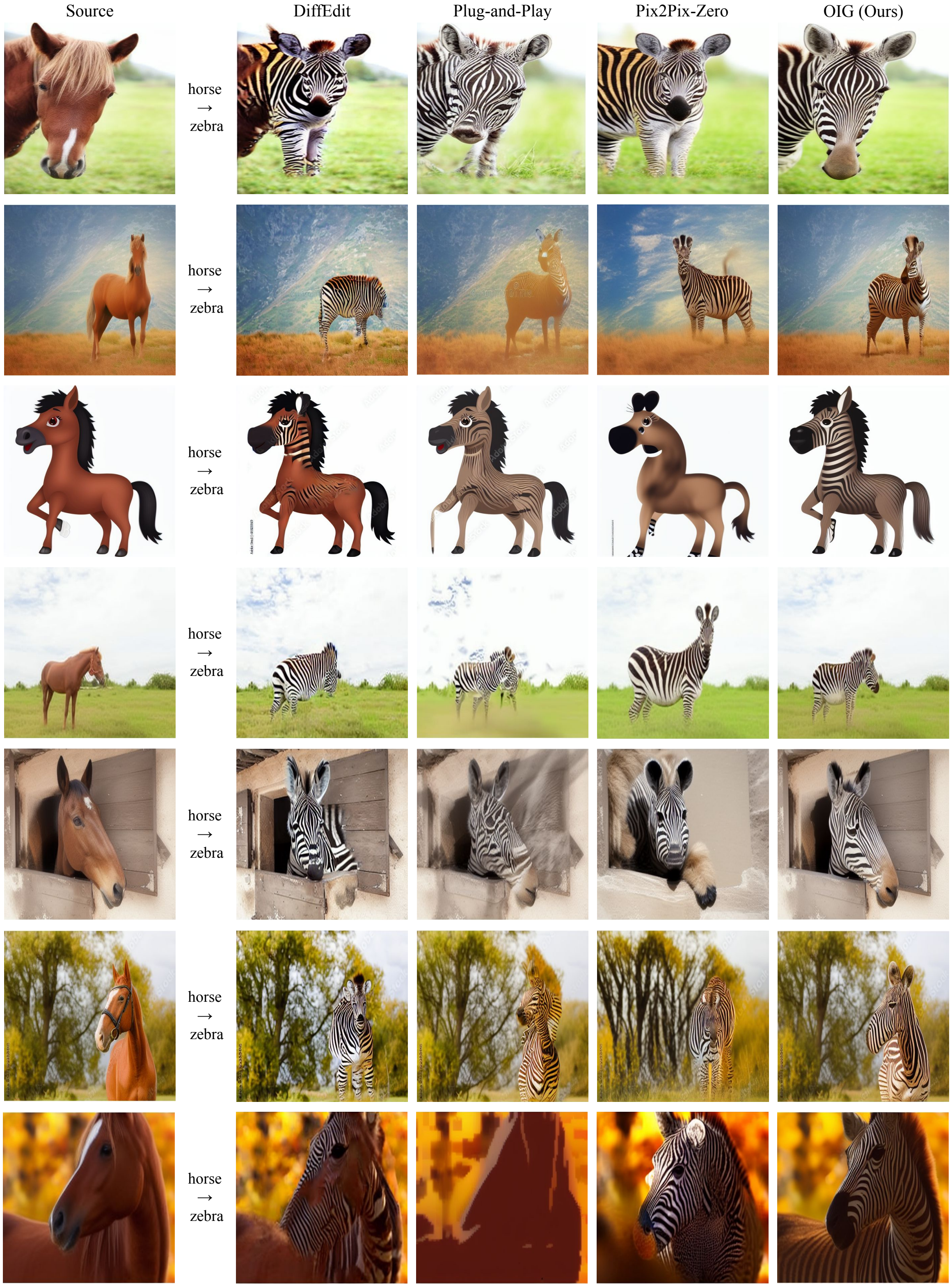


Figure 19. Additional qualitative results of the proposed method, DiffEdit [3], Plug-and-Play [12], and Pix2Pix-Zero [7] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the horse → zebra task.

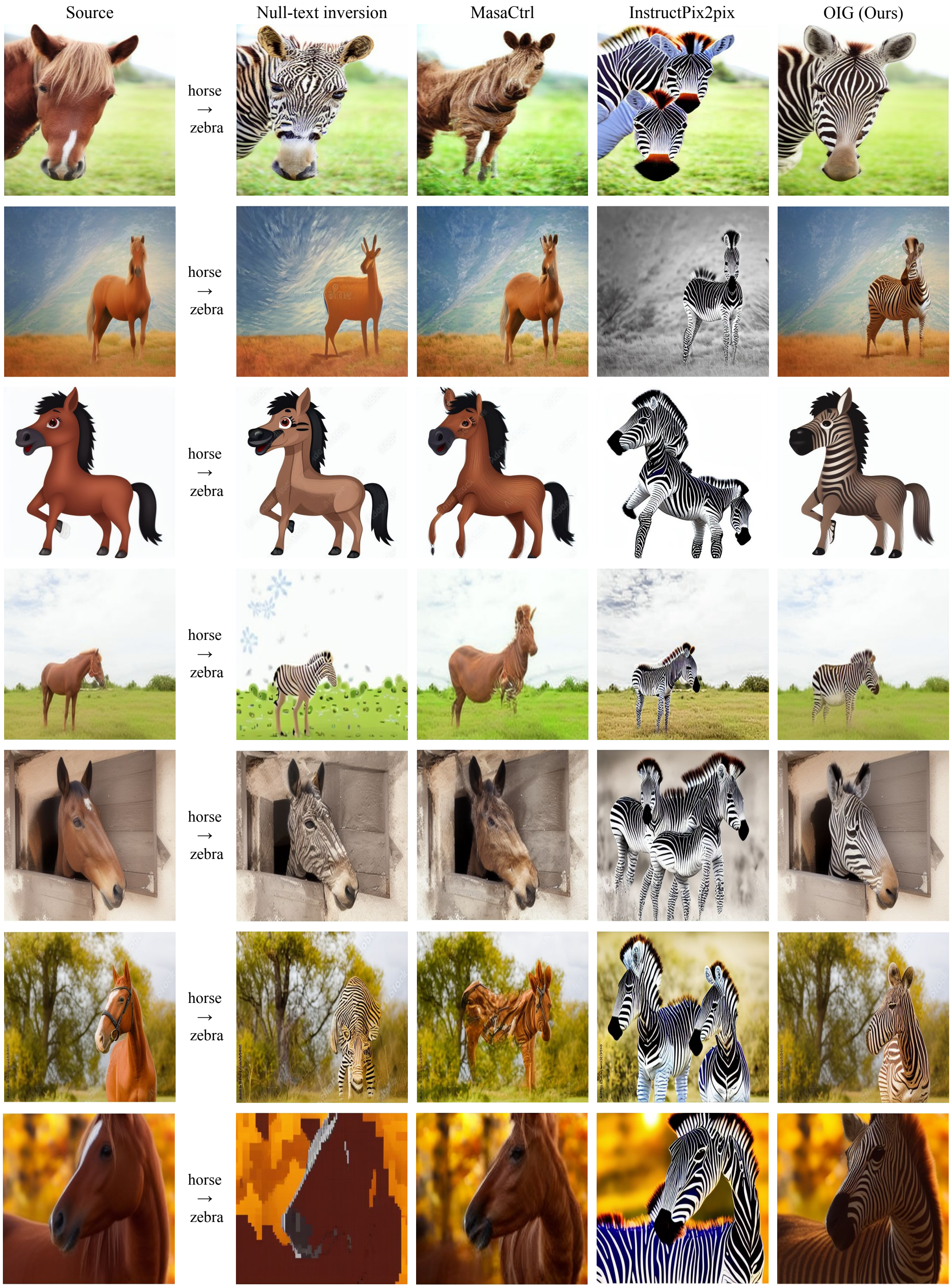


Figure 20. Additional qualitative results of the proposed method, Null-text inversion [6], MasaCtrl [2], and InstructPix2Pix [1] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the horse → zebra task.

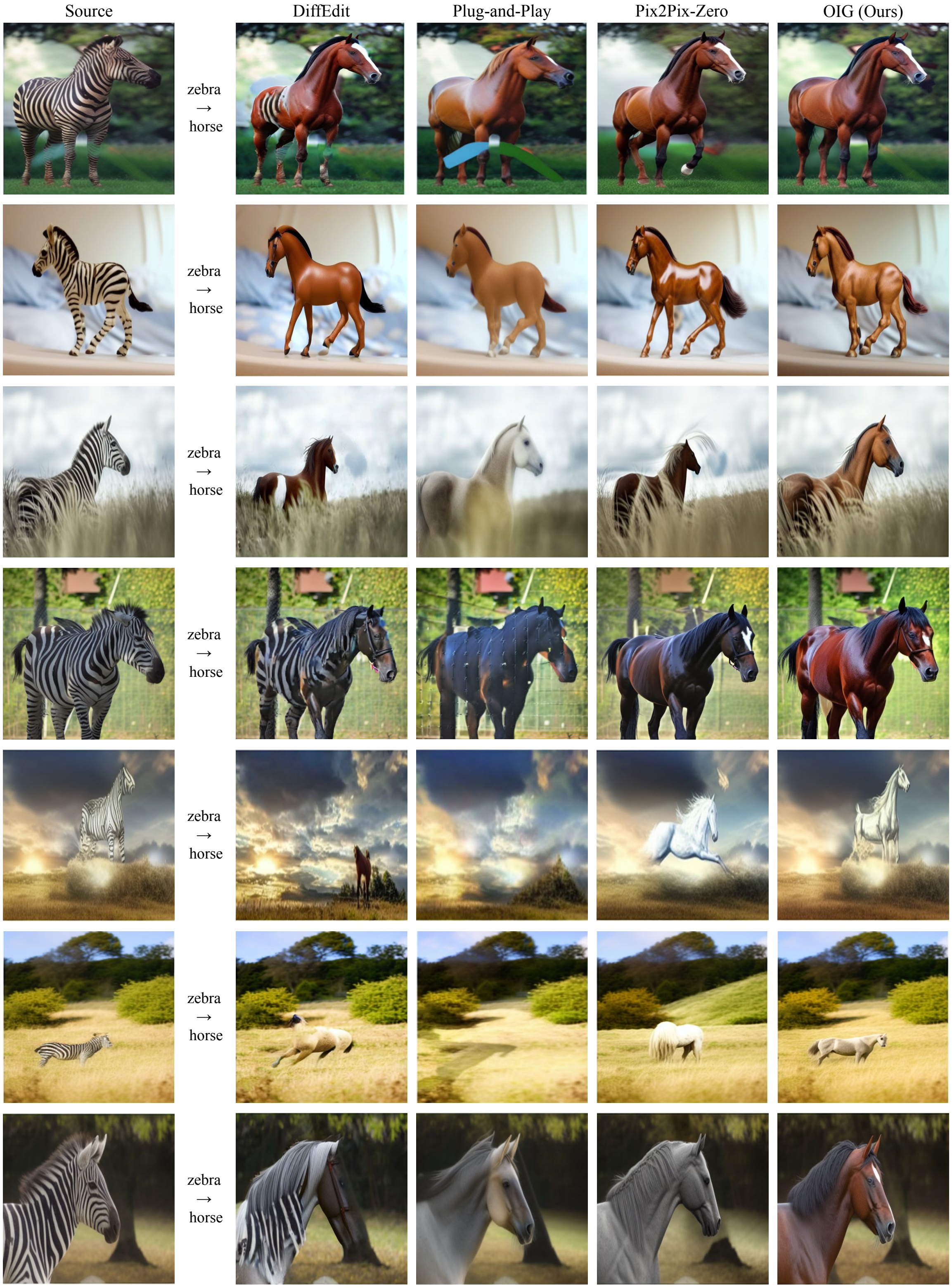


Figure 21. Additional qualitative results of the proposed method, DiffEdit [3], Plug-and-Play [12], and Pix2Pix-Zero [7] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the zebra → horse task.

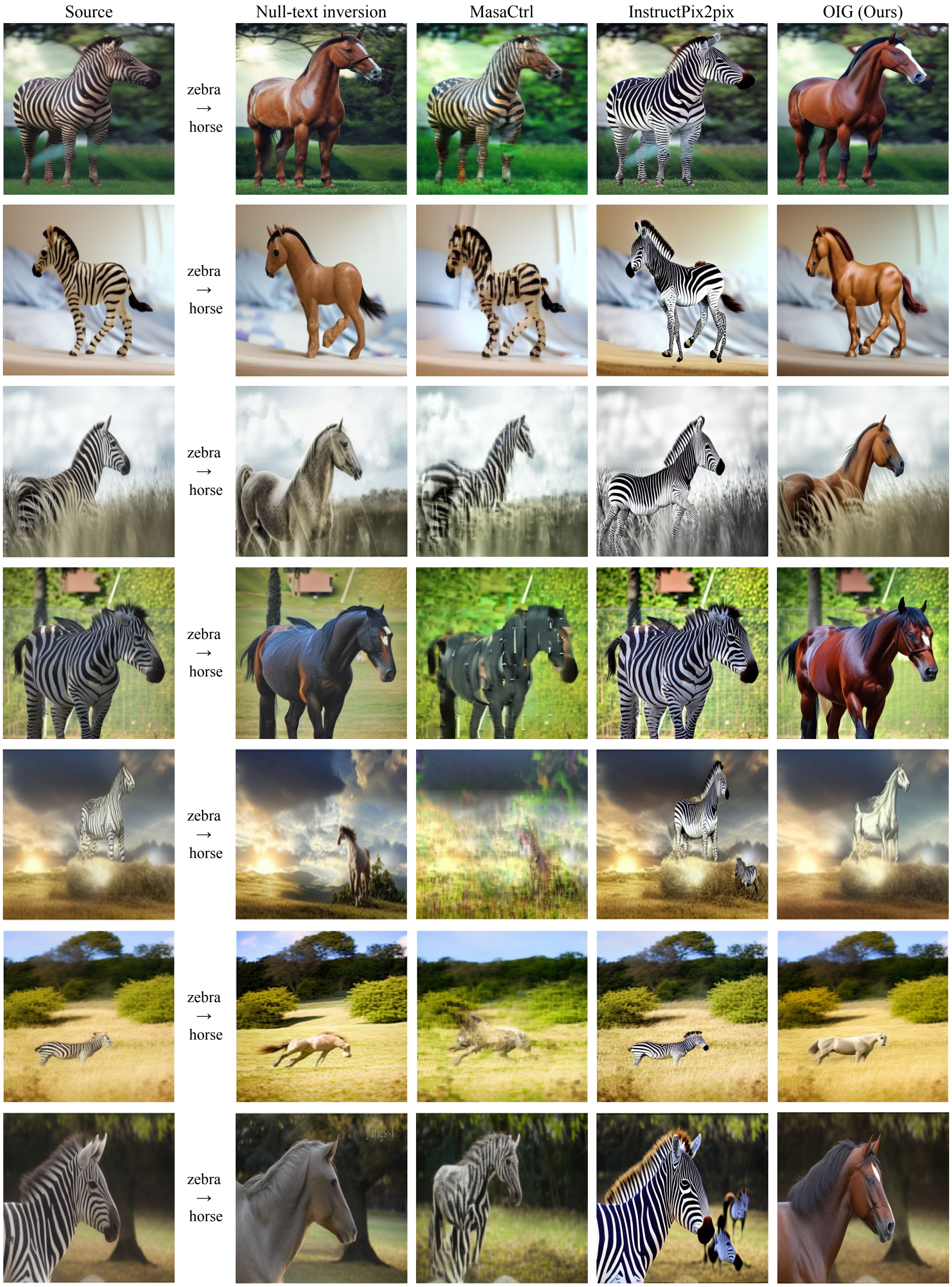


Figure 22. Additional qualitative results of the proposed method, Null-text inversion [6], MasaCtrl [2], and InstructPix2Pix [1] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the zebra → horse task.

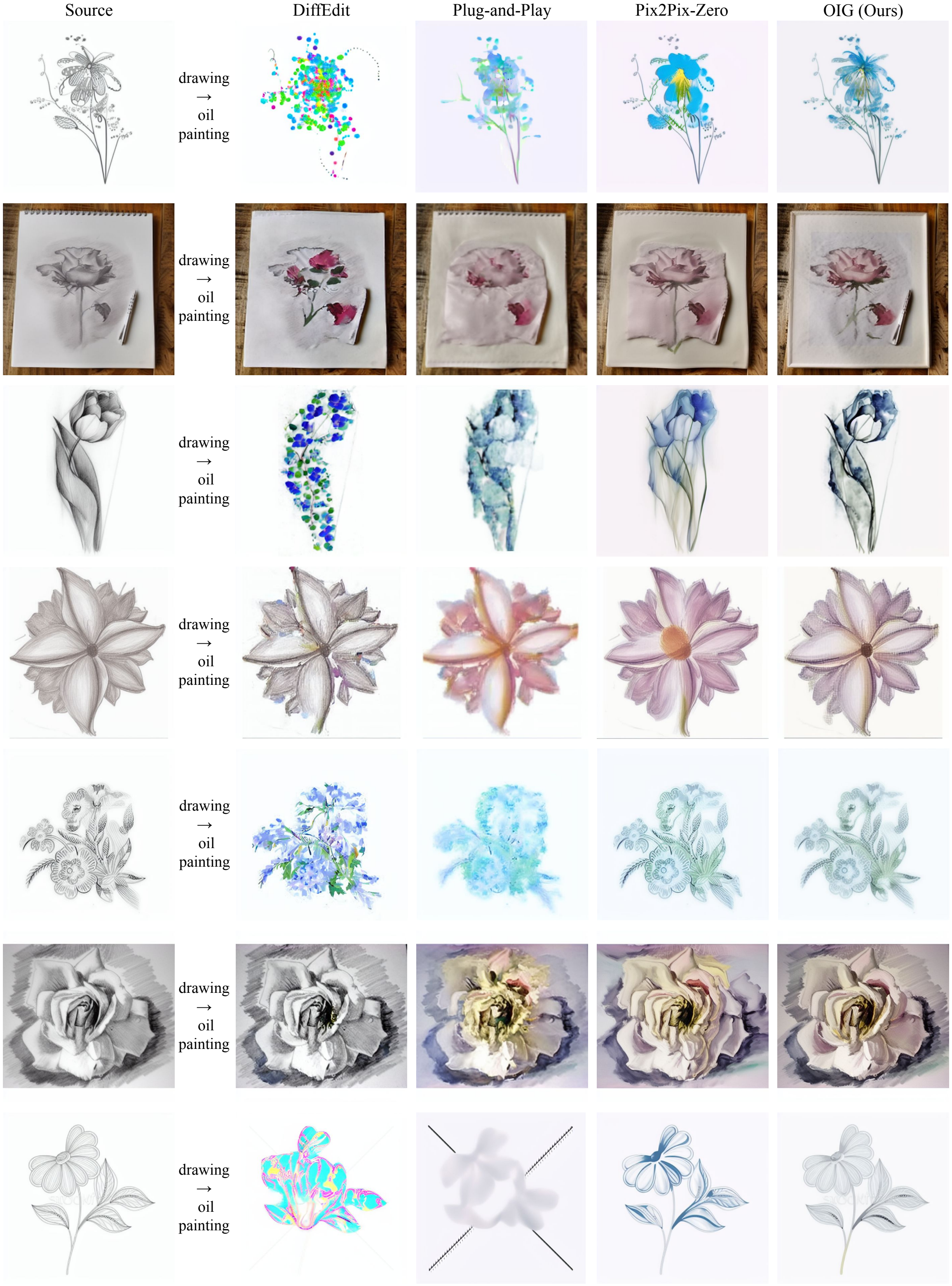


Figure 23. Additional qualitative results of the proposed method, DiffEdit [3], Plug-and-Play [12], and Pix2Pix-Zero [7] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the drawing → oil painting task.

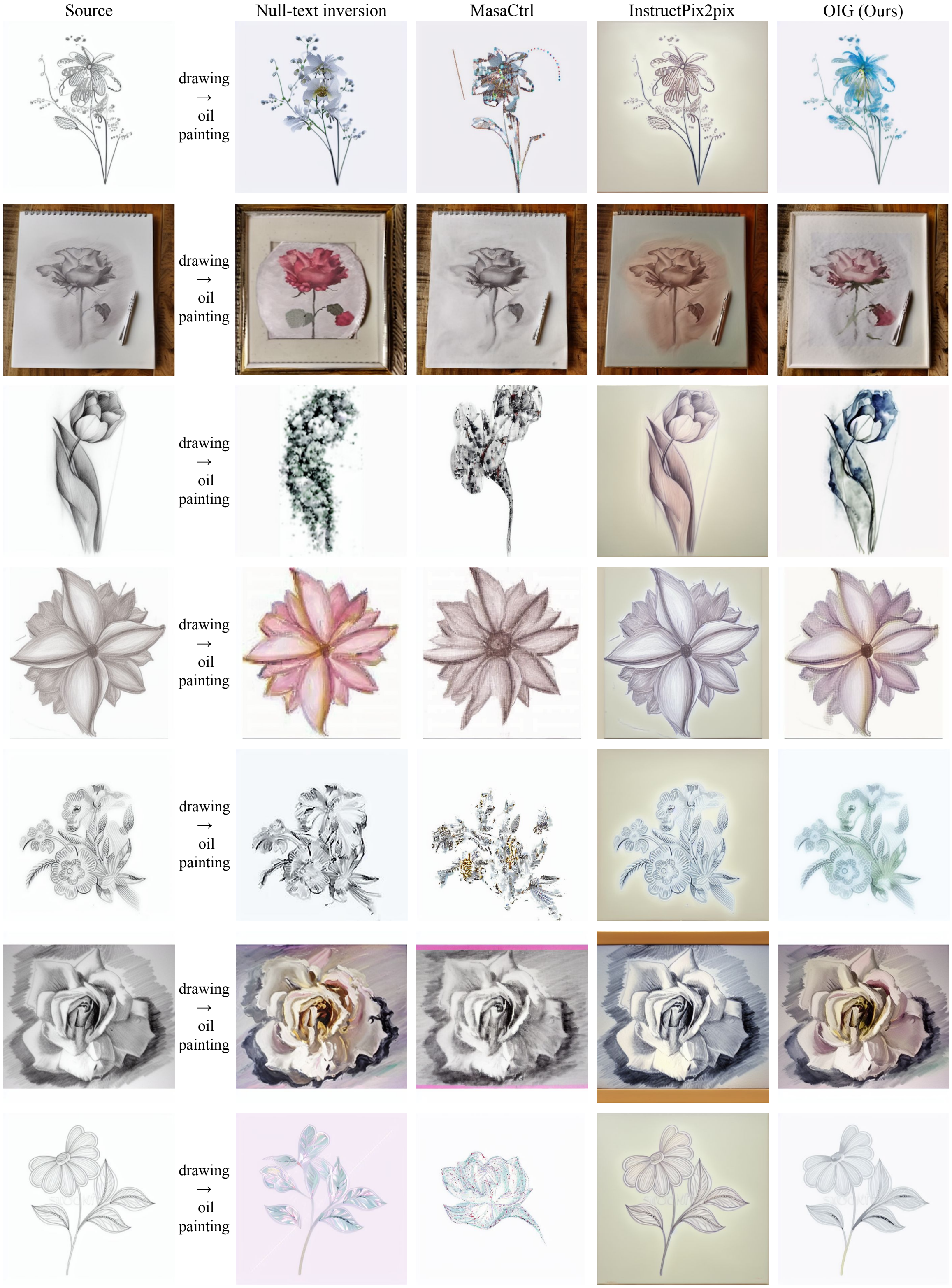


Figure 24. Additional qualitative results of the proposed method, Null-text inversion [6], MasaCtrl [2], and InstructPix2Pix [1] using the pretrained Stable Diffusion [8] and real images sampled from the LAION 5B dataset [9] on the drawing → oil painting task.

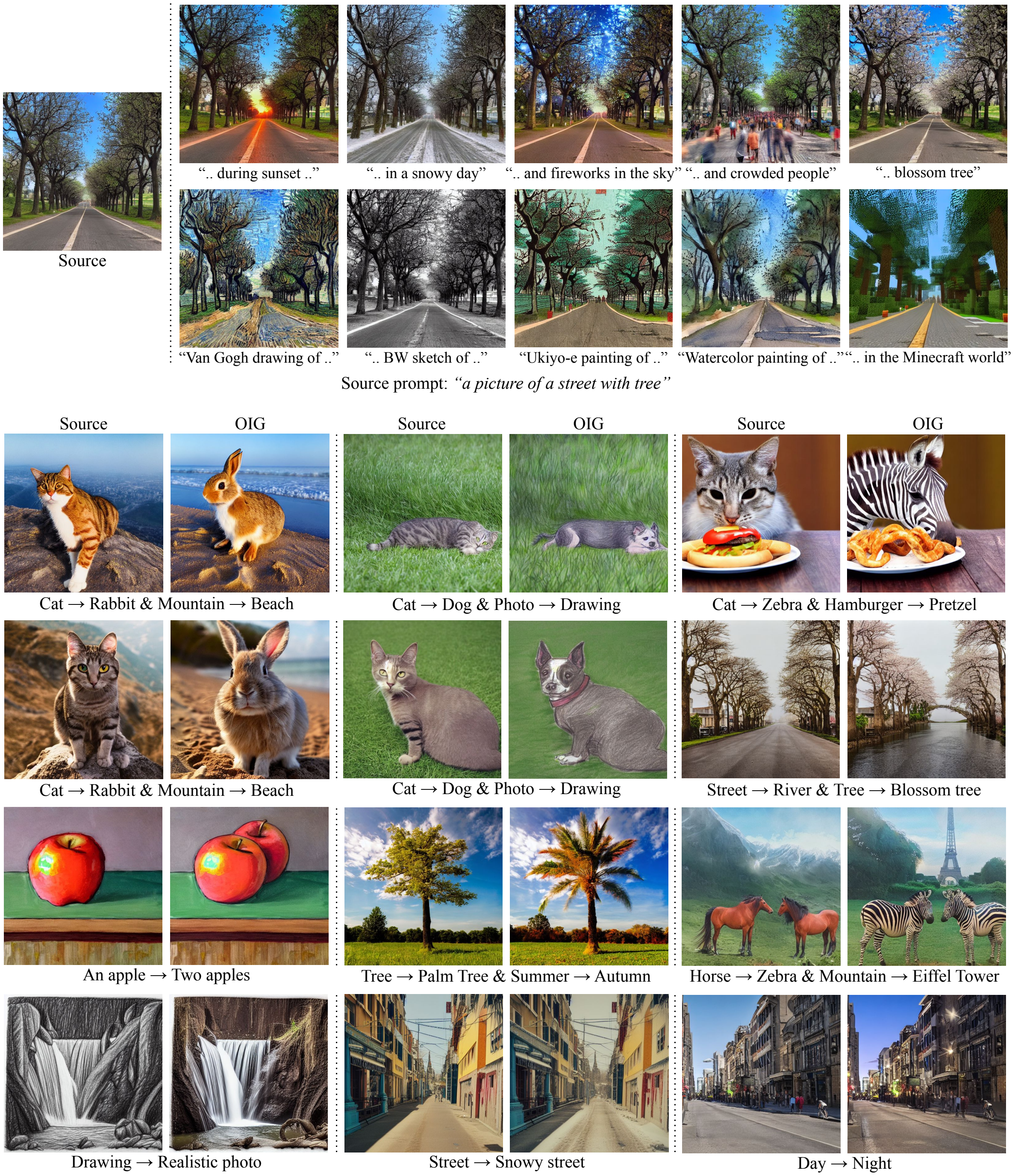

Source
".. during sunset .."
".. in a snowy day"
".. and fireworks in the sky"
".. and crowded people"
".. blossom tree"
"Van Gogh drawing of .."
".. BW sketch of .."
"Ukiyo-e painting of .."
"Watercolor painting of .."
".. in the Minecraft world"
Source prompt: "a picture of a street with tree"
Source
OIG
Source
OIG
Source
OIG
Cat → Rabbit & Mountain → Beach
Cat → Dog & Photo → Drawing
Cat → Zebra & Hamburger → Pretzel
Cat → Rabbit & Mountain → Beach
Cat → Dog & Photo → Drawing
Street → River & Tree → Blossom tree
An apple → Two apples
Tree → Palm Tree & Summer → Autumn
Horse → Zebra & Mountain → Eiffel Tower
Drawing → Realistic photo
Street → Snowy street
Day → Night

Figure 25. Qualitative results of the proposed method on synthetic images given by the pretrained Stable Diffusion [8].

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to Follow Image Editing Instructions. In *CVPR*, 2023. 4, 9, 11, 13, 15, 17, 19

[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*, pages 22560–22570, 2023. 1, 2, 4, 9, 11, 13, 15, 17, 19

[3] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance. In *ICLR*, 2023. 1, 2, 4, 8, 10, 12, 14, 16, 18

[4] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control. In *ICLR*, 2023. 1

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018. 7

[6] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text Inversion for Editing Real Images Using Guided Diffusion Models. In *CVPR*, 2023. 1, 2, 4, 9, 11, 13, 15, 17, 19

[7] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-Shot Image-to-Image Translation. In *SIGGRAPH*, 2023. 1, 2, 4, 8, 10, 12, 14, 16, 18

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

[9] 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. In *NeurIPS Datasets and Benchmarks Track*, 2022. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

[10] Charles M Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The annals of Statistics*, 1981. 3

[11] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual Diffusion Implicit Bridges for Image-to-Image Translation. In *ICLR*, 2022. 2

[12] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play Diffusion Features for Text-Driven Image-to-Image Translation. In *CVPR*, 2023. 1, 2, 4, 8, 10, 12, 14, 16, 18

[13] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *CVPR*, 2017. 2