Effective Real Image Editing with Accelerated Iterative Diffusion Inversion

Supplementary Material

Zhihong Pan Riccardo Gherardi Xiufeng Xie Stephen Huang
Oppo Mobile Telecommunications Corp.
2479 E Bayshore Rd, Palo Alto, CA, USA

1. DDIM ODE Solver

As explained in the main paper, our proposed AIDI is based on the fixed-point iteration for an implicit function $z_t = f(z_t)$ when the problem is formulated as a sampling step from an unknown z_t to a given z_{t-1}. It is also shown here that it is equivalent to the backward Euler method for an ODE when formulated as an inversion step. For the following inversion step from z_t to z_{t+1}

$$\begin{aligned} z_0^t &= (z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t)/\sqrt{\bar{\alpha}_t} \\ z_{t+1} &= \sqrt{\bar{\alpha}_{t+1}} z_0^t + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_t, \end{aligned} \tag{1}$$

it is equivalent to the following equation

$$\frac{z_{t+1}}{\sqrt{\bar{\alpha}_{t+1}}} = \frac{z_t}{\sqrt{\bar{\alpha}_t}} + \left[\sqrt{\frac{1-\bar{\alpha}_{t+1}}{\bar{\alpha}_{t+1}}} - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\right]\epsilon_t. \tag{2}$$

Using the following reparameterization,

$$\bar{z}(t) = \frac{z_t}{\sqrt{\bar{\alpha}_t}},\ \xi(t) = \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}, \tag{3}$$

the inversion step in Equation 2 is the forward Euler method for the following ODE

$$\mathrm{d}\bar{z}(t) = \epsilon(t, \bar{z})\mathrm{d}\xi(t). \tag{4}$$

Note that here t is used as a variable instead of an index t, and the previously used index $t+1$ corresponds to the variable $t + \triangle t$. As alternative to Equation 2 where $\epsilon(t, \bar{z})$ is calculated from the given $\bar{z}(t)$, the above ODE problem can be solved using the backward Euler method where the implicit $\bar{z}(t + \triangle t)$ is used instead deduct $\epsilon(t + \triangle t, \bar{z})$. Reverting to using index $t+1$ instead of variable $t + \triangle t$, the backward Euler is described as

$$z_{t+1} = \sqrt{\frac{\bar{\alpha}_{t+1}}{\bar{\alpha}_t}} z_t + \left[\sqrt{1-\bar{\alpha}_{t+1}} - \sqrt{\frac{(1-\bar{\alpha}_t)\bar{\alpha}_{t+1}}{\bar{\alpha}_t}}\right]\epsilon_{t+1}. \tag{5}$$

This is the same result derived in the main paper's eq. 7

$$z_t = \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}} z_{t-1} + \left[\sqrt{1-\bar{\alpha}_t} - \sqrt{\frac{(1-\bar{\alpha}_{t-1})\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}\right]\epsilon_t. \tag{6}$$

after replacing t and $t+1$ with $t-1$ and t respectively. Given the ODE shown in Equation 4, the simple inversion used in PTP is based on the forward Euler method while our AIDI is based on the backward Euler method. Other applicable ODE solvers would be of future research interests.

2. General Editing Test

We include here results from the recent pix2pix-zero (P2P0)[5] where applicable, in addition to those already showcased in the main text (PTP [3], NTI [4], EDICT [6] and HS-SCLIP [1, 2]).

We show in the main paper that our proposed real image editing process is able to complete a wide range of image editing tasks using only 20 editing steps, such as fully or partially changing an object, replacing backgrounds, or image-to-image translation. There is not yet however a common set of images and editing tasks established for quantitative analysis. To complement the Dog2Cat test using high-resolution AFHQ images, we adopt here the quantitative experimental settings from EDICT [6] to assess the relative performance of different methods on the same three general editing tasks.

Using the test images from 5 animal classes of ImageNet, the object swapping test includes swapping one class of animal to each of the other four classes. The background test consists of changing it to *in a parking lot* or *in the snow*, and the style transfer test aims to translate the input image as *an impressionistic painting*. To evaluate the edited results quantitatively, without a target image set available for FID calculation, we rely on CLIP and LPIPS scores to assess the trade-off between editing quality and perceptual similarity in reference to the input image. Following the practices in EDICT, the CLIP score is not calculated directly as the similarity between the edited image and the target prompt. In-

stead a number of similar prompts, the target one included, are used to calculate the softmax value of the target prompt. In this case the CLIP score used here has a range of 0 to 1 and a larger score is better.

The IDs of the 5 chosen animal classes are 386, 348, 285, 294 and 185 respectively, each associated with the following base prompts: *an elephant*, *a male sheep*, *a cat*, *a brown bear*, and *a dog*. Note that here we use *a male sheep* instead of *a ram* to avoid confusion with the vehicle. For the object

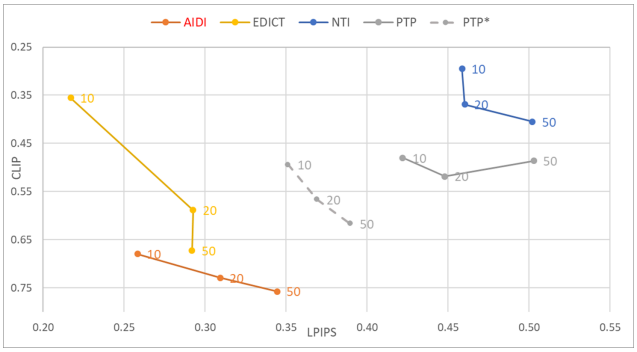

Figure 1: Quantitative assessments for general editing using LPIPS and CLIP. The data labels represent editing steps.

swapping test, the CLIP score is calculated from the same 5 base prompts. For the background replacement test, the following two sets of 5 prompts are used, where the target ones are in bold.

- ***An [animal] in a parking lot***
- *An [animal] in the wild*
- *An [animal] in the rain*
- *An [animal] in the ocean*
- *An [animal] in a sand storm*
- ***An [animal] in the snow***
- *An [animal] in the sun*
- *An [animal] in a shopping mall*
- *An [animal] in the ocean*
- *An [animal] on a football field*

For the style transfer test, the following 5 are used:

- ***An impressionistic painting of an [animal]***
- *A photograph of a [animal]*
- *A digital rendering of an [animal]*
- *A crayon drawing of an [animal]*
- *A pencil drawing of an [animal]*

As shown in Fig. 1, our method is the best overall for low LPIPS value and high CLIP score. For fast editing using 10 to 20 steps, our method is significantly better than all other methods. It is interesting to see that for EDICT, comparing 50 steps to 20, it can improve CLIP score with a good margin with minimum trade-off in LPIPS. Note that the average LPIPS value and CLIP score are calculated as the average for each of the three tests first, then averaged again over all three tests. And here the HS-SCLIP results are not available as there is no StyleGAN model trained on applicable dataset. An ablation study is also conducted for

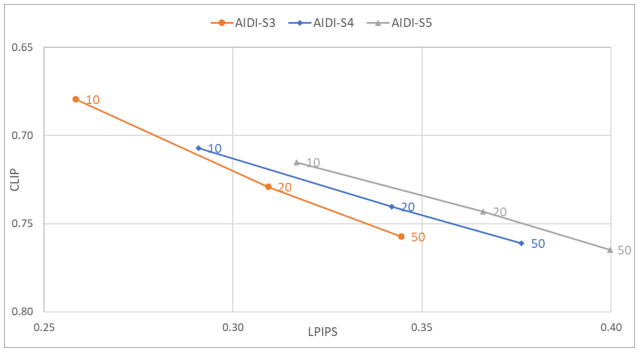

Figure 2: Comparison between different guidance scale settings in AIDI using LPIPS and CLIP for the general editing test. The data labels represent editing steps.

this test by varying the guidance scales during, while keeping the scale as 1 for inversion as designed. As shown in Fig. 2, the results for 10/20/50 editing steps are included for three different scales: 3, 4 and 5. When the number of editing step is fixed, increasing scales can only bring marginal improvement in CLIP score but result in more significant trade-off in LPIPS. For the main results shown above, same as the Dog2Cat test in the main paper, a scale of 3 is used. For the two best methods, our AIDI and EDICT, the results

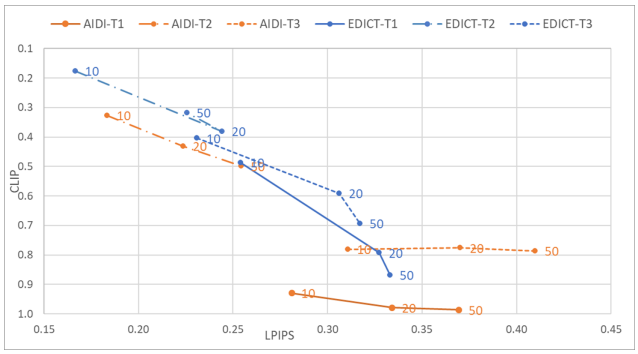

Figure 3: Comparison between different type of tests for AIDI and EDICT, including T1: object swapping; T2: background replacement; T3: style transfer. The data labels represent editing steps.

from three type of tests are illustrated separately in Fig. 3. The object swapping test, similar to the Dog2Cat test in the

main paper, has the best results comparing to the other two. For background replacement, the LPIPS values are much smaller overall, mainly because the front object often occupies the majority area of the images in this test set. A separate test set with larger background area could help evaluating the result more effectively. It is also interesting to see that for our AIDI, more editing steps doesn't result in higher CLIP values for the style transfer test.

3. Additional Dog2Cat Tests

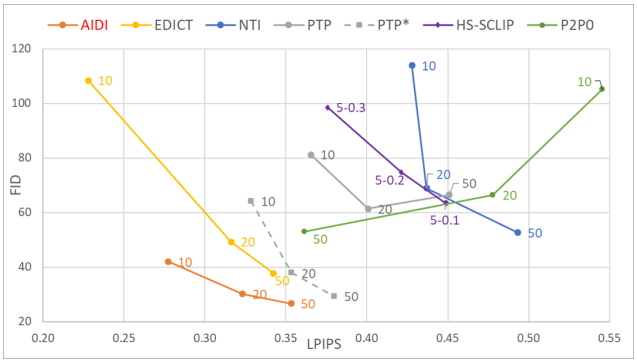

Figure 4: Quantitative assessments for Dog2Cat test using LPIPS and FID. The data labels represent editing steps for all methods except for HS-SCLIP where they are hyperparameters.

For the Dog2Cat test, we have expanded it to include results from the most recent P2P0. As shown in Fig. 4, for 50 steps, P2P0 is better than PTP for as claim in their own study, but not as good as EDICT and our AIDI. For 10-20 steps, it is worse in both LPIPS and FID, different from other methods which have degraded FID but improved LPIPS. We have also included the results from CLIP score

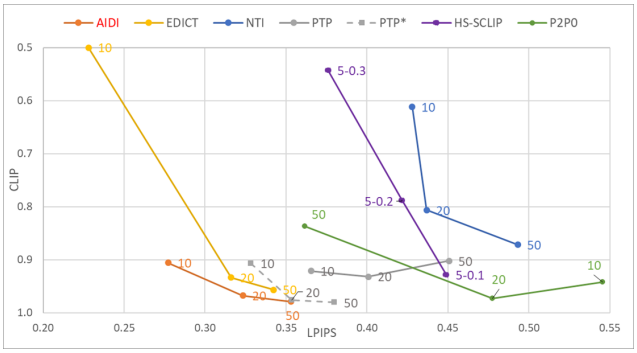

Figure 5: Quantitative assessments for Dog2Cat test using LPIPS and CLIP. The data labels represent editing steps for all methods except for HS-SCLIP where they are hyperparameters.

Original AIDI P2P0

Figure 6: Visual comparisons between AIDI and P2P0 results with similar CLIP score.

in Fig. 5 where the following 5 prompts are used.

- ***High quality photo of a cat***
- *High quality photo of a dog*
- *High quality photo of a tiger*
- *High quality photo of a lion*
- *High quality photo of a leopard*

While our AIDI is still the best overall considering both lower LPIPS and higher CLIP, its advantage in CLIP is not as significant as those in FID. From a qualitative standpoint, as shown in Fig. 6, our edited results are more consistent with the source image as compared to those from P2P0, although they have similar CLIP score using 20 editing steps. While the P2P0 results match well with the *High quality photo of a cat* prompt, the out-of-focus blurring for the neck area and the background are not consistent with the source image. FID score is a more appropriate metric to evaluate editing quality when a large set of images from the target domain is available.

4. More Visual Examples

More examples for the Dog2Cat test are shown in Fig 7. While our result is better overall, especially for 10-20 steps, we have included one failure case of our method as the third example. While ours and some other methods fail to edit one dog as one cat, the one from NTI at 50 steps has the best result.

Visual examples from the general editing test are shown Fig. 8-10. We have also included results from applying different editing tasks to the same input image and the results are shown in Fig. 11. It is shown that when the simple caption *a dog* is used as the input prompt for an image with other front objects other than the dog, it leads to incorrect editing results for different editing tasks.

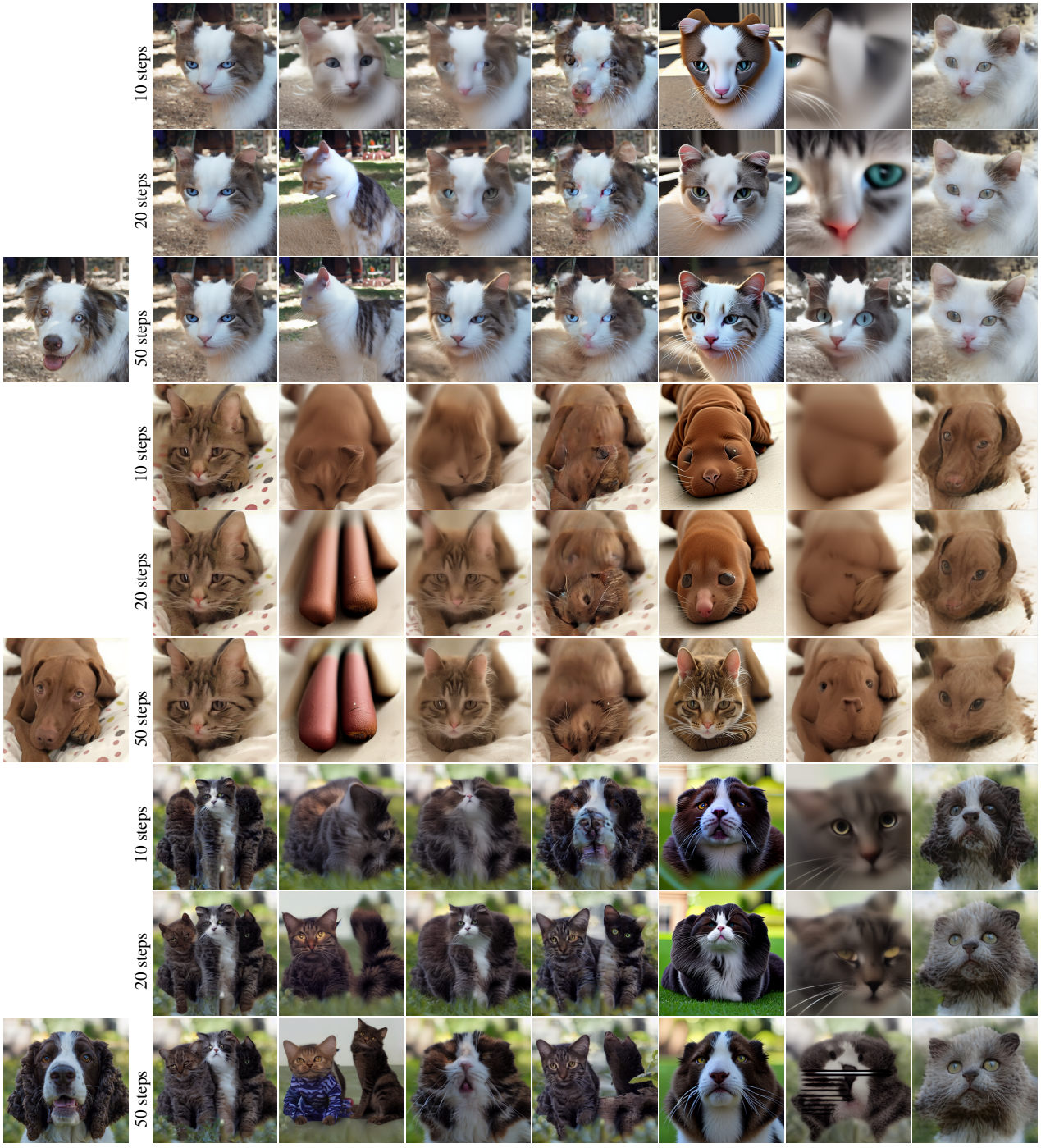

Figure 7: Visual examples for dog-to-cat editing test using AFHQ test set. Results from of different model are organized horizontally and results from different settings, 10/20/50 editing steps for all diffusion-based models or 3 hyperparameter settings for HS-SCLIP, are organized vertically.

References

[1] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. HyperStyle: StyleGAN inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, pages 18511–18521, 2022. 1

[2] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia

Yang, et al. CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1

[3] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[4] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1

[5] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 1

[6] Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: Exact diffusion inversion via coupled transformations. *arXiv preprint arXiv:2211.12446*, 2022. 1

Figure 8: Visual examples for the object swapping test using ImageNet test images.

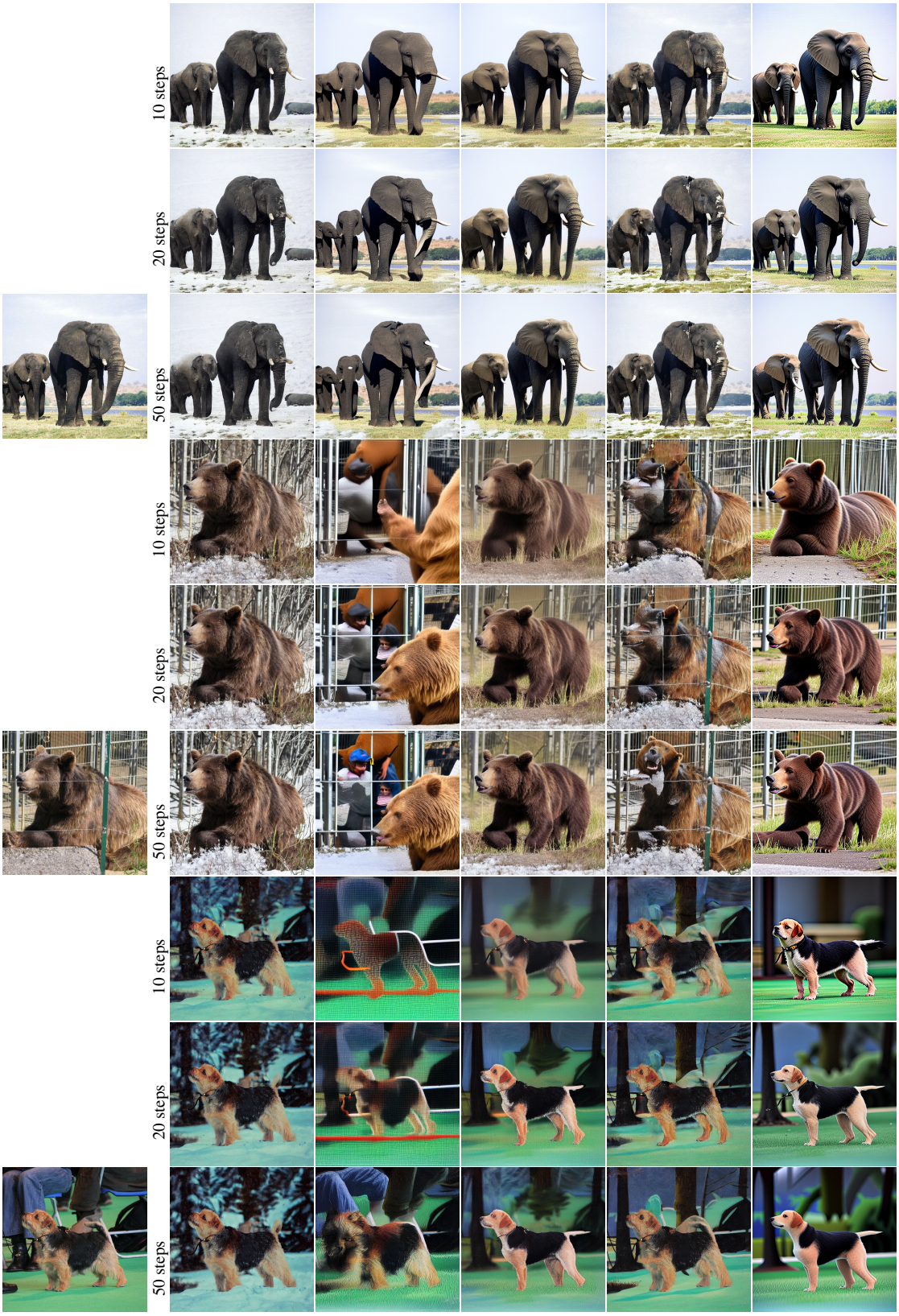

Figure 9: Visual examples for the background replacement test (*in the snow*) using ImageNet test images.

Figure 10: Visual examples for the style transfer test (*an impressionistic painting of*) using ImageNet test images.

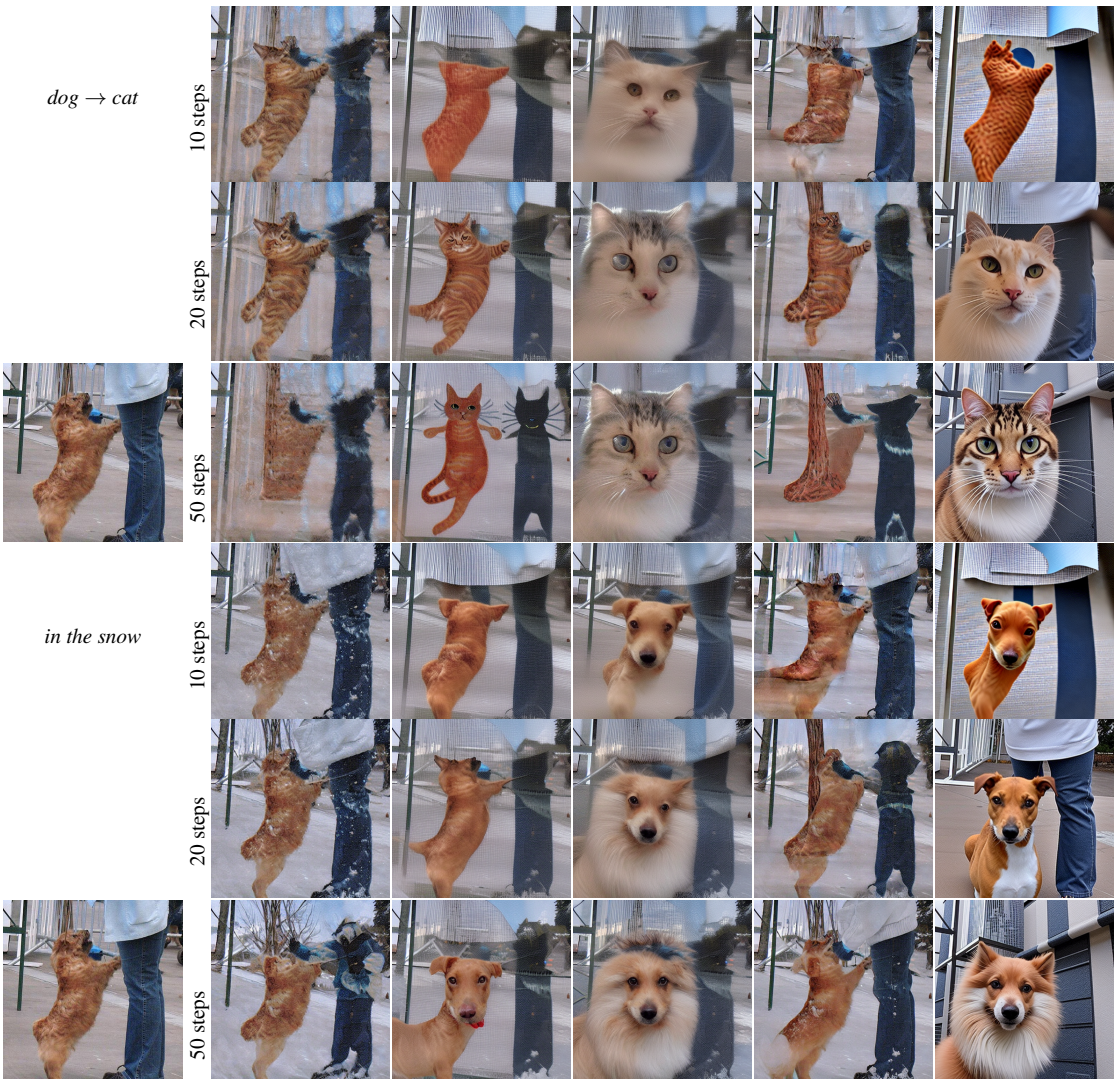

Figure 11: Visual examples for various editing tests using one input image.