Supplementary Material for EDICT: Exact Diffusion Inversion via Coupled Transformations

Bram Wallace
Salesforce Research
b.wallace@salesforce.com

Akash Gokul
Salesforce Research
agokul@salesforce.com

Nikhil Naik
Salesforce Research
nnaik@salesforce.com

The supplementary document is organized as follows:
Sec. 1: Details of the quantitative evaluation of image editing methods
Sec. 2: Human evaluation paralleling above quantitative evaluation
Sec. 3: Additional analysis of DDIM instability.
Sec. 4: Additional image editing results with EDICT
Sec. 5: Additional image reconstruction results with EDICT

1. Quantitative Experiment Details

We sample from 5 ImageNet classes (*African Elephant, Ram, Egyptian Cat, Brown Bear, and Norfolk Terrier*, validation set). Four experiments are performed, one swapping the pictured animal's species to each of the other classes (20 species editing pairs in total), two contextual changes (*A [animal] in the snow* and *A [animal] in a parking lot*), and one stylistic (*An impressionistic painting of a [animal]*). The prompt for the species edit is simply *A [animal]*. Throughout, base prompts are of form *A [animal]*. Edits are performed with inversion strength $s = 0.8$ and steps $S = 50$.

For computing CLIP score in the species example, the 5 text queries are of identical form *A [animal]*. CLIP text queries for other edits are as follows:

A [animal] in the snow

- An animal in the snow
- An animal in the sun
- An animal in the rain
- An animal in a sand storm
- An animal in the ocean

A [animal] in a parking lot

- An animal in a parking lot
- An animal in the wild
- An animal in a shopping mall
- An animal in the ocean
- An animal on a football field

A impressionistic painting of a [animal]

- An impressionistic painting of an animal
- A photograph of an animal
- A crayon drawing of an animal
- A digital rendering of an animal
- A pencil drawing of an animal

We plot the mean and median metrics for baselines on each individual benchmark experiment as well as the mean-average and median-average across experiments in Figure S1.

2. Human Evaluation

Using image edits from Fig. 8 (300 images/method), we employ labelers to study human preferences between EDICT and DDIM-(U)C. As with CLIP score, the fraction of edits where the target caption is successfully incorporated differs only marginally across methods (Q:*"Does the provided caption match this image?"*). Among target-prompt-matching edits, DDIM-C has lower faithfulness to the source (*"Does the edited image preserve the major components of the original image?"*, LPIPS analog) with its best experiment (*Painting*) having 59% the number of faithful edits vs. other methods. Increasing granularity, we perform a comparison study of EDICT vs. baselines (*"Which of these images is more faithful to the reference image?"*, comparative LPIPS analog) (Table S1), on image-pairs where *both* the baseline and EDICT are $\geq 2/3$ labeled as faithful to the source image, and as achieving the edit caption by $\geq 1/2$. Edits from EDICT are generally preferred.

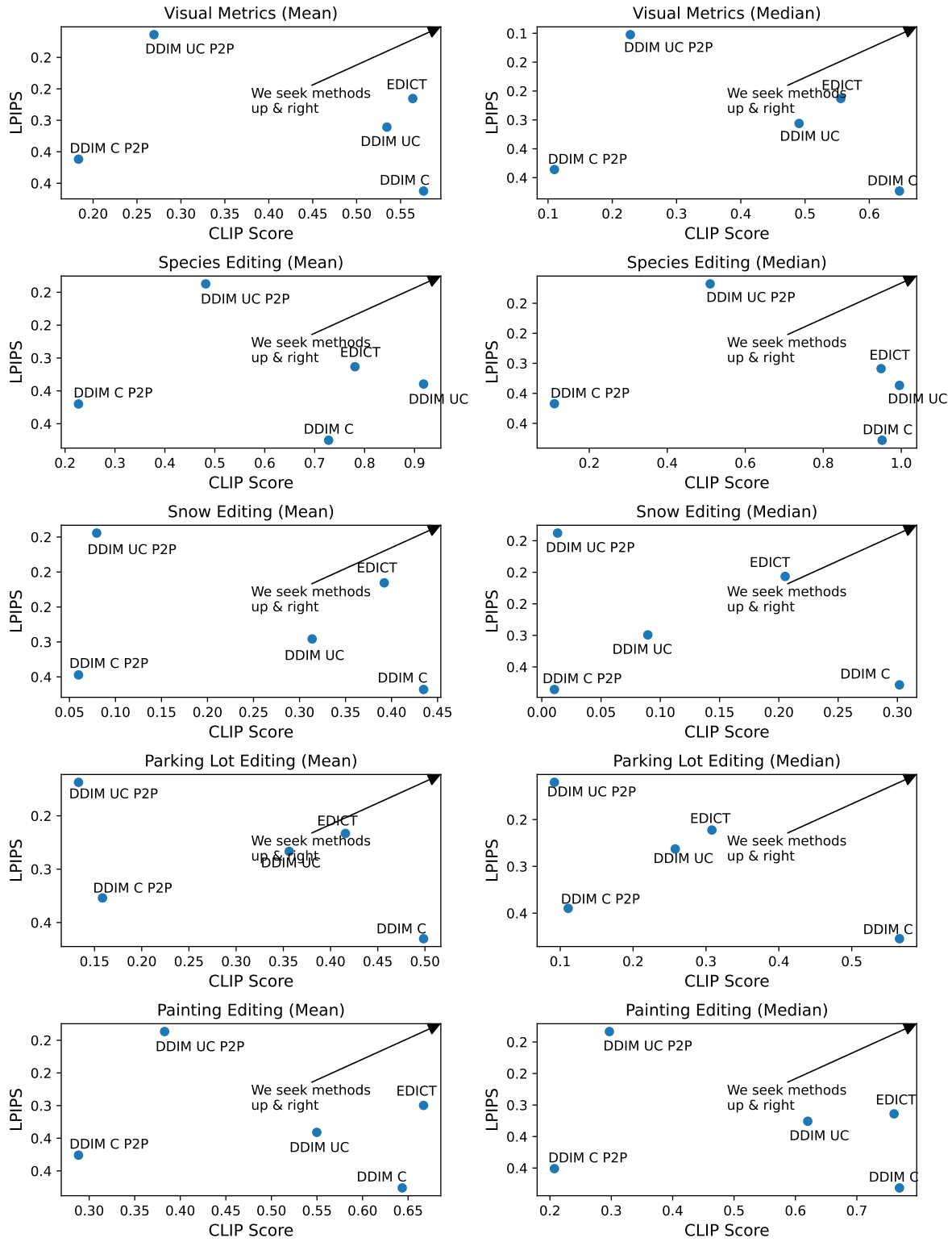

Figure S1. Median and individual plots of visual metrics (edit strength 0.8). Pareto-optimiality is achieved in all cases and for 3 of the 4 experiments (all but species editing) EDICT improves upon DDIM UC in both metrics while far outperforming DDIM UC P2P in CLIP score (achievement of edit) and DDIM C in LPIPS (perceptual faithfullness to the original image).

EDICT vs.	Animal	Snow	Parking	Painting
DDIM-C	61%	75%	100%	68%
DDIM-UC	60%	50%	57%	43%

Table S1. Head-to-head EDICT win-rates for source image faithfulness when both methods produce human-validated edits.

3. Misalignment of Pseudo-Gradient

In Section 3.2 of the main paper we claim that the pseudo-gradient of classifier-free guidance $G \cdot (\Theta(x_t, t, C) - \Theta(x_t, t, \emptyset))$ is inconsistent across time steps which drives the instability of vanilla DDIM inversion and reconstruction results. We demonstrate and analyze this instability in Figure S2 (see caption). We show that similar behavior holds for higher steps (Figure S3).

4. Edits

4.1. Additional Edit Results

In Figure S4 and Figure S5 we display further editing results.

4.2. Baselines with More Steps

In Figure S6 and Figure S7 we re-run the experiments of Figure 7 from the main paper with 100 and 250 global steps instead of the default 50. We observe minimal changes besides some instability in the final row. Note that we follow a scaling rule of $p = 0.93^{50/S}$ to maintain the same aggregate dilation/contraction factor of 0.93^{50} from the original experiments.

4.3. Extended Baselines

As noted, concurrent work such as CycleDiffusion [3], DiffuseIT [2], and DiffEdit [1] also aim to address text-prompted image editing. While the latter ([1]) does not have publicly available code to compare to, we compare to [2, 3] in Figure S8.

4.4. Dog Breeds: Extended Results

In Figure S9–Figure S15 we display additional results of dog breed editing with baselines included. EDICT x vs y are the two sequence outputs of the EDICT process to demonstrate the visually-identical convergence. We observe that EDICT consistently matches the desired output while preserving background details that baseline methods erase or alter. The base prompt is *A dog* and the target prompt is *A [target dog breed]*.

5. Reconstruction Results

In an extension of Table 1 from the main paper, we provide higher precision MSEs as well as reconstruction errors for 1000 steps in Table S2.

References

[1] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[2] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 3

[3] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. 3

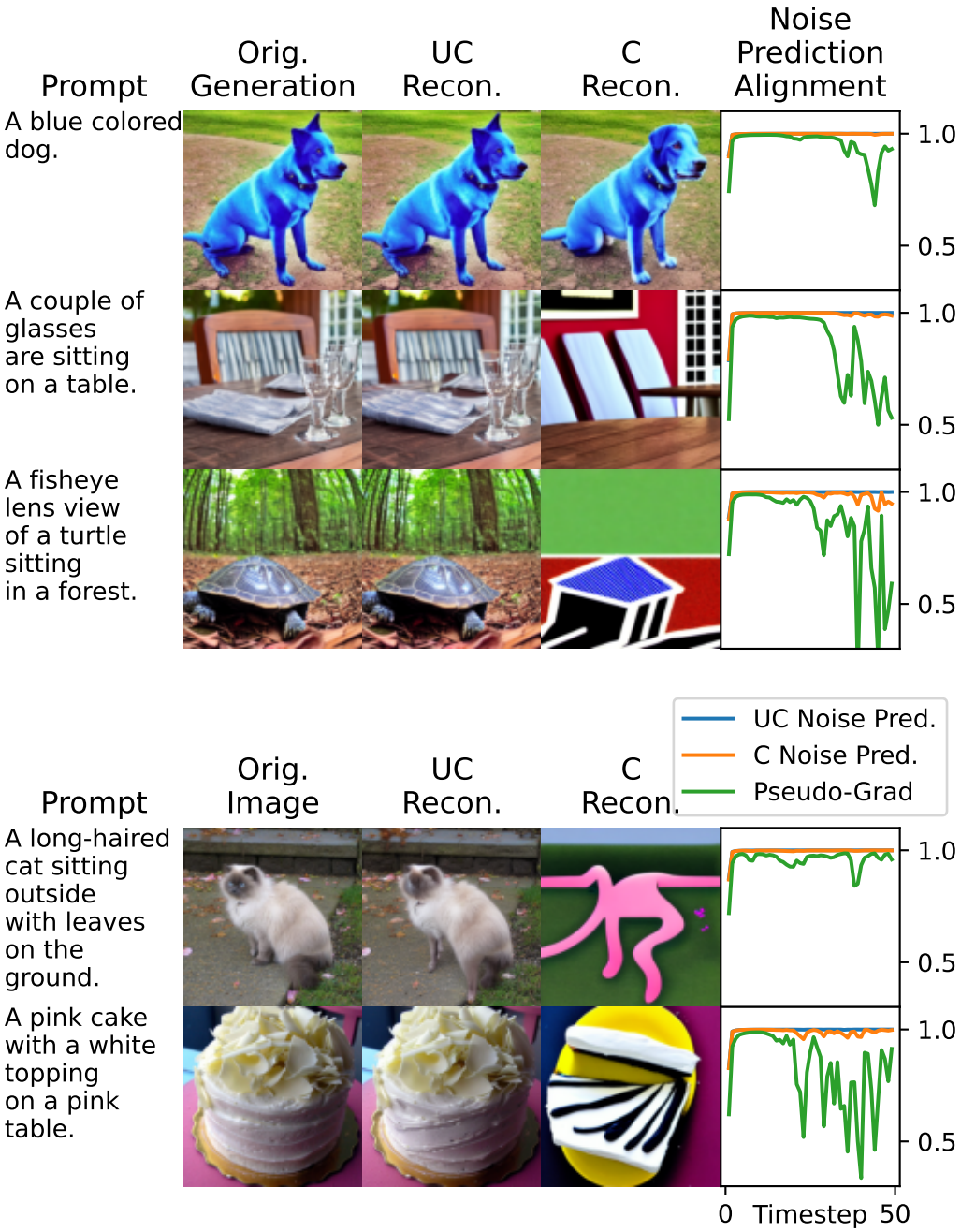

Figure S2. Top panel: Images are generated from a given prompt and displayed in *Orig. Generation*. The generation process is then inverted and re-ran both unconditionally and conditionally. While the unconditional reconstructions are near perfect, the conditional reconstructions suffer from varying degrees of instability. In the right column, we plot the cosine similarity of the noise prediction component at each timestep to the noise component at the previous timestep. Lower panel: the inversion-reconstruction is performed for real (non-generated) images. For both the unconditional and conditional component (particularly the former) we see near-perfect alignment across steps justifying the linearization assumption. For the pseudo-gradient term, however, at higher t (further-noise timesteps) the gradient becomes extremely inconsistent across timesteps. This explains the lack of stability in vanilla conditional DDIM inversion, the lack of consistency in the pseudo-gradient is counter to the linearization assumption and as such reconstructions are not faithful. All experiments run with generation-strength guidance scale of $G = 7$.

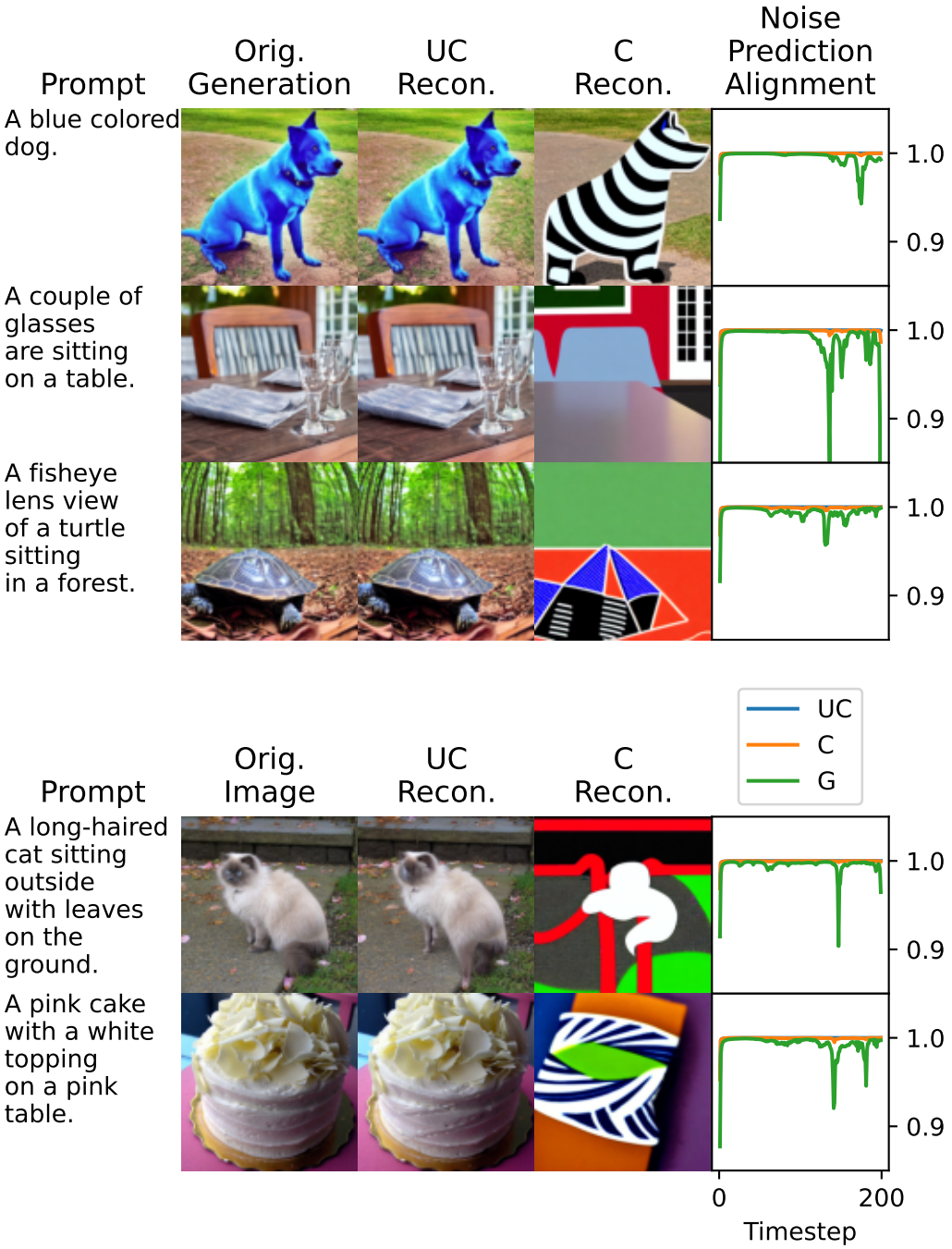

Figure S3. As Figure S2 for 200 steps. We observe that while the pseudo-gradient is more aligned between steps than previously, there are still large regions of misalignment (and given the number of steps these errors can accumulate).

Figure S4. Additional edits demonstrating EDICT's versatility. **Bold** parts of prompt are the edit.

COCO Reconstruction Error (MSE)

Method	LDM AE	EDICT (UC)	EDICT (C)	DDIM (UC)	DDIM (C)
50 Steps	0.015260	0.015260	0.015260	0.030083	0.418175
200 Steps	0.015260	0.015260	0.015260	0.023406	0.496944
1000 Steps	0.015260	0.015260	0.015260	0.018960	0.509345

Table S2. Mean-square error reconstruction results for the COCO validation set using the first listed prompt as conditioning with full-strength guidance. The latent diffusion model autoencoder (LDM AE), which is the autoencoder used to compute latents for reconstruction is the lower bound on reconstruction error.

Figure S5. A variety of types of cupcake edited with EDICT.

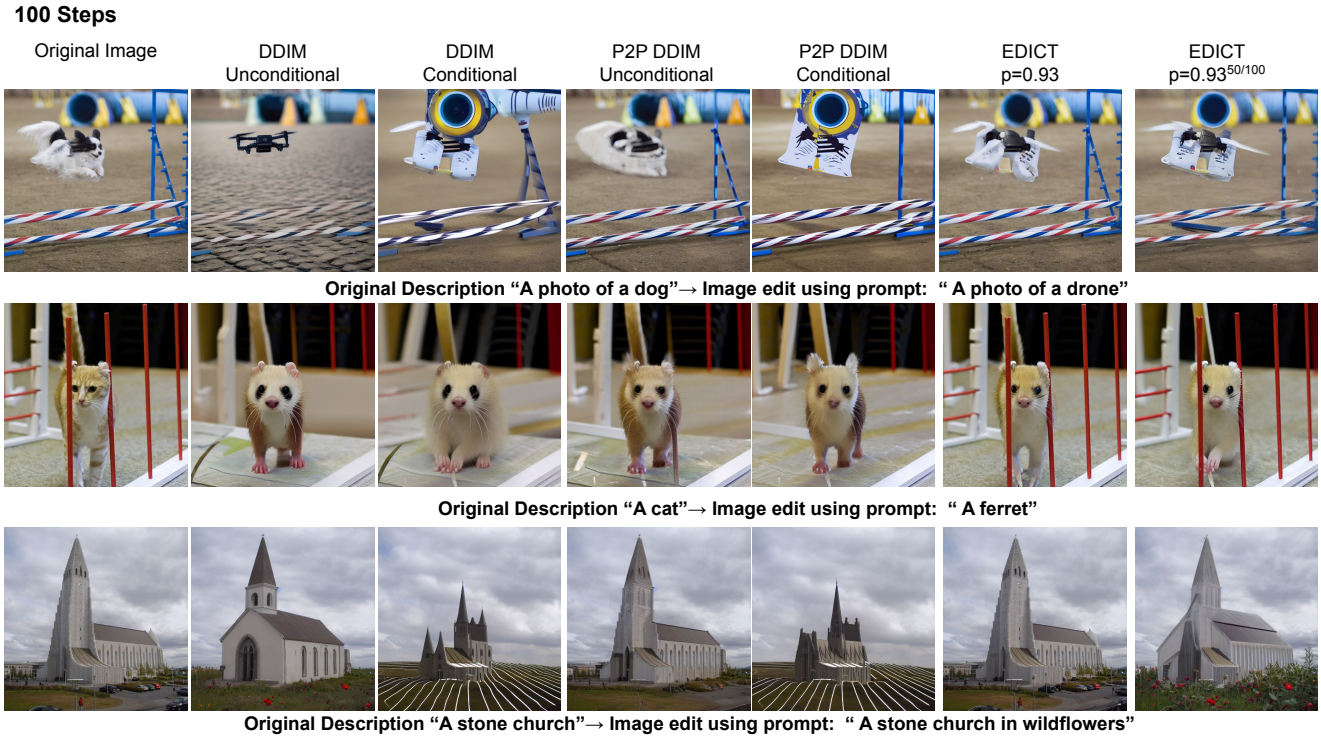

Figure S6. Baselines as in Figure 7 from the main body ran with steps $S = 100$. We note that while some detail is lost for EDICT in the bottom row it still performs well compared to the other methods and maintains far superior performance in the top two rows. The averaging/dilation factor p is chosen to have the same accumulated factor across steps and is untuned for $S = 250$.

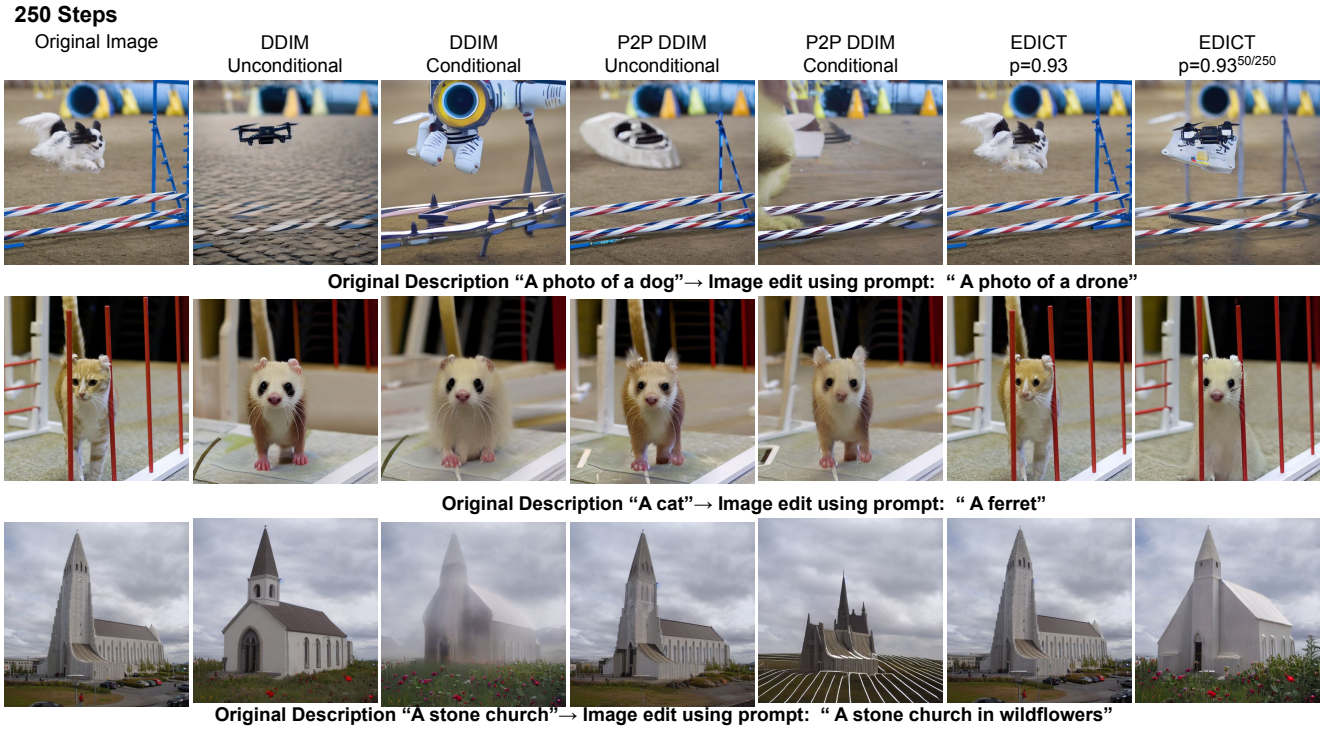

Figure S7. Baselines as in Figure 7 from the main body ran with steps $S = 250$. We note that while some detail is lost for EDICT in the bottom row it still

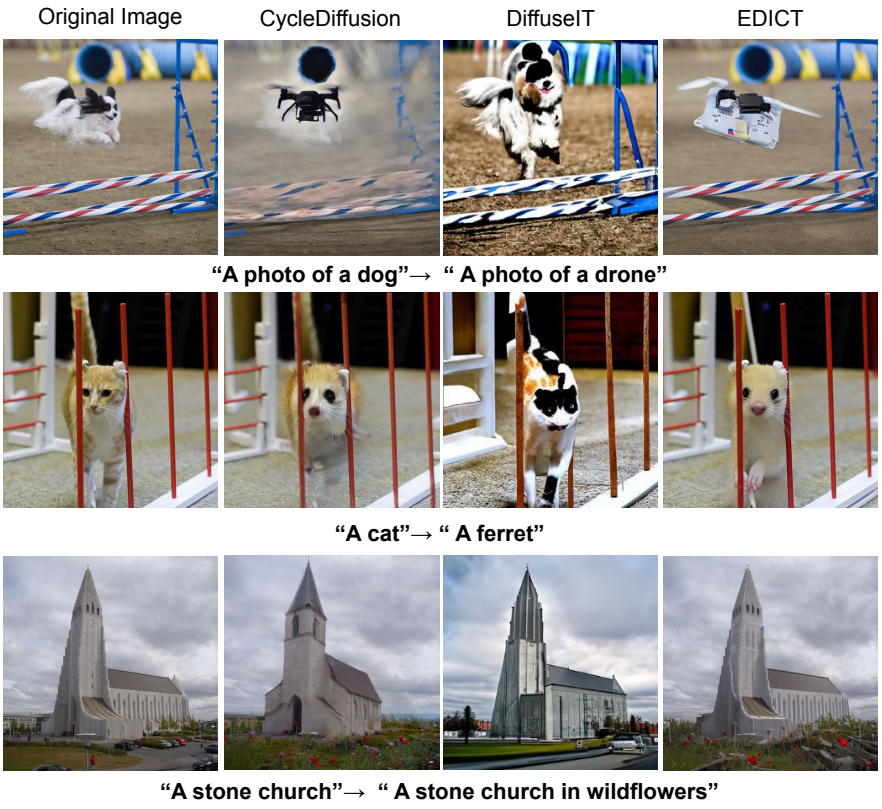

Figure S8. Comparisons to other algorithmic editing methods, CycleDiffusion and DiffuseIT, from official code implementations/suggested hyperparameters.

Figure S9. Dog breeds 1

Figure S10. Dog breeds 2

Figure S11. Dog breeds 3

Figure S12. Dog breeds 4

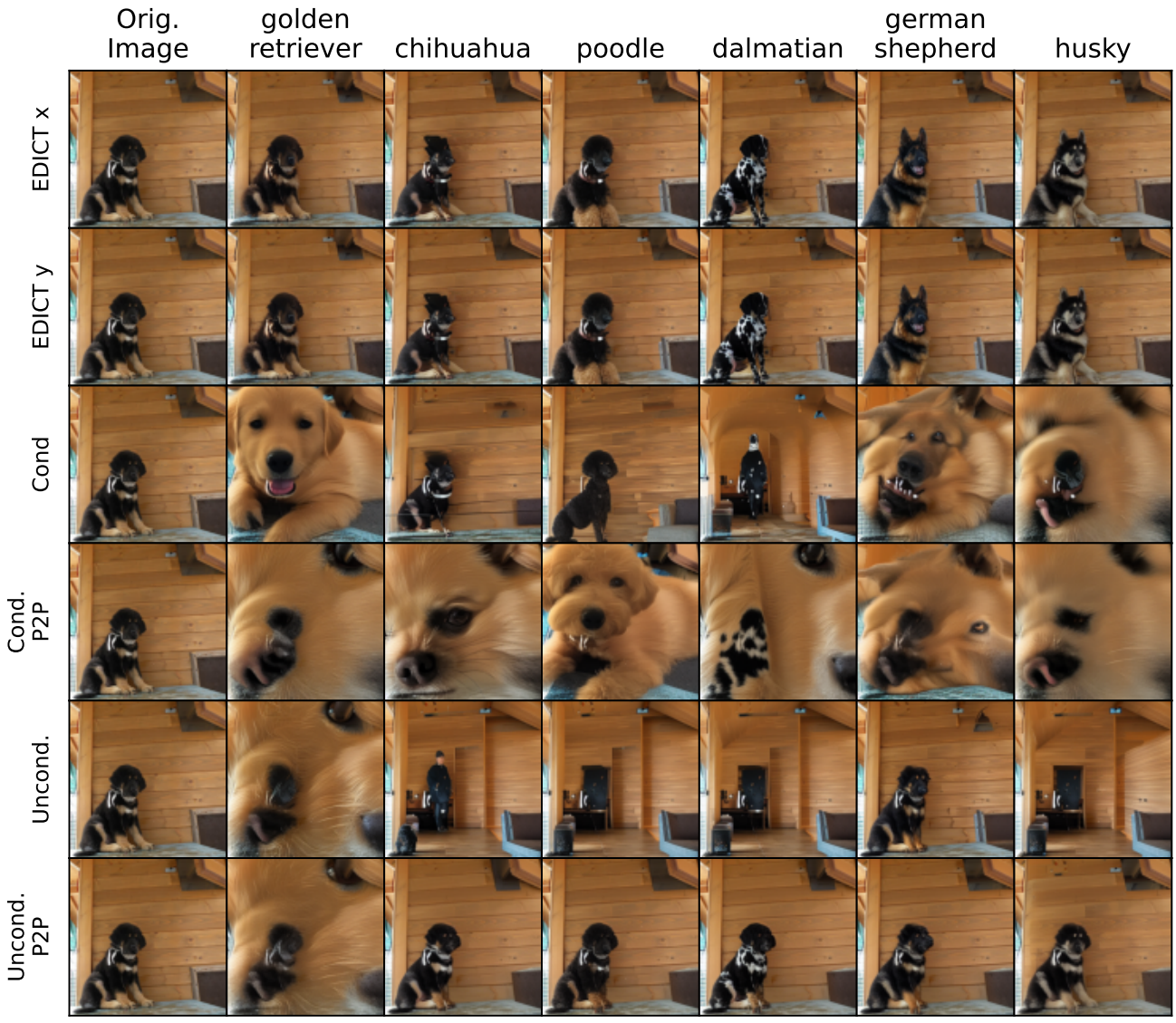

Figure S13. Dog breeds 5

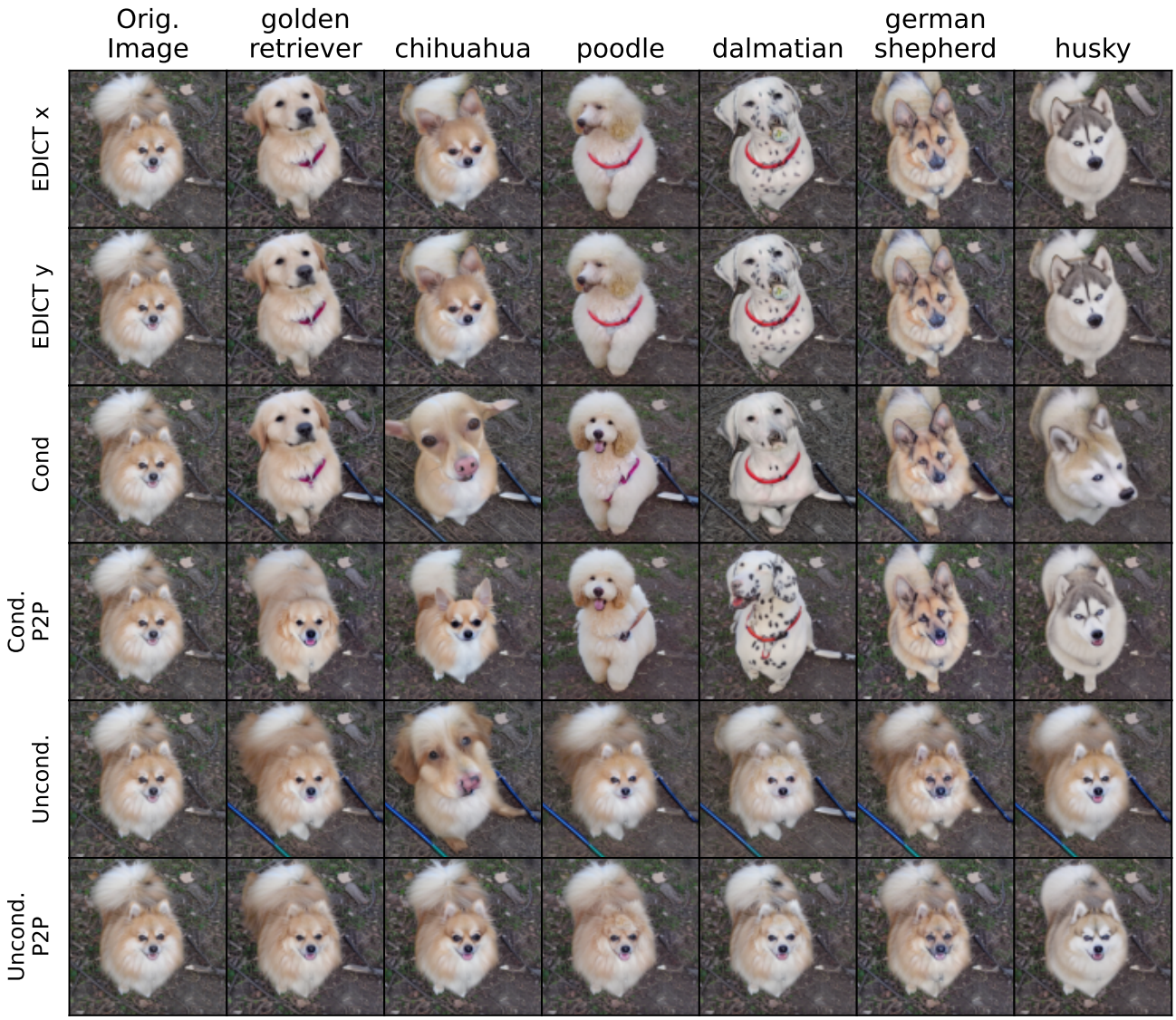

Figure S14. Dog breeds 6

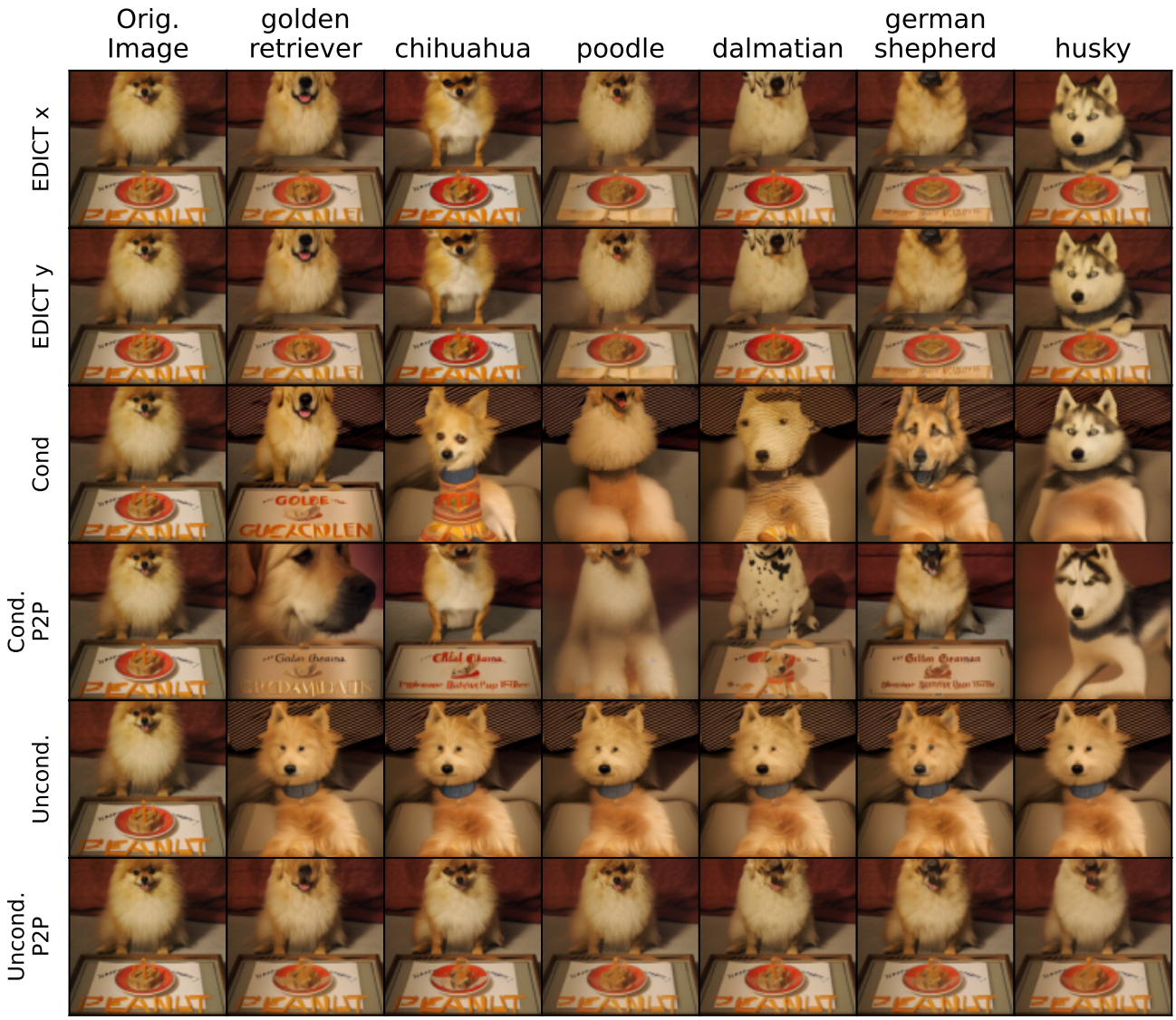

Figure S15. Dog breeds 7