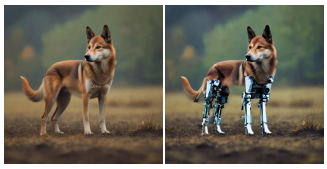
IMPROVING IMAGE EDITING MODELS WITH GENERATIVE DATA REFINEMENT

Frederic Boesel, Robin Rombach
Stability AI

Include a flock of birds and make it a vintage photograph



Change the legs to be bionic

Make it wear a white puffy jacket

ABSTRACT

Instruction-based generative image editing models allow an image to be modified based on a text prompt and have the potential to significantly improve the accessibility of image processing software. Like other generative models, they are highly dependent on the quality of their training dataset, and generating good editing datasets is an expensive task. In this paper, we show that a simple refinement of the original InstructPix2Pix (Brooks et al., 2023) dataset using SDXL (Podell et al., 2023) leads to consistent improvements in downstream models. We finetune SDXL on our refined dataset and observe competitive performance to much more cost-intensive methods. We will make the dataset and models publicly available.

1 INTRODUCTION

Image editing is an extremely popular application for which a myriad of software tools exist. Recently, fueled by breakthroughs around text-to-image diffusion models (Dhariwal & Nichol, 2021; Rombach et al., 2022), instruction-based generative models emerged (Brooks et al., 2023; Zhang et al., 2023a; Sheynin et al., 2023; Zhang et al., 2023b), which offer an intuitive interface for image editing. Their control via natural language has the potential to significantly simplify the interface for image processing and to unify many different editing tools. However, the resulting image and edit quality is sometimes still poor, and improvements have largely been achieved through expensive efforts in dataset curation (Zhang et al., 2023a;b; Sheynin et al., 2023). In this work, we demonstrate that a simple generative refinement of an *existing* instruction dataset can yield competitive results.

The InstructPix2Pix (IP2P) dataset (Brooks et al., 2023) is a large (313k samples) publicly available image editing dataset. Each sample consists of an input image an edit prompt and a target image. The pairs were generated using SD 1.5 (Rombach et al., 2022) with Prompt2Prompt (Hertz et al., 2022). The dataset faces limitations in image quality and edit faithfulness that can be observed in the resulting model, through the relatively low image quality. In this paper we study how a refinement of the IP2P dataset affects the downstream image editing performance. Furthermore we demonstrate that a refinement process in combination with a stronger base model can produce competitive results when compared to the much more data engineering heavy Emu Edit (Sheynin et al., 2023).

Refining an Instruction Dataset We apply an image-to-image refinement of each image in the dataset with the help of the text-to-image diffusion model SDXL Podell et al. (2023). Our goal is to improve upon image quality and leverage the better prompt following ability of SDXL. To do so, we first use HAT (Chen et al., 2023) to upsample each image from 512x512 to 1024x1024 pixels. Next, we apply noise to the input and output image down to a log snr of -4.649 before denoising the images with the SDXL-Base (18 steps) and Refiner model (15 steps) using the original prompts that are provided in the dataset and the same seed for an image pair. We show examples of the results of this refinement process in 3, demonstrating that the resulting images are (i) generally of

higher quality that (ii) better align with the original prompts used to generate the original dataset. In addition to the refinement we employ a CLIP ViT-G/14 (Radford et al., 2021; Ilharco et al., 2021) based filtering to reduce the proportion of unaligned images or bad edits. In particular, we follow Brooks et al. (2023) and filter by an image similarity of 0.7 and a edit direction similarity of 0.2.

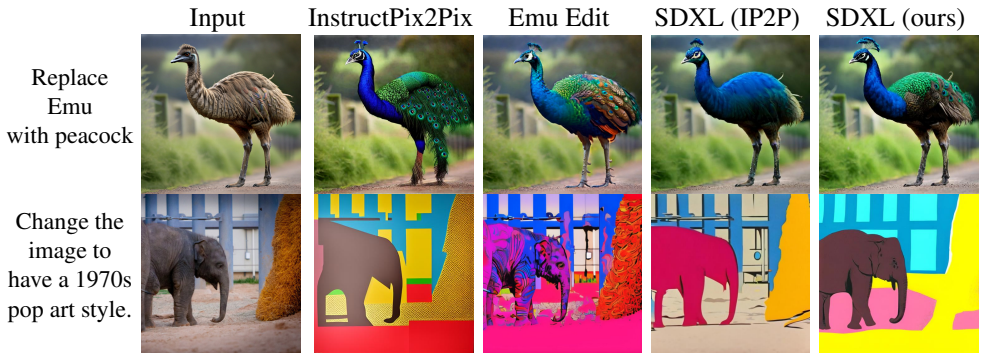


Figure 1: Qualitative comparison of our model to baselines.

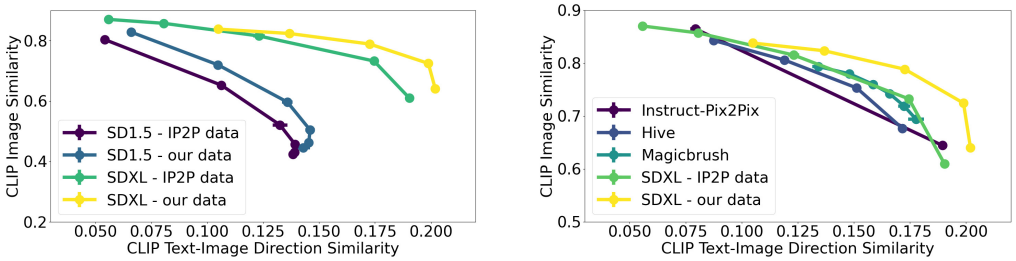


Figure 2: Image Similarity vs Text-Image Direction Similarity for different image classifier free guidance values, evaluated on PIE Bench (Ju et al., 2023). The standard deviations over 3 inference runs are reported for each datapoint. Left: Comparison of SD1.5 and SDXL trained on the IP2P dataset and our refined version. Right: Comparison of our SDXL versions with other benchmarks.

2 Results and Discussion

To assess the effect of our data refinement pipeline, we instruction-tune SD1.5 Rombach et al. (2022) and SDXL Podell et al. (2023) on both the original IP2P dataset and our refined version. For fine-tuning the models, we add four input channels to the input of the UNet to allow for processing the conditioning image and denoise on the edit target. We evaluate our model by computing CLIP image similarity (input image vs edited result) and text-direction similarity (image dir vs text dir); see Figure 2. Moreover, Figure 4 provides a qualitative comparison of the SDXL variants with Emu Edit (Sheynin et al., 2023) and IP2P. We observe a clear qualitative jump in performance when training on the refined data, backing up the results in Figure 2. For a variety of edit tasks we observe similar qualitative performance to Emu Edit, limited to the distribution of tasks in the IP2P dataset.

Our experiments suggest that a simple refinement of the synthetic IP2P dataset with the help of SDXL improves downstream model performance for image editing. We furthermore show that in combination with a stronger base model, compared to SD1.5 used in IP2P, fine-tuning on this refined dataset shows competitive performance with Emu Edit, a model that is trained on a large variety of additional tasks and a $50\times$ larger edit dataset. We note that the inherent limitations coming from the original dataset such as locally targeted editing or targeted colorization are still apparent in the refined dataset and are thus still present in our resulting model; see Figure 5.

Conclusion We hypothesize that a generative refinement process is a promising data cleaning step that can be applied to a variety of other tasks. Furthermore, our refinement process can be used as a cost-reducing measure when creating a synthetic dataset that is subject to heavy filtering, such as the IP2P dataset where 100 samples were generated for each prompt before filtering. Such a dataset could initially be created using a faster and less expensive model before refining the filtered dataset.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22367–22377, June 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2328–2337, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *arXiv preprint arXiv:2306.10012*, 2023a.

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023b.

A Appendix

A.1 Related Work

There are a variety of diffusion model based approaches for image editing, that can be broadly categorised into three groups. Guided Approaches (Hertz et al., 2022; Parmar et al., 2023; Meng et al., 2021) that guide the denoising process of the edited image with information from the source image e.g. via losses or attention copying, image optimisation based methods (Hertz et al., 2023) and model based approaches(Brooks et al., 2023; Zhang et al., 2023a; Sheynin et al., 2023; Zhang et al., 2023b), where a text-to-image diffusion model is fine-tuned based on a paired image dataset. IP2P (Brooks et al., 2023) used Prompt2Prompt (Hertz et al., 2022) to generate a paired image dataset of 450k of which 313k are released publicly. The paired prompts for the IP2P dataset were generated with the help of a fine-tuned GPT3 (Brown et al., 2020) model. Recently there have been a variety of efforts building upon the approach, proposed by Brooks et al. (2023). Magicbrush Zhang et al. (2023a) built a manually created dataset of $\sim$10k edited COCO (Lin et al., 2014) images. Hive (Zhang et al., 2023b) generated more data in a similar fashion to IP2P and trained an edit preference model, which was used during model training. Emu Edit (Sheynin et al., 2023) added generative and descriptive tasks ranging from segmentation to generation based on a depth map. Additionally a high quality dataset of 10 million samples with a masked based filtering approach was generated for a series of edit tasks.

A.2 Implementation details

For the dataset refinement we use EDM Discretization (Karras et al., 2022) with the DPM++ Solver (Lu et al., 2022) as our sampler. Our refined version of the dataset contains 220k samples, the IP2P dataset contains 313k samples. The models are trained with a learning rate of $1e^{-5}$. We train the SDXL variant on a resolution of 512x512 with a batch size of 160 for 20000 steps. The SD1.5 model is trained on 256x256 with a batch size of 400 for 12000 steps. We adopt the dual classifier free guidance (Ho & Salimans, 2022) solution proposed in IP2P. We adjust the values for the SDXL variant to 1.5 for image and 5.0 for text guidance.

A.3 Future Work

As future works we see an opportunity to explore how generative data refinement can be leveraged for other tasks. Additionally, the current method can be improved by a) studying the applied image-to-image strength, which could benefit from an adaptive strategy and b) applying new state of the art text-to-image models.

A.4 ACKNOWLEDGEMENTS

We would like to thank Jonas Müller for feedback on the draft, and typesetting and Axel Sauer for helpful discussions.

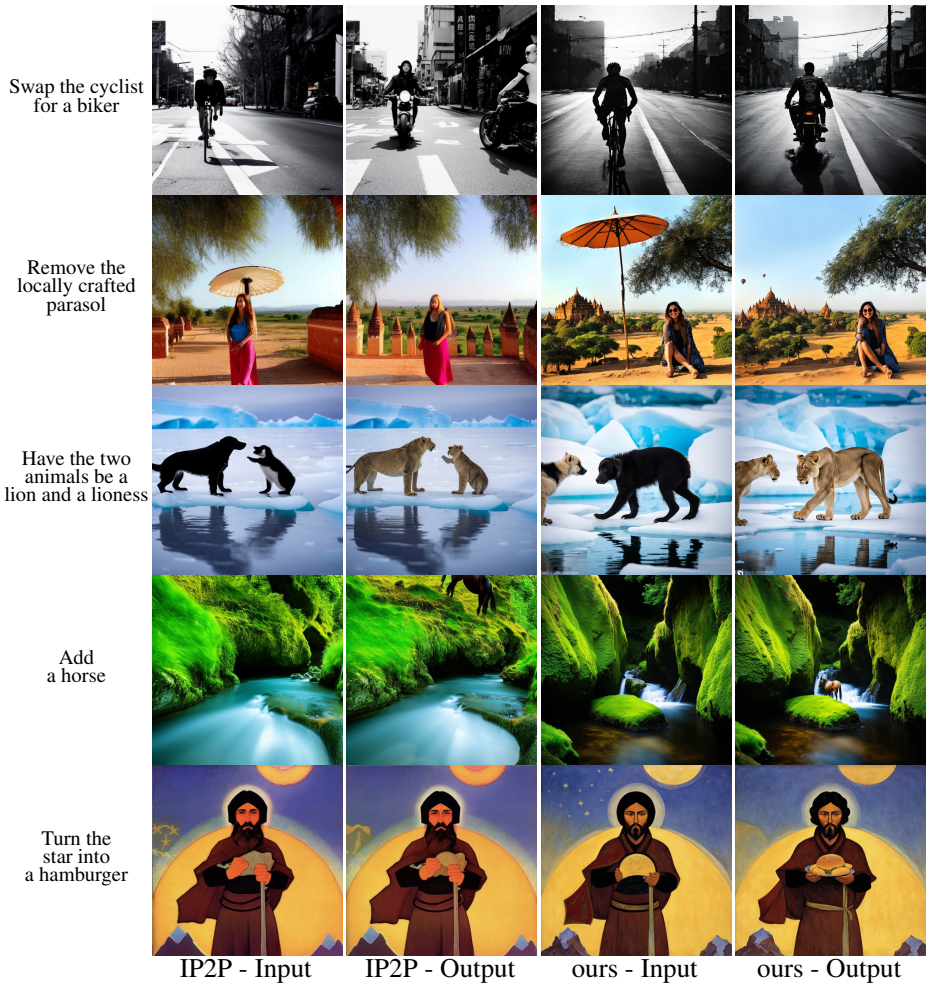


Figure 3: Comparison of samples in the training dataset.

Figure 4: Qualitative comparison of our model to baselines.

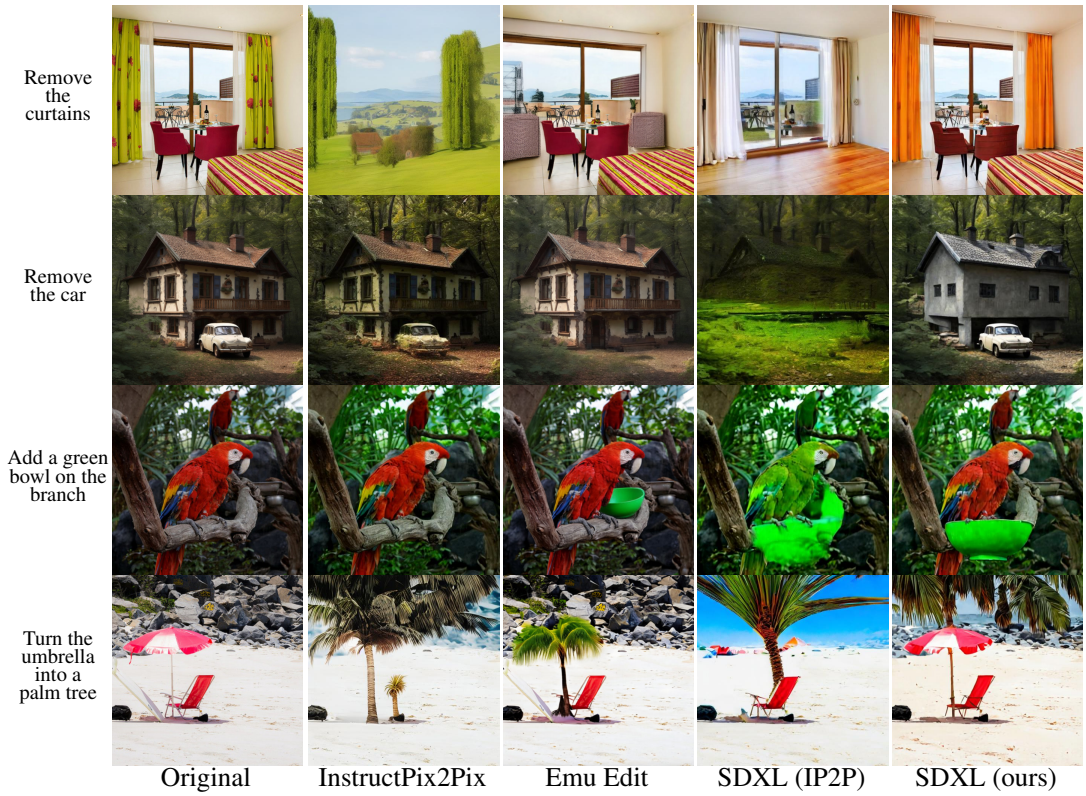


Figure 5: Failure cases. Removing objects, although more reliable, is still inconsistent. The same limitations apply to targeted placement/replacement.

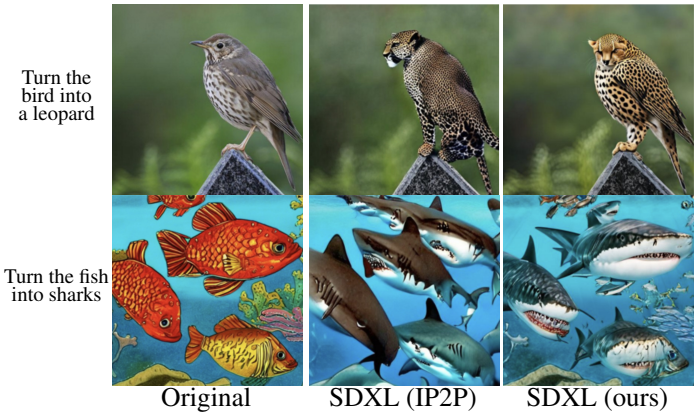


Figure 6: Comparison of samples from SDXL fine-tuned on IP2P data and our data, showing the qualitative visual difference, although using the same base model.

IP2P: Input

IP2P: Output

Figure 7: Example of a sample from the InstructPix2Pix dataset that got filtered out. Instruction: ”Replace the cover with a magnifying glass.” The edited image replaced the person instead, providing a bad signal to the model.