On Consistency Training for Language-Based Image Editing Interface

Youngwon Lee[*], **Ayoung Lee**[*], **Yeonjoon Jung**, **Seung-won Hwang**[†]
Seoul National University
ywlee@ldi.snu.ac.kr, {aylee2020,y970120,seungwonh}@snu.ac.kr

Abstract

This paper studies the training of an image editing interface using language instructions, without requiring expensive human annotations. Such a process involves making necessary changes while preserving the original content as required. For example, when learning an instruction like "change the suitcase to a wine glass," the editing model should be provided with a pair of images with a suitcase and a wine glass, respectively, where the two images share the common background. To obtain such training data, the existing approach capitalizes on a large pretrained language model in tandem with a text-to-image model. Together, they generate a pair of images from the edited caption, derived from the original caption and the instruction. Although this process imposes cross-attention based regulation, towards effectively constraining the Euclidean distance between the images, we posit that this control is still somewhat weak, insufficient for adequately steering the editing interface model in distinguishing where to modify and where to preserve. Our distinctive approach lies in enforcing greater consistency through the utilization of automated object detection and inpainting within a unified pipeline, thereby ensuring the preservation of context. The robust empirical results obtained with our proposed method can be attributed to enforcing "cycle consistency." This signifies that the reverse editing instruction should possess the capability to reconstruct the original image. Our code is publicly available at github.com/aylee2008/ConsEdit.

1 Introduction

Image editing, the task of altering an image based on various contextual cues such as another image or human-annotated masks, has been a central focus in both computer vision and graphics research. The advent of powerful image generation models like GANs and diffusion models, coupled with advancements in jointly learning text and image embeddings as exemplified by CLIP (Radford et al., 2021), has spurred efforts towards incorporating text in image editing. However, many of these endeavors still rely on parallel captions or labels, additional images, human-drawn masks, or per-example fine-tuning (Bar-Tal et al., 2022; Gal et al., 2023; Hertz et al., 2023; Kawar et al., 2023; Rombach et al., 2022; Couairon et al., 2022b). These dependencies limit their capacity to provide a 'natural' language interface for image editing.

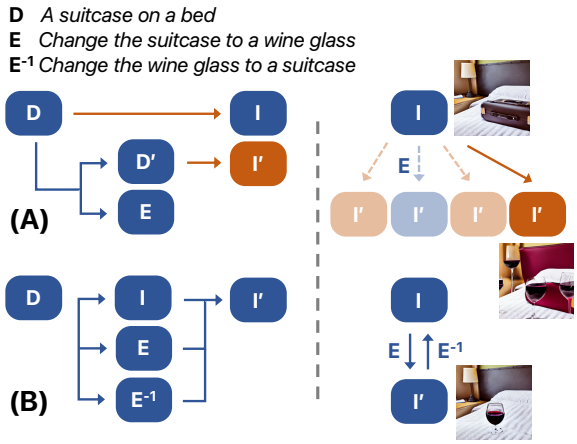

Figure 1: Unlike (A) the baseline method, (B) our method ensures causal relationship between the edit instruction E and the resulting image I', through generating the target image by editing the source image I in the way the reverse edit E^{-1} is well-defined.

In contrast, we focus on editing by a language instruction, as an intuitive means of guiding editing, which affords users two crucial advantages: firstly, supported editing operations can be straightforwardly diversified to arbitrary tasks, as opposed to earlier models focused on singular tasks like style transfer; and secondly, the ability to freely adjust the expressiveness and level of detail in the instruction. The closest work to our focus is Brooks et al. (2023) training a conditional diffusion model

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 22–30
November 1–4, 2023.

that takes an image and the edit instruction in text to generate the edited image. Dubbed InstructPix2Pix, this approach requires **training triplets (original image I, edit instruction E, edited image I')**, for which they leveraged a large pretrained language model and text-to-image generation model.

Figure 1A describes their data creation process where I is an image of a suitcase on a bed obtained from the corresponding description D, E is the edit instruction "change the suitcase to a wine glass," and the resulting image I' ought to be an image of the same bed with a wine glass instead of the suitcase. With the description of the image after editing D', a text-to-image model, namely stable diffusion, is utilized to produce the pair of images (I, I') from both of the description. As there is no assurance that the resulting images closely align in terms of content, cross-attention based regulation as outlined by Hertz et al. (2023) was adopted to force them to look alike. This regulation essentially constrains the Euclidean distance between the images by compelling certain components of the latent image representation associated with specific tokens to be shared by both images. However, as can be seen in the example, the resulting image fails to keep nearby objects such as those on the nightstand or the headboard itself, because such regulation cannot guarantee that the resulting images are close enough in terms of content that should not be affected by the object-level edit E.

To this end, we present a consistency-aware method that leverages LLM, object detection, and inpainting in one shot, which effectively avoids text-to-image generation which is the key source of inconsistency. Our method first generates a list of possible image captions, and edit operations E that can take place in those images as well as the reverse instructions E^{-1} using a finetuned LLM. Then, those generated image captions are provided to stable diffusion model to obtain source images I, which subsequently get masked for the object of interest by running YOLOv7 (Wang et al., 2023) object detection model. Finally, the stable diffusion model is used again for inpainting the masked out area, creating an edited version I' of the image with most of the background that should not be affected by the edit instruction intact. As described in Figure 1B, our proposed method can effectively achieve the very consistency required to provide accurate supervisory signals to image editing interface, with its consistency- and causality-aware design.

Our method combines the strengths of working with language-based instructions while accurately localizing modifications within the input image. Empirical results demonstrate that our proposed approach effectively executes object-level edits with a significantly higher success rate, all while preserving the background intact. Additionally, our model maintains comparable performance across image-level edits as well.

Our contributions can be summarized as follows:

- We highlight the inadequacy of Euclidean distance-based consistency regularization in obtaining high-quality data for training image editing interfaces.
- We introduce the concept of cycle consistency as a solution to this problem.
- We target object-level edits where it is non-trivial to define the inverse edit instruction for enforcing cycle consistency.
- We propose an effective pipeline that successfully achieves cycle consistency by avoiding translating the edited caption back to image.

2 Method

2.1 Baseline: InstructPix2Pix

InstructPix2Pix (Brooks et al., 2023) consists of two phases as described in Figure 2(A). First, LLM is fine-tuned, specifically GPT-3, using human-annotated (source description D, edit instruction E, edited description D') triplets. An 'edited caption' refers to the text description that appropriately corresponds to the edited image achieved by following the edit instruction. This fine-tuned language model is then utilized to generate potential edit instructions and the resulting edited captions from real image captions found in the LAION-Aesthetics dataset (Schuhmann et al., 2022).

Second, a text-to-image generation model, specifically stable diffusion, is employed to generate image pairs (I, I'). These pairs of images form the set of triplets (I, E, I'), which are ultimately used to train the image editing model in a supervised manner. The main challenge with this process lies in ensuring consistency between the two generated images. The text-to-image generation model lacks an inherent mechanism for this, which can adversely affect the training of the image

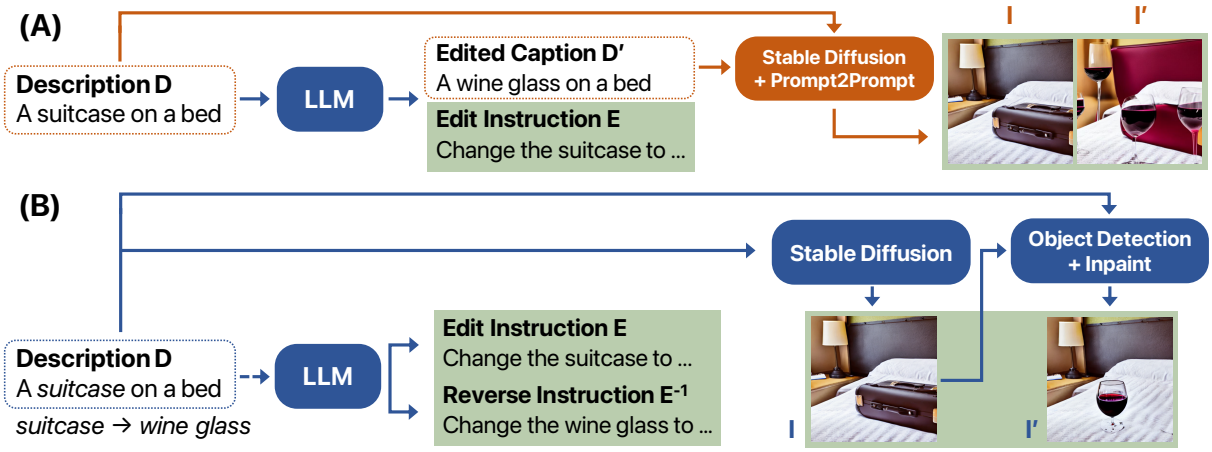

Figure 2: An overview of the training data generation process of (A) InstructPix2Pix (Brooks et al., 2023) and that of (B) our proposed method. Although regulated with mutual distance, generating a pair of images (I, I') from (D, D') as in A (denoted as orange path) induces undesirable changes while our method is free of such problem. Shaded area with green represents the data used to train the editing interface.

editing model. It is crucial to learn object-level edits, distinguishing between what should be changed and what should be preserved.

To mitigate this inconsistency, Brooks et al. (2023) employed a technique introduced by Hertz et al. (2023), which involves regulating the cross-attention weights between the tokens in the edit instruction text and the images. This ensures a more consistent image generation process.

However, as illustrated in both Figure 1 and 2, this approach falls short of effectively aligning the pair of images to a level suitable for use as ground-truth edit demonstrations in the editing interface. We delve further into this specific limitation of InstructPix2Pix from the perspective of cycle consistency in Subsection 2.2.

2.2 Proposed: Consistency-aware image pair generation

To ensure sufficient consistency between the images in each pair, allowing the editing model to properly learn causal modifications, our proposed method has two distinctions: First, we create edit instruction in both directions, to enforce cycle consistency (Subsection 2.2.1). Second, to explicitly preserve the background when generating an edited version of a source image, we propose an integration of object detection and inpainting techniques (Subsection 2.2.2).

2.2.1 Edit, reverse edit instructions generation

We categorize typical types of object-level edit operations as follows:

- Erasing an object;
- Creating a new object; and
- Changing an object to another.

We note that the last one, transformation, differs from a simple combination of removal and insertion edits. In this case, the newly introduced object must precisely replace the space occupied by the original object.

These edits come with natural reverse operations, that is, erasing an object is the reverse of creating one, while changing an object X to Y is the reverse of changing Y to X. Based on such categorization, a finetuned LLM is used to produce the edit instruction E and reverse instruction E^{-1} suitable for a given description D and a target class of object (which can be 'None' for erases). Note that class of the source object – the object to be removed or transformed – can be easily derived from D. For the object classes, COCO (Lin et al., 2014) classes were chosen for that their taxonomy is equipped with moderate granularity and that well-performing object detection systems trained on those data are publicly available.

Using the reverse instruction is inspired by Zhu et al. (2017), introducing the notion of cycle consistency for a different problem context of translating images between domains without paired examples. The idea of cycle consistency, enforcing that the editing of image should be able to be 'back-translated' into the original image with its reverse edit instruction, enhances the consistency in our

(a) Source image

(b) 'Edited' image

Figure 3: An example of image pair which failed validation and thus discarded. (`origin_class`, `target_class`) was ('bear', 'toothbrush').

proposed generation. Prompt-to-prompt (Hertz et al., 2023), regulating the distance between the edited and the source image, fails to meet the cycle consistency requirement for that applying the reverse instruction would corrupt the surroundings again in 'random direction.' Finally, we note that considering the reverse edit E^{-1} as well not only exposes the editing model to the concept of cycle consistency, but also essentially serves the role of doubling the number of training examples.

We used GPT-3 `curie` as the LLM for this process. 400 manually curated examples of description and instructions were used to finetune this LLM. After generating and filtering out results with wrong formatting with this LLM, we obtained nearly 199k examples.

2.2.2 Edited image generation with object detection and inpainting

Now, by providing the description D as input to stable diffusion (Rombach et al., 2022) model, the source image I is obtained. Then, we filter this image with YOLOv7 (Wang et al., 2023) to ensure that each image contains a single object of the source class, while objects of other classes are free to appear. In this process, the position of the source class object is located as a byproduct of object detection, which is masked out to prepare input for the subsequent inpainting phase. The source image, mask which masks out the targeted object, and appropriate text prompt describing how the masked area should be infilled to stable diffusion inpainting model. We simply chose to use "Erase it" for erase operations and "Replace it to [target class]" for transforming operations.

Finally, the resulting inpainted image I' is filtered once again using YOLOv7 in the same way as before to assure the operation is performed correctly. An example of image pair which failed this test can be found in Figure 3. After all this process, about 42k training examples were obtained. We randomly split the data into train and test split with 9:1 ratio.

3 Results

We validate whether the proposed consistency-aware training improves the editing interface both quantitatively and qualitatively.

3.1 Quantitative results

3.1.1 Object-level editing

We begin the comparison of our and baseline models' editing capability regarding object-level edits with presenting the trade-off curve between LPIPS and CSFID metrics. This curve was used by Couairon et al. (2022b) to assess the performance of inpainting models.

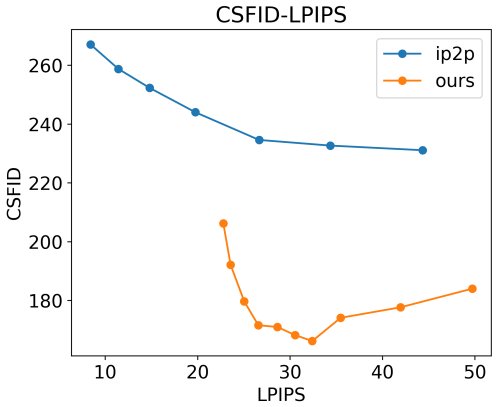

Figure 4: Trade-off curves regarding the LPIPS perceptual distance and class-conditional FID score computed over object-level edits. The steepness of the curve observed when LPIPS distance approaches lower in ours (orange) is due to our model's much higher edit success rate.

LPIPS (Zhang et al., 2018), which stands for 'Learned Perceptual Image Patch Similarity,' calculates the perceptual distance between input and output images based on the feature maps obtained from those images. CSFID (Couairon et al., 2022a), class-conditional Frechet Inception Distance, is the mean FID (Heusel et al., 2017) score of each class. CSFID score becomes lower when the class distribution of input images is similar to that of output images. Thus, LPIPS and CSFID scores are in a trade-off, since it is more difficult for the editing model to change the input image to fully accommodate the target class object when it is required

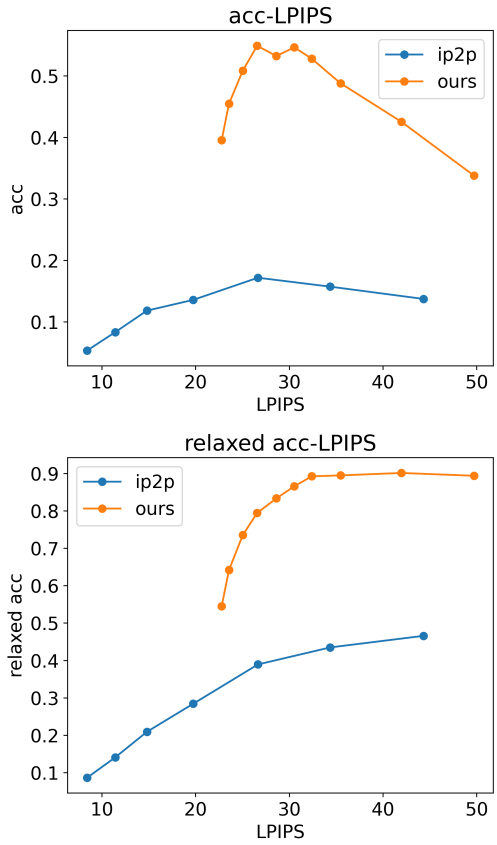

Figure 5: Trade-off curves regarding the LPIPS perceptual distance and (top) accuracy or (bottom) relaxed accuracy, computed over object-level edits.

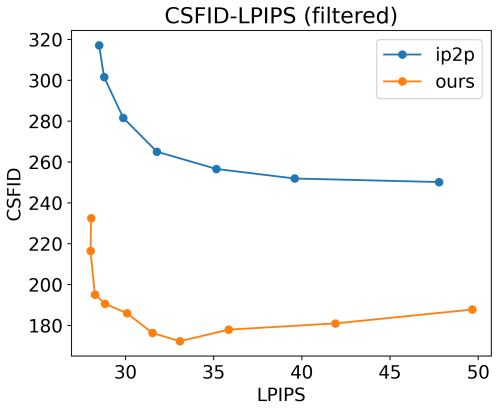

Figure 6: Trade-off curves regarding the LPIPS perceptual distance and class-conditional FID score computed for object-level edits after filtering out edit failures. Upon comparing these results with Figure 4, it becomes evident that the narrowed range of LPIPS values in the outputs of the InstructPix2Pix model indicates two significant points: (1) successful editing leads to an increase in LPIPS distance, and (2) a substantial portion of InstructPix2Pix edit attempts prove to be unsuccessful. This finding aligns with the observation depicted in Figure 5.

to minimize the distance between input image and output image at the same time. This results in a downtrend curve in CSFID-LPIPS plot, as presented in Figure 4, where the more the curve moves towards the origin it is considered more superior.

For computing CSFID score, which is typically reported using ImageNet data, such setting renders it difficult to be adopted to evaluating our model as those images are not accompanied by any edit instructions. Instead, we used the test split of our generated data to calculate CSFID.

Additionally, as part of a comprehensive evaluation protocol tailored specifically for object-level edits, we present a curve depicting the 'edit success rate' versus LPIPS distance, as illustrated in Figure 5. Similar to the automated filtering of generated training data described in Subsection 2.2.2, the decision of edit success hinges on a comparison of the detected objects. This success rate is labeled as 'accuracy' in Figure 5.

Furthermore, we introduce 'relaxed accuracy,' which employs a more lenient criterion for validating successful edits. When performing erasures, we verify solely if the targeted object of the original class has been removed, without considering the fate of other objects, if any were present. Similarly, for creation, we only assess whether a new object of the desired class has been added. In the case of transformations, we check if the object of the original class has been replaced with an object of the target class. The trade-off curve obtained with this relaxed accuracy can also be found in Figure 5.

InstructPix2Pix exhibits much higher edit failure rates than ours does, which explains the vacant area on the left of the orange curve in Figure 4. This is again validated in Figure 6, where the same CSFID-LPIPS trade-off curve as in Figure 4 is plotted again after filtering out failed edits. The resulting curve from InstructPix2Pix resembles the shape of our curve with reduced LPIPS distance span and inferior CSFID scores.

For plotting the points on the curves, we followed Brooks et al. (2023) for choosing the values of guidance scales, fixing $s_T = 7.5$ and varying $s_I \in [1.0, 2.2]$ over a range of numbers uniformly spaced. As our model has a tendency to be less sensitive to the changes in the value of s_I in terms of the perceptual distance of the output image from the original image, we experi-

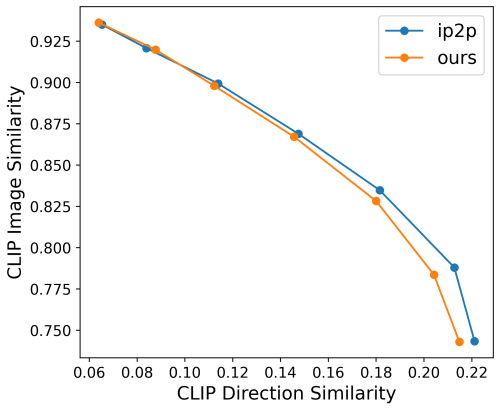

Figure 7: Trade-off curves regarding text-image directional similarity and image-image similarity computed over arbitrary edit instructions from InstructPix2Pix validation dataset.

mented with $s_I \in [0.2, 3.8]$ with twice the strides ($0.2 \rightarrow 0.4$) between points to obtain a curve with enough length.

3.1.2 Image-level editing

While not targeted, our proposed method performs well in the widely studied image-level editing scenarios as well. To make this point, following Brooks et al. (2023), here we report trade-off curvee considering text-image directional similarity and image-image similarity using CLIP (Radford et al., 2021) embeddings.

Text-image directional similarity, introduced by Gal et al. (2022), calculates the alignment between the change in input-output description embeddings and the change in input-output image embeddings. The image similarity is the similarity between CLIP embeddings of input and output images. Indeed, these two scores are also in a trade-off relationship, since the embedding of the output image must diverge from that of the input image if it needs to faithfully reflect the change in corresponding text description.

We use the validation split of dataset generated by InstructPix2Pix for this purpose by choosing 1000 samples randomly, as our data do not have the description for edited image. The result is depicted in Figure 7. Overall, ours achieves comparable performance compared to the baseline model.

3.2 Qualitative comparison

Some of the generated examples from our test split are presented in Figure 8. InstructPix2Pix frequently fails to (1) target the correct part to change and/or (2) preserve the background color and other details.

In order to substantiate our method's qualitative effectiveness, we conducted an additional demonstration involving the image-text retrieval task. This task aims to accurately identify the ground-truth caption corresponding to a given query image.

As widely known, the frequent co-occurrence of objects and backgrounds (e.g., fish and water) often leads to spurious correlation for image-text retrieval models, resulting in the bias of backgrounds in the retrieval (Wang et al., 2020). For example, when an image of a "fish in the water" is provided as the query, its relevance can be biased to the occurrence of water in the background, which may lead to incorrectly retrieving the image of "ship in the water."

In order to show the effect of such a bias, we constructed a challenging evaluation set, consisting of image-caption pairs with large caption overlaps (e.g., more than 3 words), as in the "ship/fish in the water" example. CLIP finetuned with image-caption pairs generated from InstructPix2Pix exhibits a significant drop in retrieval performace on this challenging set. As presented in Table 1, MRR@1 on the challenge set is only 47.84, while it was as high as 84.33 on a generic retrieval scenario.

In contrast, our proposed paired generation, which enforces consistency on the generated image pairs, such as changing the object from fish to wine without changing backgrounds, contributes to debiasing such an effect, as reported in Table 1. Our proposed method consistently and significantly outperforms finetuning on InstructPix2Pix pairs, and the performance drop from a generic retrieval scenario is reduced noticeably as well.

4 Related work

Diffusion-based image editing models Recent advancements in techniques that facilitate joint representations of text and image have transformed the landscape of diffusion models. These developments enable diffusion models to operate within a latent embedding space, rather than the image space (Rombach et al., 2022). This shift not only alleviates the complexities associated with high-resolution images but also enhances their ability to capture semantic nuances. Furthermore, it empowers the manipulation of images in this latent space, offering the potential to leverage the model's inpainting capabilities for image editing. Neverthe-

Method	MRR@k				
	k=1	k=3	k=5	k=10	k=20
InstructPix2Pix (all)	84.33	91.82	91.86	91.86	91.86
InstructPix2Pix (challenge)	47.84	73.00	73.13	73.16	73.16
Ours	**68.26**	**83.58**	**83.65**	**83.66**	**83.66**

Table 1: MRR@k for image-text retrieval task. The retrieval model trained with our data consistently and significantly outperforms baseline over all values of k, particularly with larger margins for smaller k's.

less, limited attention has been given to the utilization of edit instructions beyond conventional methods such as user-drawn masks (or strokes) (Meng et al., 2022), additional images, or full descriptions (Bar-Tal et al., 2022; Kawar et al., 2023; Hertz et al., 2023) as inputs.

Image editing with instructions In response to this gap, InstructPix2Pix (Brooks et al., 2023) has taken a distinct approach by fine-tuning a stable diffusion model in a supervised manner, to enable them to follow edit instructions. This process involves creating a triplet comprising an original image, an edited image, and an edit instruction based on Prompt2Prompt (Hertz et al., 2023) framework.

However, as elaborated previously, this approach has encountered challenges related to ensuring sufficient consistency between the two generated images, which are intended to serve as ground truth examples for the editing model. In a more recent study conducted by Zhang et al. (2023), the generated images were first evaluated by a trained reward model. The quantized scores were subsequently incorporated as additional context during the training of the editing model.

Our distinction It is important to note that our approach distinguishes itself by avoiding the exposure of the editing model to poor-quality training examples lacking consistency. This is achieved by obtaining the target image from the source image in a systematical, consistency-aware editing pipeline that ensures strong consistency between the two images by design, as opposed to generating both images simultaneously in a weakly regulated and less controllable manner.

5 Conclusion

We proposed a new pipeline for training image editing interface that receives natural language instructions solely as input. Based on the observation that the Euclidean distance based consistency regulation using the text-image cross-attention weights of latent diffusion model resulted to poor quality training data which hinders the performance of editing interface, we established that the edited image should preserve contexts that need to be remain invariant, or, it should be able to be 'reversed edited' to produce the source image. The proposed method naturally encodes this idea by creating pairs of edit instructions and their respective reverse operations based on target object classes of interest, exposing the editing model to the concept of cycle consistency. Extensive quantitative and qualitative analyses demonstrate the benefits of our method.

References

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, volume 13675 of *Lecture Notes in Computer Science*, pages 707–723. Springer.

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

Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022a. Flexit: Towards flexible semantic image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18270–18279.

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

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022.

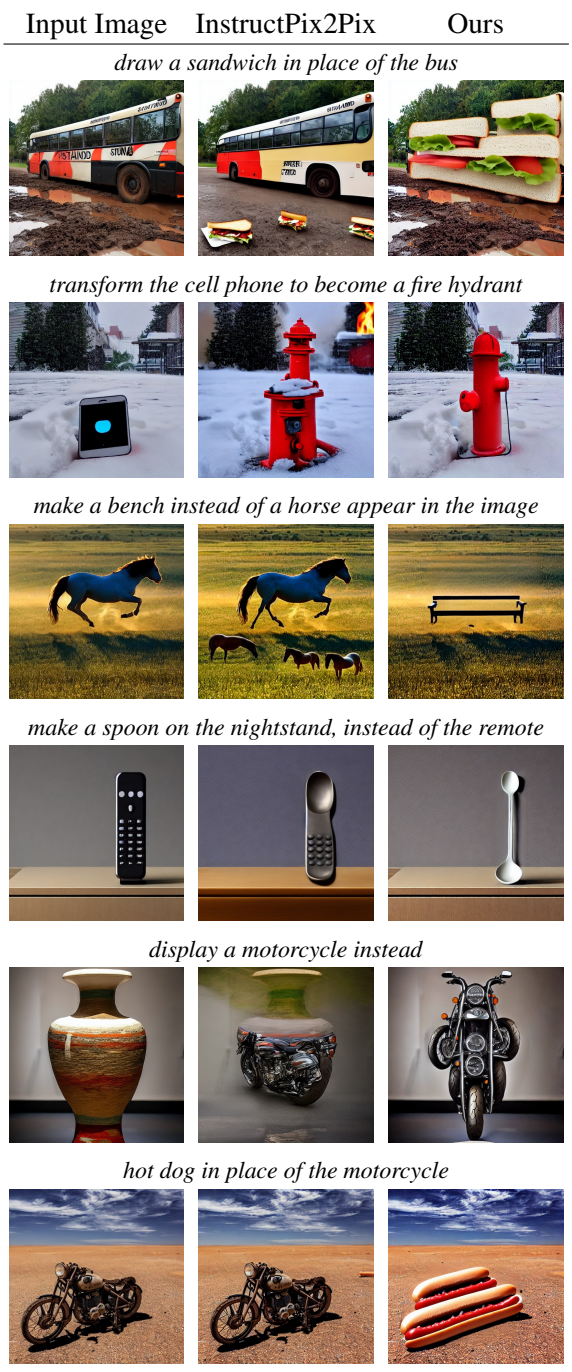

Figure 8: We present some qualitative examples of edited images generated from InstructPix2Pix and our model. We compare two models using localized, object-specific edits using $s_I = 1.4$ for InstructPix2Pix and 2.6 for ours.

Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4).

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6007–6017. IEEE.

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

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural*

Information Processing Systems, volume 35, pages 25278–25294. Curran Associates, Inc.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. 2023. Hive: Harnessing human feedback for instructional visual editing.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society.