REALEDITPLUS: High-fidelity Synthetic Expansion of Instruction based Image Dataset

Akshan Krithick
University of Washington
akshan3@uw.edu

Archit Ganapule
University of Washington
aganap@uw.edu

Keejay Kim
University of Washington
keejayk@uw.edu

Peter Sushko
Allen Institute of AI
peters@allenai.org

Abstract

Generative models such as Stable Diffusion excel at creating realistic images, but struggle with fine grained edits required in real world applications. The REALEDIT [1] dataset (Sushko et al) is a collection of images and paired prompts that describe real image edits that reddit users have created. However, REALEDIT is relatively small (48K posts) and collecting more examples is tedious. A pipeline that could synthetically and significantly extend the REALEDIT dataset while maintaining the natural language tone and visual alignment found in real user requests would enable the development of more capable and generalizable image editing models that better reflect real user needs.

1. Introduction

Image generation and editing have experienced significant improvement through recent advancements in diffusion models. The current text-guided image editing models do not fulfill the requirements of real-world users. A major problem exists because training datasets consist of either limited-size collections or synthetic modifications that fail to match the diversity and actuality of real user demands. Sushko et al introduced RealEdit as a revolutionary dataset that solved this problem. It gathered genuine image edit requests together with their corresponding human-edited outputs from Reddit which formed a dataset containing 48K training examples alongside 9.3K test cases of actual user-driven edits.

Models trained on RealEdit achieved superior performance in realistic editing tasks. Although RealEdit contains 48K examples, these numbers might not fully represent the complete range of possible edits. The process of obtaining human-edited pairs requires extensive manual effort which creates an obstacle to increase the dataset size beyond its current limits.

To improve the scalability and general applicability of text-based image editing we need to explore synthetic data augmentation techniques. Following InstructPix2Pix we suggest using generative models to grow the RealEdit training data because this method demonstrated its ability to teach models about image editing instructions through completely synthetic data. This approach addresses two main goals: (1) Scale – Generative models create hundreds of thousands of training examples at low costs thus addressing the dataset size constraints. (2) Diversity – Through the synthesis of edits on multiple images and instructions we can introduce rare or missing edits such as unusual object modifications and stylistic changes which will help expand the model's capabilities.

The synthetic data is grounded in real images which are the original photos from RealEdit to preserve realism, thus addressing concerns about artificial distributions. This project extends the RealEdit dataset through a combination of state-of-the-art generative tools that generate a large number of new data. The training set images from RealEdit are captioned and receive new possible editing instructions through large language models for each original image. The instructions receive execution through image generative models Stable Diffusion and an instruction-following diffusion model to produce new edited images.

Automated filters monitor the editing process to maintain semantic consistency with the instructions throughout the entire operation. Our goal is to create a massive synthetic image edit triplet dataset (original image, edit instruction, edited image) that exceeds RealEdit in scale by one order of magnitude while

maintaining high quality and realistic edit representations.

Our research establishes a vital connection between authentic human-edited data and the expansive capabilities of synthetic data. This paper evaluates the potential uses of such data for training editing models. We aim to develop more powerful image editing models through RealEdit expansion with synthetic data.

2. Related Work

2.1 Text-Guided Image Editing

There is a growing body of work on editing images based on natural language instructions. Early methods often required paired training data or specific architectures (e.g. image captioning plus GAN inversion). More recently, diffusion models have enabled flexible text-driven edits. For example, SDEdit [2] performs edits by perturbing an input image with noise and refining it with a new text prompt, while Null-Text Inversion in Stable Diffusion finds a latent code that preserves the original image content for better prompt-based edits. Prompt-to-Prompt [3] is another notable approach that allows fine-grained control over generated images by manipulating the diffusion model's attention maps, preserving image structure while changing specified details. InstructPix2Pix [4] by Brooks et al. took a different approach. They trained a diffusion model explicitly to follow image editing instructions. Their model learns to apply a described edit to an input image, without requiring a full description of the output image. Notably, since obtaining large, paired datasets of (image, instruction, edited image) is difficult, Brooks et al. generated their own synthetic training set by combining a language model (GPT-3) to produce instructions and a text-to-image model (Stable Diffusion) to generate edited images. InstructPix2Pix demonstrated compelling results and the ability to generalize to real user-written instructions. This highlights the potential of synthetic data generation for instruction-driven image editing.

2.2 Real vs. Synthetic Datasets

The success of InstructPix2Pix notwithstanding, concerns remain about the realism of synthetic training data. The RealEdit dataset was created in direct response to the shortcomings of earlier synthetic data. RealEdit's authors argue that models trained on artificial edits lacked ecological validity and struggled on genuine user requests. RealEdit provided the first large-scale human-annotated image editing benchmark, and models fine-tuned on it achieved better human-rated outcomes. This underscores the value of real, human-generated edits. On the other hand, purely human-collected data is limited in size (e.g. 48K training pairs in RealEdit) and may not cover the full diversity of edits users might want (for example, creative or rare edits might be under-represented). Recent research in other domains shows a nuanced picture: synthetic data can boost performance if it is of high fidelity and diversity, but naive synthetic data may fail to match real data distributions [5]. Kim et al. note that simply increasing quantity with generated images does not always yield gains, as synthetic sets often miss important real-world characteristics. Thus, a hybrid approach is emerging where synthetic data is used to augment real data rather than replace it. Our work aligns with this strategy: we leverage real images and context from RealEdit as a backbone, and use generative models to create new edited examples. By grounding synthetic edits in real content, we aim to preserve realism while benefiting from virtually unlimited augmentation.

2.3 Synthetic Dataset Generation Techniques

The idea of using generative models to create training data has gained traction across vision tasks. For instance, stable diffusion has been used to generate diverse images for classification, with CLIP-based filtering to select only high-fidelity examples [6]. InstructPix2Pix's pipeline is a prime example of synthetic paired data generation. GPT-3 was prompted to produce an edit instruction and a before/after image caption. Stable Diffusion (with Prompt-to-Prompt constraints) then generated a before-image and after-image pair, and a CLIP-based metric filtered out inconsistent pairs. We build on similar principles, but with some differences; rather than generate a new random image for each instruction, we start from an existing real image , ensuring realism and variety, and we apply edits on it using both a direct editing model (InstructPix2Pix) and diffusion-based image-to-image generation. Using a real source image for every synthetic edit grounds the result in realistic content (backgrounds,

placement of objects, etc.) that pure text-to-image generation might struggle to faithfully reproduce.

2.4 Evaluation Metrics for Image Editing

Evaluating how well an edited image matches the instruction is a non-trivial problem. Simple pixel-wise metrics (L2, PSNR) fail to capture semantic correctness. CLIP-based scores have become a common proxy for assessing image-text alignment. For example, CLIPScore compares the CLIP embedding of a generated image with the embedding of the target description, and has shown good correlation with human judgments in captioning tasks [7]. In editing, one can similarly measure how well the edited image and the instruction align in CLIP space, or even use the directional CLIP similarity between the original and edited image captions to gauge if the intended change was achieved [8]. InstructPix2Pix employed such a directional CLIP metric to filter its synthetic data, removing cases where the change in the image did not match the change described in text. RealEdit's benchmarking protocol goes further, using learned vision-language models to ask question-answer pairs about the edit (VIEScore) and even GPT-4 for open-ended assessment. In our pipeline, we incorporate CLIP-based filtering as a lightweight automatic check on each generated example's quality. While CLIP scores are coarse (and can saturate for fine-grained differences), they are useful for pruning clearly invalid outputs at scale. Ultimately, a thorough evaluation of our expanded dataset will require both automatic metrics and human judgment, as discussed later.

3. Methods and Experiments

Our data generation pipeline expands the RealEdit dataset by creating new edit examples for each original image. Figure 1 gives an overview of the process.

3.1 Data Acquisition

Each original image from REALEDIT is downloaded via its URL (Imgur/Dropbox) and checked for validity. Verified images serve as the inputs for synthetic editing.

3.2 Caption Generation

The original image is fed to GPT-4o model, which has vision capability, hence produces a descriptive caption. We found GPT-4o's descriptions to be detailed and reliable, providing a strong basis for generating edit instructions. Stable Diffusion's text encoder is a CLIP-based model, so it cannot ingest arbitrarily long text. To overcome this, we truncate each caption to at most 77 tokens. This ensures the caption will not be cut off by the CLIP tokenizer. The captioning step thus yields a pair (image, original_caption) for each original image.

3.3 Instruction Generation

Given the original image caption, GPT-3.5 -Turbo generates 10 distinct edit instructions that a user might request, each paired with an expected outcome description (edited caption). These instructions mimic real user tone and intent, covering a variety of plausible edits (e.g., object removal, color changes, background swaps). The edited caption is essentially a hypothetical description of the image after the edit is applied. This is useful for two reasons: (a) it helps in guiding image generation if we use a text-to-image model, and (b) it provides a reference for evaluation (we later compare the edited image to this caption via CLIP). We generate up to 10 such instructions per image to maximize diversity and dataset size. We chose 10 based on a balance between having many examples and not overloading a single image with too many similar edits. We instruct GPT-3.5 to avoid NSFW or nonsensical edits.

3.4 Image Editing

For each edited caption, we condition a Stable Diffusion v1.5-based image-to-image model on both the original and edited captions, starting from the encoded latent of the original image. Our pipeline adapts the InstructPix2Pix framework but uses a standard Stable Diffusion model with an MSE-finetuned VAE for improved reconstruction quality. We experimented with initializing each edit from the original image latent to preserve spatial consistency, but found that this does not fully maintain pixel-level alignment across edits; nevertheless, it helps maintain semantic grounding. For each prompt, we generate a pair of images: one conditioned on the original caption and one on the edited caption. K-diffusion samplers are used for stable, high-quality outputs. This pipeline structure so far enables us to systematically generate consistent and diverse image-

edit pairs suitable for training text-guided image editing models.

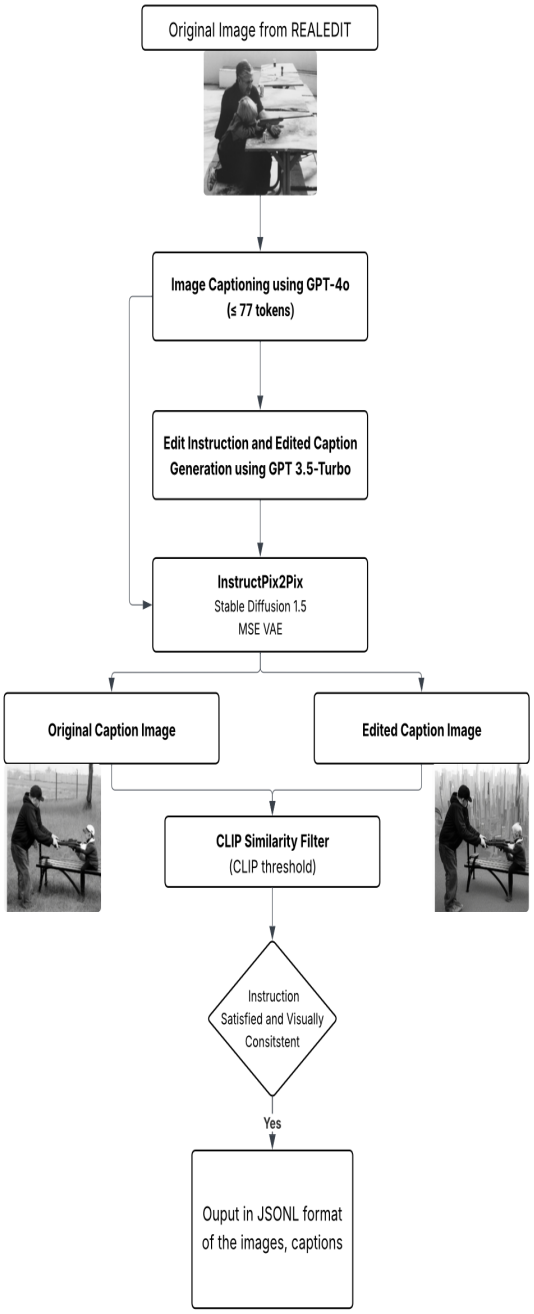


Fig.1. Architecture Diagram of REALEDITPLUS Pipeline

3.5 CLIP Filtering

Each edited image is evaluated using a CLIP similarity check. We compute the similarity between the original and edited image (in the CLIP embedding space) to ensure the edit has not drifted too far from the source content. Edits that fail to reflect the instruction or significantly alter the image beyond the intended change are filtered out. This step acts as a quality control, keeping only realistic and instruction-faithful edits. This is called the CLIP directional similarity score. We embed the original caption and edited caption with CLIP's text encoder, and the original image and edited image with CLIP's image encoder. We then check if the change from original to edited in text space is reflected by a similar change in image space. This technique is used in InstructPix2Pix's data generation and gives a measure of whether the edit happened as described. If the score is below a threshold, we consider that a failure to accomplish the instruction and discard that pair.

We also use absolute image-text similarity as a secondary check; the edited image's CLIP embedding should match the edited caption's embedding with a cosine similarity above a minimum value. We found CLIP filtering to be effective at catching obvious mismatches or hallucinations. It is less sensitive to very fine details, so results may still include some subtle edits that CLIP can't detect, which is acceptable as long as the instruction is followed.

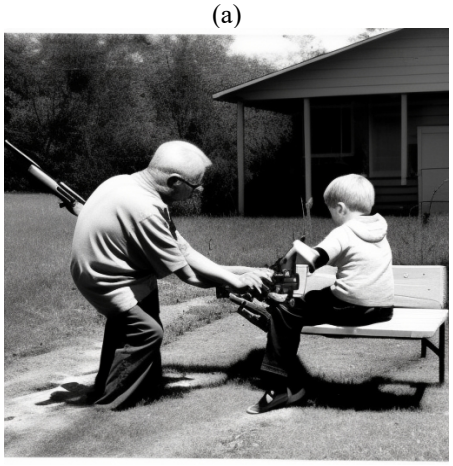
3.6 Output Assembly

Finally, the pipeline records the results in a structured format. For each original image, we store its original caption and the set of successful edits. Each edit has the instruction, the edited caption, and the edited image. The data is saved in JSONL format, one entry per edited image, linking it back to the original.

3.7 Results

Using the above pipeline, we have generated a synthetic expanded dataset. Figure 2 (c) shows a few examples of the synthetic edits produced. Each example is formatted as the original image with its GPT-4o caption, followed by one of the GPT-3.5-generated instructions and the resulting edited image with its caption. We emphasize that none of the Figure 2 (b) and (c) images existed originally, and they are entirely model-generated based on the original photo from RealEdit dataset.

(a)



(b)

(c)

Fig.2. (a) Original Image from RealEdit dataset. (b) Images generated from Original Image Caption. (c) Images generated from Edited Image Captions.

Figure 2 (a) is the original image retrieved from the RealEdit dataset, which is captioned using GPT-4o. Figure 2 (b) images were generated based on the caption of the original image. That is, these images are of the same caption as the one GPT-4o provided. The captions of both Figure 2 (a) and (b) images are "*A young child aims a rifle with the assistance of an*

older adult on a bench. The setting appears outdoors with other benches, scattered soda cans, and containers on the tables. The image is in black and white".

Figure 2 (c) images were generated based on edited captions while conditioning on their respective generated original caption images' latent space. Figure 2 (c) edited captions are: "*A young child aims a rifle with the assistance of an older adult on a bench. The setting appears outdoors with other benches, scattered soda cans, and containers on the tables. The image is sepia-toned* " , "*A young child aims a rifle with the assistance of an older adult on a bench. Colorful balloons decorate the background, adding a cheerful atmosphere to the scene*", "*A young child aims a rifle with the assistance of an older adult on a bench. The setting transforms into a futuristic cityscape with skyscrapers and flying vehicles. The image is in black and white*" respectively.

We emphasize that both Figure 2(b) and Figure 2(c) images are fully model-generated outputs; only Figure 2(a) is an original RealEdit dataset image. We also used CLIP-based similarity metrics to filter outputs and assess edit faithfulness, with results indicating that the edits closely follow the intended caption changes.

4. Discussion

The research demonstrates how to scale a real-world user-driven dataset like REALEDIT through the combination of strong language and diffusion models. The combination of GPT-4o-generated captions with GPT-3.5-produced edit instructions proved successful in capturing realistic user intent. The modified Stable Diffusion v1.5 pipeline with its fine-tuned VAE and k-diffusion samplers produced high-quality synthetic edits which demonstrated the potential for automated large-scale dataset expansion.

The CLIP-based filtering system proved effective in preserving high semantic coherence between the original images and their edited versions. The filtering process protects the edits from becoming irrelevant or unrealistic thus minimizing the presence of synthetic edits that lack quality. The system demonstrated strong overall performance yet struggled with accurately processing edits that required major transformations or delicate artistic details.

We tested different architectural approaches and latent reuse methods to achieve spatial consistency in our results. The consistent use of original image latents resulted in degraded visual quality because it created a trade-off between maintaining semantic consistency and achieving pixel-level fidelity. The selected method of independent image generation through caption conditioning enables semantic relevance preservation without compromising image quality.

5. Limitations

While our pipeline effectively generates diverse and realistic synthetic edits, several limitations emerged:

1. **Complex and Extreme Edits:** The model faces difficulties with complex transformations (e.g., detailed background replacements or significant style changes) and sometimes produces subtle artifacts or incomplete modifications.

2. **CLIP Filtering Sensitivity:** The CLIP-based filtering mechanism is strong, but it sometimes eliminates valid, substantial edits because of its strict similarity thresholds. The system fails to detect valid extreme edit modifications when the edit instruction itself is extreme, which causes the extreme edit.

6. Future Work

The research directions for future development include:

1. **Advanced Diffusion Models:** Exploring the integration of newer diffusion architectures, such as Stable Diffusion XL or diffusion models explicitly trained on complex editing tasks, will improve visual quality and extend editing capabilities.

2. **Aspect Ratio Diversity:** Addition of images with different aspect ratios and resolutions would increase dataset diversity which could lead to better model robustness and generalization.

3. **Data Augmentation Techniques:** Exploring image transformation methods (e.g., rotation, scaling, or color augmentation etc.) could further diversify synthetic edits and enhance model capabilities.

4. **Consistent Original Images:** Keeping the generated original image constant across multiple edits, ensuring edit variations are isolated, making it easier to compare.

By addressing these areas, our pipeline can become even more effective at producing large-scale, high-quality, and highly realistic synthetic datasets, significantly benefiting future research in text-guided image editing.

References

[1] Sushko, Peter, Ayana Bharadwaj, Zhi Yang Lim, Vasily Ilin, Ben Caffee, Dongping Chen, Mohammadreza Salehi, Cheng-Yu Hsieh, and Ranjay Krishna. "REALEDIT: Reddit Edits As a Large-scale Empirical Dataset for Image Transformations." *arXiv preprint arXiv:2502.03629* (2025).

[2] 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).

[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).

[4] Brooks, T., Holynski, A., & Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18392-18402).

[5] Kim, J. M., Alaniz, S., Schmid, C., & Akata, Z. (2025). LoFT: LoRA-fused Training Dataset Generation with Few-shot Guidance. *arXiv preprint arXiv: 2505.11703*.

[6] Dunlap, L., Umino, A., Zhang, H., Yang, J., Gonzalez, J. E., & Darrell, T. (2023). Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in neural information processing systems*, *36*, 79024-79034.

[7] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

[8] Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, *41*(4), 1-13.