Adressing the lack of data in prompt-based interior design image editing task

Victor-Eugen Zarzu
Babeş-Bolyai University, 1, M. Kogălniceanu street
Cluj-Napoca, Romania
victorzarzu@gmail.com

***Abstract*—I propose a method for creating a dataset aimed to improve the prompt-based interior design image task editing in a supervised way with no prior existing data. The image editing task implies given an input image and an edit instruction for the image to output the image resulted by altering the initial image with respect to the given instruction. Data scarcity is a significant issue in the image editing task as it is difficult to collect pairs of images after and before applying a certain editing. However, even though the state-of-the-art model performs well on average, it does not perform well on the interior-design case.**

I. Introduction

The presented method shows the possiblity of creating data for prompt-based image editing task for a interior-design case, with the posibility of extension to various other cases. Training data for image-editing task is hard to acquire at scale, but as presented [1] it can be generated using a Large Language Model and a text-to-image model starting with initial descriptions of the images that need to be edited. However, for interior-design case, there are no such public existent descriptions datasets. For generating the captions prior to generating the images, GPT-3.5 it's used and, for generating the images, a Stable Diffusion model is used.

II. Related work

A. InstructPix2Pix

In [1] it is proposed a method for generating data for training a model for prompt-based image editing task based on two pre-trained models: a model language model and a text-to-image model. The presented approach relies on a pre-existent big dataset with descriptions of images for the desired task. My approach is similar to prior work, but it is also addressing the lack of an existing dataset for interior design room or object descriptions by leveraging the knowledge of recent Large Language Models. Additionally, it presents a fast method of creating data that can be extended to a wide variety of use cases (e.g. cartoons).

B. SSCR: Iterative Language-Based Image Editing via Self-Supervised Counterfactual Reasoning

In [2] it is presented a method for addressing data scarcity in Iterative Language-Based Image Editing that achieves by using just 50% of the data, a comparable result to using complete data. This method can be used for further training for a prompt-based image editing model on the dataset created with the proposed approach. This is preferable for reducing the bias and limitations of context of the Large Language Models and for letting the model explored alternatives that were not captured by the LLM model.

III. Method

The instruction-based image editing is treated a supervised learning problem. The dataset generation consists of 2 parts: the generation of text editing instructions and the generation of pair of images based on those captions.

A. Text editing instructions generation

The text editing instructions for generating the dataset consists of 3 elements: 1) the description of the initial room or object 2) the edit instruction 3) the initial description modified by the editing instruction. For addressing the absence of room descriptions, the Large Language Model was queried to generate all 3 components (see Figure 1), compared to [1] where the last 2 components of the tuple is generated based on a previously known description. By leveraging the knowledge of language model, there is no need of fine-tuning it, but experiments in this area were done. With the proposed approach, by just presenting to the language model the format of the desired output and 3 other examples of the format, it is able to generate a big amount of data in the desired form with a great variety in responses. Additionally, the presented method has the advantage of enabling creation of data in a hierarchical way of difficulty for the editing model: it first creates paired editing captions for single objects followed by paired captions with a description of rooms with more objects. Additionally, compared to[1], the presented method can be extended and used for any other special case of prompt-based image editing, without the prior need of data. For the experiments GPT-3.5 was used as text generator.

B. Generating images from paired editing instructions

Starting from the paired editing instructions generated with the previous method, a text-to-image model is used for generating the dataset in a supervised way: the image before and after edition. However, generating one image for each instruction does not guarantee that they are consistent. For addressing this issue, similar to the approach presented in [1], a number of 30 pairs of images are generated for each

Fig. 1: Example of prompt for GPT-3.5

Generate JSON objects with other 10 such examples, different than the previous ones, but with for living room for interior design editing, but with just 1 feature change at a time (change, addition or removal), different from the previous answers. Respect the most succinct format like this: JSON is on a single line in format {”input:”, ”edit:”, ”output:”}, with one JSON per line, without other text between JSONs and without any other message. {”input”: ”An elegant dining room with a long table that seats eight, a statement chandelier, and plush velvet chairs.”, ”edit”: ”Replace the statement chandelier with a modern pendant light.”, ”output”: ”An elegant dining room with a long table that seats eight, a modern pendant light, and plush velvet chairs.”}

pair of captions, followed by a CLIP-based metric filtering introduced by *Gal et al.* [3]. This metric measures the consistency between the change of two images with respect the change between the two captions that describe the images. So, using this metric, after the images are generated using a Stable Diffusion [4], only the top 4 pairs of images that are above the image similarity threshold of 0.7 are kept. Compared to the [1], where for every pair of captions 100 image pairs are generated before filtering, my approach split the generation in 2 parts for reducing the time of generation: for the images with single objects 30 image pairs are generated and for rooms 50 pairs are generated. While these shows

IV. Results

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros, *Instructpix2pix: Learning to follow image editing instructions*, 2023.

[2] Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang, *Sscr: Iterative language-based image editing via self-supervised counterfactual reasoning*, 2020.

[3] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or, *Stylegan-nada: Clip-guided domain adaptation of image generators*, 2021.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, *High-resolution image synthesis with latent diffusion models*, 2022.