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

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

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Related work

A. InstructPix2Pix

[1] proposes a method 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 levereging the knowledge of recent Large Language Models.

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 resulted dataset. This is necessary and preferable for reducing the bias and limitations of context of the Large Language Models.

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 editing instructions.

A. Text editing instructions generation

The text editing instructions consists of 3 elements: 1) the description of the room or object to modify 2) the edit instruction 3) the intial description modified by the editing instruction. For addressing the absence of room descriptions, the Large Language Model was queired to generate all 3 components, 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 to fine-tuning it. 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. Additionally, the presented method creates 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.

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, followed by a CLIP-based metric filtering introduced by *Gal et al.* [3] is used. This metrics measures the consistency between the change of two images with respect the change between the two captions that describe the images.

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.