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
victor.zarzu@stud.ubbcluj.ro

***Abstract*—**

The prompt-base image editing task got a lot of attention in recent years due to its immense potential in various applications, allowing users to perform natural language based image editing. However, the process of collecting the data for training models on this task is difficult and time consuming, yet essential for achieving optimal performance. Furthermore, for the interior-design prompt-based image editing case, the top existent models do not show a good performance due to the scarcity of relevant data. So, for improving the performance, we propose a solution for generating a dataset without the need for prior data .The approach is exploiting the remarkable capabilities of the large language models to provide all the required information. This method offers a way of generating a great amount of training samples, making it extensible for multiple other cases.

I. Introduction

The problem of data scarcity is a recurrent problem in prompt-based image editing task because of the difficulty in collecting images before and after a specific edit instruction at scale. Previous methods [1] showed a way of generating such data starting from text descriptions of the initial images by fine-tuning a large language model to generate edit prompts and the resulting edited description based on these prompts, followed by a text-to-image model (Stable Diffusion [4]) for generating the images. This methods are available when a large dataset of initial descriptions is available, which is not the case for the interior-design case. The impact of this lack of data is directly visible in the state-of-the-art model's results, which does not shows good results for this task. The proposed approach shows a solution to this unavailability by leveraging the knowledge of the recent large language models for generating all the necessary textual data, showing impressive results.

II. Related work

A. InstructPix2Pix

In [1] a method for generating data for training a model for prompt-based image editing task is proposed, 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. The proposed approach is similar to prior work, but it is also addressing the lack of an existing dataset for interior design room or object descriptions. We leverage the knowledge of recent Large Language Models by including the initial description in the data generation, showing inpressive results. Additionally, it presents a faster 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] a method for addressing data scarcity in Iterative Language-Based Image Editing task is presented, approach 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 explore alternatives that were not captured by the LLM model.

C. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

The method of generating a caption based on an initial image presented in [7] can be used for generating the initial prompts based on a given image based on the available image datasets. This would be followed by a generation of edit prompts with a large language model and a generation of output images from this prompts. However, compared to the proposed approach, this method would have 2 downsides. The first one is related to the amount of available data, which is limited to around 6,000 samples by combining the Interior Design IKEA dataset [5] and the House Rooms Image Dataset. The second one is related to the difficulty of generating the output image consistent with the initial one based on the edited prompt generated by the large language model. The presented method aims to generate as much data as needed for the task without any restrictions on existing data.

III. Method

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

TABLE I: GPT-3.5 Editing Caption Generation

Original Caption	Edit Instruction	Edited Caption
A farmhouse-style bathroom with a distressed wood vanity, a copper sink, a clawfoot bathtub, and vintage-inspired fixtures.	Change the copper sink to a porcelain sink.	A farmhouse-style bathroom with a distressed wood vanity, a porcelain sink, a clawfoot bathtub, and vintage-inspired fixtures.
A cozy living room with a fireplace, a plush sectional sofa, and a coffee table.	Remove the coffee table.	A cozy living room with a fireplace, a plush sectional sofa.

2) the edit instruction
3) the initial description modified by the editing instruction.

For addressing the absence of initial descriptions, the Large Language Model was queried 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 the 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 a 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 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 to 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, the proposed 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.

IV. Results

A. Generating editing captions

The generating of all text components using GPT-3.5 shows impressive results, being able to build captions for multiple contexts (see Table I) and provide multiple edit instructions for the same context. Furthermore, it shows that a fast generation of such data is possible by leveraging the impressive abilities of large language models without the need for prior datasets.

B. Alternatives explored

1) GPT-4:

For generating the pairs of captions, the newly released GPT-4 was used, but the results were not much different from the GPT-3.5's ones. The new model aims to build descriptions from multiple contexts, but the lack of diversity is still present similar to the data generated by GPT-3.5. However, the data generation with GPT-4 was significantly slower than with its previous model due its complexity.

2) Llama2 13B:

The option of fine-tuning the last released Llama2 [6] open-source model was explored as well. For the sake of experimentation, the 13B version was used and fine-tuned on the data generated by the GPT model. Nonetheless, due to the small size of the model compared to the 70B or GPT-4, the lack of diversity and small number of samples in the previously generated data, Llama2 was not able to produce good results. The data used for fine-tuning can be found **here**.

Fig. 1: Images created based on the generated captions

(a) **(b)** **(c)**

Images generated based on descriptions generated by GPT-3.5: (a) A farmhouse-style bathroom with a distressed wood vanity, a copper sink, a clawfoot bathtub, and vintage-inspired fixtures. (b) A cozy living room with a fireplace, a plush sectional sofa, and a coffee table. (c) A modern kitchen with stainless steel appliances, granite countertops, and a farmhouse sink.

C. Image generation

Based on the prompts generated by the large language model, the pipeline formed of a Stable Diffusion Model followed by a CLIP filter shows good results in terms of diversity and correctness. Having a CLIP threshold of 0.7 for the similarity between images, having 40 sampling steps for each image

Fig. 2: Comparison between the generated data and InstructPix2Pix performance on it

and generating 50 pairs of images per prompt gives high quality data in relative short time (see Figure 1). Furthermore, the generated data exposes the limitations of InstructPix2Pix model in generalizing for the interior-design case and its poor performance on it (see Figure 2).

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.

[5] Ivona Tautkute, Aleksandra Możejko, Wojciech Stokowiec, Tomasz Trzciński, Łukasz Brocki, and Krzysztof Marasek, *What looks good with my sofa: Multimodal search engine for interior design*, Proceedings of the 2017 federated conference on computer science and information systems, 2017, pp. 1275–1282.

[6] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom, *Llama 2: Open foundation and fine-tuned chat models*, 2023.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio, *Show, attend and tell: Neural image caption generation with visual attention*, 2016.