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 and available data, we propose a solution for generating a dataset without the need of prior data. The approach is leveraging the remarkable capabilities of the large language models to provide all the required textual information for generating images based on captions, showing promising results in dataset creation for the task in discussion. This method offers a way of generating a great amount of training samples, making it extensible for multiple other cases.

I. Classification

A. ACM

1) I.2.7 Natural Language Processing
2) I.2.10 Vision and Scene Understanding
3) I.4.6: Segmentation

B. AMS

1) 68T45: Machine vision and scene understanding
2) 68U10: Image processing
3) 68T07: Artificial neural networks and deep learning
4) 68T50: Natural language processing

II. Introduction

The problem of data scarcity is a recurrent problem in prompt-based image editing tasks 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 [7]) for generating the images. These methods are available when a large dataset of initial descriptions is available, which is not the case for the interior-design setting. The impact of this lack of data is directly visible in the state-of-the-art model's results, which does not show good performance for the task in discussion. 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. Additionally, it shows the incapabilities of the current models in performing well in the interior-design setting and raises the question if a model can be trained on this task with no prior existing data. The paper has two main parts: the creation of paired captions and the generation of images based on them, followed by experiments with various LLMs and a method of data augmentation for reducing the noise.

III. 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 large language model and a text-to-image model. The presented approach relies on LAION-Aesthetics V2 6.5+ [9], a pre-existent big dataset with different text descriptions, used as the starting point for text generation. 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 impressive results. Additionally, it presents a faster method of creating data that can be extended to a wide variety of use cases (e.g. cartoons).

Fig. 1: Images created based on the generated captions

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

Images generated based on descriptions generated by GPT-4: (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.

TABLE I: GPT-4 Editing caption generation

Original Caption	Edit Instruction	Edited Caption
A luxurious foyer with marble flooring, an ornate console table with a vase of flowers, and a large painting on the wall.	Remove the vase of flowers from the console table.	A luxurious foyer with marble flooring, an ornate console table without a vase of flowers, and a large painting on the wall.
A contemporary living room with a wall-mounted TV, a beige sofa without pillows, and a sleek black floor lamp.	Add a set of blue velvet pillows on the sofa.	A contemporary living room with a wall-mounted TV, a beige sofa with a set of blue velvet pillows, and a sleek black floor lamp.

B. Emu Edit

In the Emu Edit paper [10] introduced by Sheynin et al. a new method of generating training data for the prompt-based image editing task is proposed that tries to solve the lack of diversity and introduced noise in the generated dataset. Their dataset offers more diversity by creating a generative model for each subtask (Region-Based Editing, Free-Form Editing and Vision task) and less noise by creating a more comprehensive filtering pipeline consisting of task reassignment, CLIP, L1 distance between the depth map of the input image and the edited image and image detectors to validate the existence, absence or replacement of the desired elements from the edit instruction. The proposed approach also tries to capture the last part of the filtering from the text generation, but, however, there is still noise introduced by the Stable Diffusion model which does not always respect the descriptions (see Figure 3). Furthermore, the approach presented in Emu Edit still relies on a pre-existent set of text descriptions which is a limitation for the interior-design setting.

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

In [4] a method for addressing data scarcity in Iterative Language-Based Image Editing task is presented, an 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.

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

The method of generating a caption based on an initial image presented in [13] 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 these prompts. However, compared to the proposed approach, this method would have 3 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 [11] and the House Rooms Image Dataset. The second one is generated by the image-to-text model which will also introduce some noise to the data by not creating a reliable description of the image. And the last 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. Compared to this one, the presented method aims to generate as much data as needed for the task without any restrictions on existing data.

IV. 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
2) the edit instruction
3) the initial description modified with respect to the editing instruction.

For addressing the absence of initial descriptions, GPT-4, the large language model used for experiments, 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 in that format, it is able to generate a big amount of data in the desired form with a great variety in responses. Moreover, for reducing the noise, GPT-4 was instructed to clearly state that the object is missing or exists in the original description for the add and remove action, respectively. 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 captions for single objects followed by the ones with a description of rooms with more objects. 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.

	MTLD	Dugast's U^2	Guiraud's Index	Yule's K
GPT-3.5	28.13	12.83	3.87	356.49
GPT-4	**32.72**	**13.73**	**4.52**	**278.80**

TABLE II: Comparison of text diversity in the textual data generated by GPT models

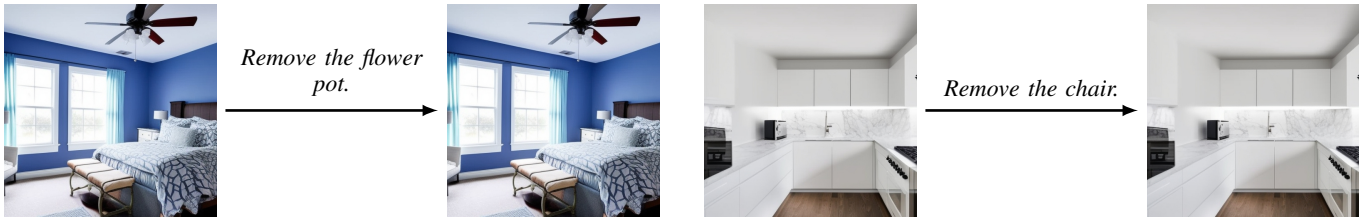

Fig. 2: Example of data augmentation with no change in the output image for reducing the noise

B. Generating images from paired editing instructions

Starting from the paired editing instructions generated with the previous method, Stable Diffusion 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 their consistency. 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.* [5]. 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 [7], only the top 4 pairs of images that are above the image similarity threshold of 0.75 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 with multiple objects 50 pairs are generated. Furthermore, for reducing the noise introduced by the models used for generation and for increasing the dataset size, samples with an edit instruction that does not alter the initial image are introduced for augmentation (see Figure 2).

V. Experiments

A. Generating editing captions

The generating of all text components using GPT-4 shows impressive results, being able to build captions for multiple contexts 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. However, just 2,700 caption pairs were generated for the running experiments due to the resource limitations. For performing large scale experiments, at least 100,000 needs to be generated and computed a measurement of text diversity over them.

B. Image generation

Based on the prompts generated by the large language model, the pipeline formed from 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 and generating 50 pairs of images per prompt gives high quality data in relative shorter time (see Figure 1) compared to time taken using the hyperparameter values from InstructPix2Pix [1] data generation. Furthermore, the generated data exposes the limitations of the InstructPix2Pix model in generalizing for the interior-design case and its poor performance on it (see Figure 4).

C. Alternatives explored

1) GPT-3.5:

For generating the pairs of captions GPT-3.5 was also tried, but the results were less impressive than the newer one. It aims to build descriptions from multiple contexts, but they have less diversity than the ones generated by GPT-4 (see Figure II). Furthermore, the duration of the process was a little faster due the complexity of GPT-4.

Fig. 3: Stable Diffusion introduces noise in the generated images

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

Images generated based on descriptions generated by GPT-4: (a) A rustic living room with a stone fireplace, leather sofas, a wooden coffee table, and a bear skin rug on the floor. (b) A bright living room with clean lines, a blue sofa, a geometric rug, and a large potted plant in the corner. (c) A rustic living room with a stone fireplace, leather armchairs, and a pine coffee table with a bowl of pinecones as a centerpiece.

2) Llama2 13B:

The option of fine-tuning the last released Llama2 [12] open-source model was explored as well. For 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 useful results. The data used for fine-tuning can be found **here**.

Fig. 4: Comparison between the generated data and InstructPix2Pix performance on it

VI. Conclusions and Future Work

A. Caption generation

The proposed approach shows impressive results in data generation for the prompt-based images editing with no prior data. By just leveraging the power of LLMs, using a text-to-image model and applying CLIP filtering we are able to generate data with no limit in quantity. However, due to the lack and limitations of context there are still problems in terms of diversity in image setting which is a direct consequence of the generated descriptions. Considering the fast evolution of the large language model these limitations should slowly disappear, but there is still a lot to wait till then. For now, there are still other models to explore for generating the text descriptions like Llama2-70B, Mistral 7B [6] and the newly released **Mistral 8x7B**. Additionally, breaking the big task into smaller subtasks and utilizing in-context learning for creating task-specific LLMs can improve the diversity in the generated text as presented in [10].

B. Image generation and filtering

The generation of the before and after images with respect to the edit instruction works similar to the one presented in [1], but the reduced number of samples created per caption pair impacted the amount of images present in the result dataset. Furthermore, Stable Diffusion fails of generating correct images from text descriptions (see Figure 3), so other experiments can be run for various other text-to-image models like Imagen [8] and Muse [3] and. Additionally, the CLIP filtering is not enough for assuring that the difference between the initial and the output images lies just in the one specified in the given instruction. For improving the filtering, having the subtasks and the task-specific LLMs defined as presented in [10], there can be added a task predictor for reassigning samples with instructions to the correct task in case of an error and a detector to validate that the presence, absence or replacement of an object in the output image for every type of task, respectively.

C. InstructPix2Pix's performance

The small amount of data and the resource limitations did not make possible the fine-tuning of the InstructPix2Pix on the generated data and running analysis on it. However, having powerful resources at hand is preferable and will create the opportunity to validate the quality of the data generated with the proposed approach. For testing the actual advantages of the presented approach at least 5,000 samples need to be used for test data. Additionally, for fully validating the advantages of the presented approach, a comparison between the base model and the one fine-tuned with the generated data is necessary and can be more on multiple metrics like CLIP (with its variances CLIP_{im}, CLIP_{dir}, CLIP_{out}) and DINO [2].

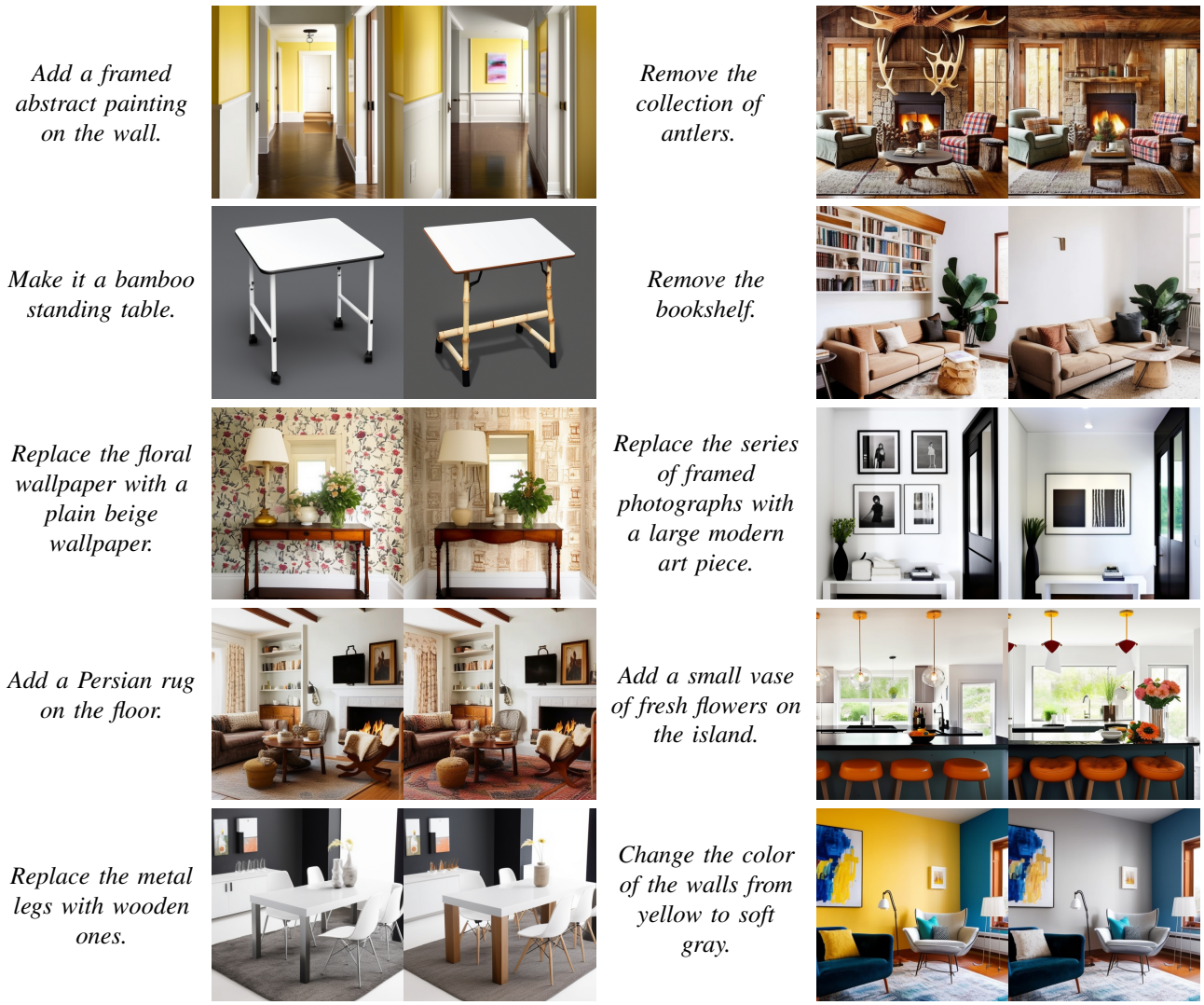

Fig. 5: Additional generated trainig samples for the interior-design prompt based image editing task.

REFERENCES

[1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros, *Instructpix2pix: Learning to follow image editing instructions*, IEEE/CVF conference on computer vision and pattern recognition, CVPR 2023, vancouver, bc, canada, june 17-24, 2023, 2023, pp. 18392–18402.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, *Emerging properties in self-supervised vision transformers*, 2021 IEEE/CVF international conference on computer vision, ICCV 2021, montreal, qc, canada, october 10-17, 2021, 2021, pp. 9630–9640.

[3] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan, *Muse: Text-to-image generation via masked generative transformers*, International conference on machine learning, ICML 2023, 23-29 july 2023, honolulu, hawaii, USA, 2023, pp. 4055–4075.

[4] Tsu-Jui Fu, Xin Wang, Scott T. Grafton, Miguel P. Eckstein, and William Yang Wang, *SSCR: iterative language-based image editing via self-supervised counterfactual reasoning*, Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, november 16-20, 2020, 2020, pp. 4413–4422.

[5] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or, *Stylegan-nada: Clip-guided domain adaptation of image generators*, ACM Trans. Graph. **41** (2022), no. 4, 141:1–141:13.

[6] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed, *Mistral 7b*, CoRR **abs/2310.06825** (2023), available at 2310.06825.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, *High-resolution image synthesis with latent diffusion models*, IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, new orleans, la, usa, june 18-24, 2022, 2022, pp. 10674–10685.

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi, *Photorealistic text-to-image diffusion models with deep language understanding*, Neurips, 2022.

[9] 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, *LAION-5B: an open large-scale dataset for training next generation image-text models*, Neurips, 2022.

[10] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman, *Emu Edit: Precise Image Editing via Recognition and Generation Tasks*, CoRR **abs/2311.10089** (2023), 1–10, available at 2311.10089.

[11] Ivona Tautkute, Aleksandra Mozejko, Wojciech Stokowiec, Tomasz Trzcinski, Lukasz Brocki, and Krzysztof Marasek, *What looks good with my sofa: Multimodal search engine for interior design*, CoRR **abs/1707.06907** (2017), available at 1707.06907.

[12] 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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom, *Llama 2: Open foundation and fine-tuned chat models*, CoRR **abs/2307.09288** (2023), available at 2307.09288.

[13] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio, *Show, attend and tell: Neural image caption generation with visual attention*, Proceedings of the 32nd international conference on machine learning, ICML 2015, lille, france, 6-11 july 2015, 2015, pp. 2048–2057.