STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXIX**, Number 2, 2024
DOI: 10.24193/subbi.2024.2.02

LANGDES: A NEW APPROACH FOR IMPROVING THE PERFORMANCE OF PROMPT-BASED IMAGE EDITING IN INTERIOR DESIGN SETTING

VICTOR-EUGEN ZARZU

Abstract. The topic of instruction-based image editing has gotten a lot of attention in recent years with a lot of research conducted due to its immense potential in various applications such as removing unwanted details present in existing images or improving them. However, one of the main problems in addressing this problem is acquiring a dataset for model training. Several methods and variations were proposed, but all of them rely on already-existent data. We propose a method to address this problem by creating a context-specific dataset for interior design with no previously available information by leveraging the knowledge of large language models (LLM). Furthermore, we test and prove the efficiency of the generated dataset on InstructPix2Pix which starts to compute better results for the interior-design setting after the fine-tuning. Moreover, we propose an alternative solution for enhancing the localization of the edit region through cross-attention map regularization based on a text-based segmentation mask.

1. Introduction

Prompt-based image editing is the problem of modifying an input image concerning a natural language edit prompt. Applications of this task consist of reducing the effort in professional image editing (e.g. removing a person from an image will transform into writing a phrase) and increasing the efficiency in graphics.

Received by the editors: 11 October 2024.

2010 *Mathematics Subject Classification.* 68T05, 68T45.

1998 *CR Categories and Descriptors.* I.2.6 [**Learning**]: Subtopic – *Connectionism and neural nets*; I.2.10 [**Vision and Scene Understanding**]: Subtopic – *3D/stereo scene analysis*.

Key words and phrases. Diffusion models, Prompt-based image editing, Deep learning, Attention, Data generation.

The major challenge of this problem is generating a dataset for training and evaluation. While each instance in the dataset needs to consist of an input image, an edit prompt, and the image resulting from the edit of the original one to the prompt, it is challenging to create such a dataset at scale because of the costs involved. Furthermore, reducing the unwanted modifications in the background and objects is also an important part of the problem, being intensely researched and correlated with the noise in the training dataset and the incapability of the model to map the edit instruction to the correct objects in the image.

The main focus of this paper is distributed among three aspects. The first part tries to answer whether a robust, high-quality, and context-specific dataset for the task in a discussion can be generated without previously available data. Secondly, it is proved that this dataset is qualitative and improves the IP2P model in the chosen context (interior design). Lastly, we aim to answer if a referring expression-based image segmentation with the object(s) under edit improves the performance of the InstructPix2Pix model in the general case through cross-attention map regularization.

In summary, we aim to answer the following research questions.

RQ1. *How to generate data for context-specific prompt-based image editing tasks with no previously known data?*

RQ2. *Does the generated data improve the performance of instruction-based image editing in the specific context?*

RQ3. *Does a referring expression-based image segmentation with the object under edits improve the performance of the InstructPix2Pix model in the general case through cross-attention map regularization?*

The rest of the article is structured as follows. Section 2 presents the previous work in the area of prompt-based image editing as well as text-based image segmentation. Afterward, the methodology that aims to respond to the addressed research questions is presented in Section 3. Along with discussions, we showcase the experimental results of the proposed approaches in Section 4, while the article ends with the conclusions and directions for future work in Section 5.

2. Related work

2.1. **Prompt-to-Prompt.** Introduced by Hertz et al. [11], Prompt-to-Prompt is an approach for generating two similar images based on two given prompts based on diffusion models. The method relies on the fact that the geometry and spatial layout of any generated image using text-guided diffusion models depend on the cross-attention maps. The approach generates the two images

simultaneously, but while the generation of the first one is normal, in the second one, the cross-attention maps from the first generation are injected. This integration lasts for $\mathcal{T}$ timesteps and controls how similar the resulting images should be, its value variation depending on the area that has to be edited. However, while it tackles the problem of image editing, it is not able to edit an already-existent image, but it opens the possibility of generating a dataset for the prompt-based image editing task for supervised training.

2.2. **InstructPix2Pix.** IntructPix2Pix (IP2P) is an approach introduced by Brooks et al. [3] for editing an already-existent image based on a given prompt. It relies on using a Stable Diffusion [21] checkpoint, incorporating also the input image as conditioning and applying classifier-free guidance [12]. The guidance is done based on the initial image conditioning c_I and edit prompt conditioning c_T. Furthermore, two guidance scales are introduced, one for the image s_I and one for the text s_T, which can be adjusted to trade off the importance of each conditioning in the generated sample. The training is done in a supervised fashion on a dataset generated in two steps by leveraging the knowledge of GPT-3 [4] and Stable Diffusion (SD) [21] combined with the Prompt-to-Prompt approach. In the first step, GPT-3 is fine-tuned, given an initial caption from the LAION-Aesthetics V2 6.5+ [23], to output an edit prompt and the caption edited following this prompt. Secondly, based on the initial and edited caption, Prompt-to-Prompt is applied to generate 100 pairs of images followed by filtering using CLIP [20] to keep the most consistent pairs.

2.3. **Emu Edit.** Introduced by Sheynin et al. [24], Emu Edit addresses the issue of inaccurately interpreting and executing the edit instructions of InstructPix2Pix by being trained on a variety set of problems including classic computer vision and image editing tasks. The model architecture and training are similar, but the diffusion model used is Emu [6], and learned task embeddings are injected into the U-Net architecture to enhance the accuracy of the edit application. The data generation pipeline is also based on an LLM and Prompt-to-Prompt but followed by a more comprehensive filtering approach to reduce the noise and increase the consistency of the data. To generate the textual data, the approach proposes creating via in-context learning task-specific Llama2-70B [27] agents that, given an initial caption, are prompted to return an edit instruction and the final edited caption along with a list of the objects that are edited. The initial Prompt-to-Prompt solution is enhanced to reduce the noise of the original method by binary injecting the masks of the edited objects from the initial image during the editing process to increase image consistency.

2.4. **LIME: Localized Image Editing via Attention Regularization in Diffusion Models.** LIME is a solution for reducing the unwanted modification in the edited image through cross-attention map regularization proposed by Enis et al. [25]. The approach relies on the property that diffusion models can be used for text-based image segmentation tasks by leveraging their intermediate features as introduced by PNVR et al. [19]. With this, the method extracts the feature from various layers of the IP2P's U-Net architecture, followed by their fusing in three steps. Afterward, having a final attention map, the Region of Interest (RoI) is computed based on the top 100 pixels in probability followed by introducing all segments that overlap at least one of these pixels as well. Having the binary mask M computed, the method regularizes the attention scores (QK^T) within the RoI of the unrelated tokens to the edit denoted as S (e.g. <*start of text*>, *stop words*)) as in Formula (1), where α is a large value.

$$R(QK^T, M) = \begin{cases} QK^T_{ijt} - \alpha, & \text{if } M_{ij} = 1 \text{ and } t \in S \\ QK^T_{ijt}, & \text{otherwise} \end{cases} \tag{1}$$

2.5. **GRES.** Introduced by Liu et al. [14], **Generalized Referring Expression Segmentation** (GRES) is a new benchmark that addresses the limitations of the original task by allowing the segmentation of multiple objects within the same image and returning an empty mask if the referred object is not present. They also propose a model called ReLA, achieving state-of-the-art performance in the new GRES benchmark and the original RES one.

3. Methodology

Prompt-based image editing is the problem of computing the image resulting from the editing of an original image based on a given instruction. Existent methods deal with this problem in the general case with no special focus on the interior design setting, and this paper aims to improve the performance in such a setting.

This section introduces the proposed methodology for improving the text-based edit in the interior design setting as well as the edit localization based on a text-based segmentation mask through cross-attention map regularization, approaches that will facilitate the answering to the addressed research question. Firstly, we propose a pipeline for generating context-specific datasets with no previous data, focusing on the text and images in two different sections. Afterward, to show the efficiency of the created dataset, we fine-tune

the base InstructPix2Pix model on it. Lastly, we propose a method for improving the edit localization based on ReLA's text-based segmentation mask through cross-attention map regularization.

3.1. **Dataset generation.** 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. As stated before, the currently available proposal for acquiring a large dataset for this task was the generation of it. Still, all of them rely on an initial description of the original image. Of course, there is still the option of manually creating such a dataset, but it involves high costs and it is not scalable for large and high-performance models. For a context-specific case like interior design, this data is limited and not diverse enough for a good generalization of the problem.

So, we propose a method for creating such a dataset with no previous data to respond to research question RQ1 by creating a context-specific dataset in this fashion. The proposed approach is composed of the following two steps similar to previous approaches: the textual data generation followed by the creation of the paired images based on the text. Since the instruction-based image editing is treated as a supervised learning problem, each dataset instance will be made of an edit prompt and two images, the original one and the image modified concerning the given prompt.

3.1.1. *Text editing instructions generation.* The initial generated text samples will mimic the structure of the final dataset's instance, but the images are replaced with their textual descriptions. As such, each text instance will be made of three elements: (i) the description of the initial room or object, (ii) the edit instruction, and (iii) 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 [3] where the last 2 components of the tuple are generated based on a previously known description. By leveraging the knowledge of the language model, there is no need to fine-tune it.

Knowing that different types of rooms usually have different objects that characterize them, room-specific agents can be created via in-context learning. With the proposed approach, by just presenting to the language model the format of the desired output (here JSON) and 3 other examples in that format, it is able to generate a large amount of data in the desired format with a great variety of responses. Moreover, to reduce the noise, GPT-4 [17] was instructed to clearly state that the object is missing or exists in the original description for the add and remove actions respectively. With this approach, 8,831 text samples were generated and published on HuggingFace [29]. Intuitively, this

method can be used for any specific context by just providing the language model with examples from the targeted setting.

Additionally, the presented method has the advantage of enabling the creation of data in a hierarchical way of difficulty for the editing model: it first creates paired captions for single objects captions followed by the ones with a description of rooms with more objects. Compared to [3], the presented method can be extended and used for any other special case of prompt-based image editing, without the prior need for data, hence independence on the existent datasets.

Moreover, for the generation of the text captions and instruction GPT-3 was also used as an alternative before GPT-4 became publically available. While both of them produced overwhelming and diverse results, GPT-4 was more diverse in its outputs by different criteria computed using the site introduced by Runker [1] as seen in Table 1.

	MTLD [15] ↑	Dugast's U^2 [8] ↑	Guiraud's Index [7] ↑	Yule's K [9] ↓
GPT-3.5	28.13	12.83	3.87	356.49
GPT-4	**32.72**	**13.73**	**4.52**	**278.80**

TABLE 1. Comparison of diversity in the textual data generated by GPT models.

3.1.2. *Generating images from the paired editing instructions.* Starting from the paired editing instructions generated with the previous method, the Prompt-to-Prompt based on the Stable Diffusion model approach is used for generating the dataset samples in a supervised way: the image before and after the edit. However, generating one image for each instruction does not guarantee their consistency. For addressing this issue, similar to the approach presented in [3], a large number of image pairs is generated for each pair of captions with different values of p that controls the similarity between the images, followed by CLIP-based metric filtering introduced by Gal et al. [10]. Only the top four pairs of images that are above the image similarity threshold of 0.75 are kept.

Compared to the values used for InstructPix2Pix, where for every pair of captions 100 image pairs are generated before filtering, the proposed approach splits the generation into two parts to reduce the time of generation: for the images with single objects, 30 image pairs are generated, while for rooms with multiple objects, 50 pairs are generated. The choice for fewer pairs for single object images comes from the idea that the fewer the number of objects in the image, the less diversity in the images of the same pair will be present in the

Prompt-to-Prompt generation. With this approach, 4,259 train samples and 1,129 test samples were generated. The training and testing datasets can be accessed on HuggingFace [34, 33]. Furthermore, the generated data, not only that is visually appealing and diverse, but it also exposes the limitations of the InstructPix2Pix model in generalizing for the interior-design case and its poor performance as shown in Figure 1.

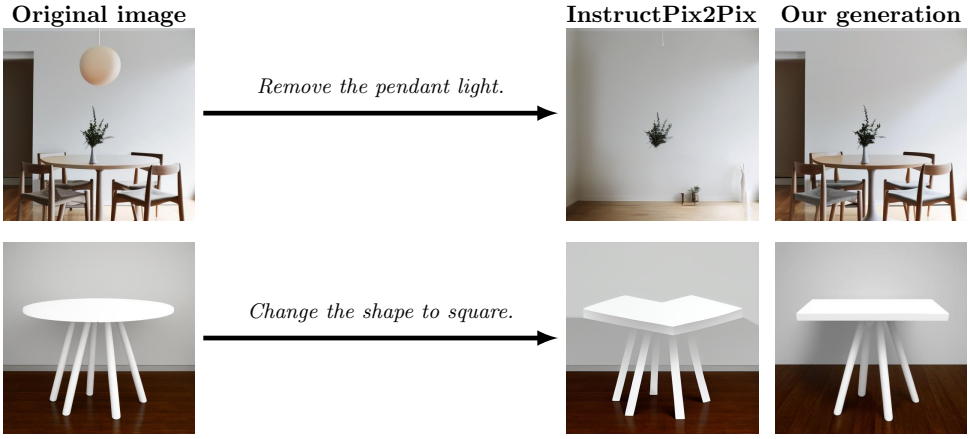


FIGURE 1. Comparison between the generated data and InstructPix2Pix's performance on it.

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). Intuitively, this will also enforce the model to correctly identify the Region of Interest for the edit and to learn that in some cases the given prompt can be misleading, a problem that was not addressed in the previous approaches. This additional dataset can be found at [29], and its effects will be studied in the following sections. Moreover, these types of prompts with no effect on the image were also introduced in the initial test set.

3.2. **Fine-tuning InstructPix2Pix on generated dataset.** Having the previously generated data, we investigate its benefits when used to fine-tune InstructPix2Pix in order to answer research question RQ2. However, due to resource limitations, the training setup was modified to satisfy the computing capabilities. For this, the training was run in float16 precision, with a batch of only two images, and the images were resized from an initial dimension of 512x512 to 256x256. Nonetheless, even with these restrictions, the overall training was not affected drastically, and the results are promising. Using the

training data generated with the method presented in Section 3.1, InstructPix2Pix was fine-tuned on 300 epochs with a learning rate of 10^{-5}.

3.3. **Enhancing Region of Interest detection using a Referring Expression Segmentation model.** This section presents an alternative approach to the one introduced by Enis et al. [25] by leveraging the overwhelming performance of the recent text-based segmentation model, ReLA. This section aims to respond to research question RQ3 and to explore if such a method is improving the edit application in a general setting.

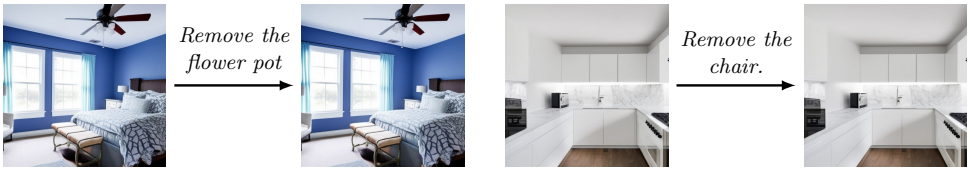


FIGURE 2. Data augmentation with samples containing no change in the output image

The proposed approach differs from the one introduced in LIME, just by how the segmentation map of the region(s) under edit is computed, here using ReLA. The usage of ReLA enhances the edit localization by being state-of-the-art in this task, and, additionally, it allows context-dependent references like *"The right blue chair."*.

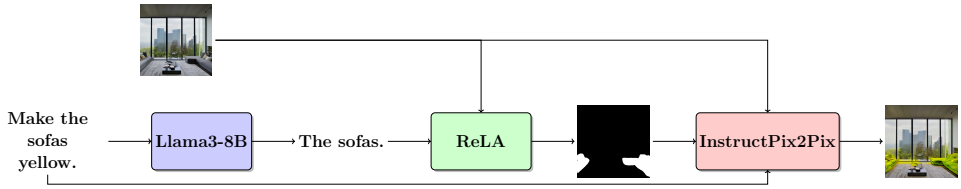


FIGURE 3. The pipeline for computing the edit through cross-attention regularization using ReLA's segmentation mask.

In order to be able to use ReLA in the editing pipeline, we propose the solution showcased in Figure 3. To use the text-based segmentation model, we first need to create a reference to the object(s) to be edited out of the initial edit prompt. This is achieved by creating an LLM agent via in-context learning by injecting the task description and a couple of examples in the model's system prompt as highlighted in Figure 4. Here, we use the 8B version of the most recent Llama3 [16] model. Afterward, we compute the text-based segmentation map determined by ReLA and feed it along with the initial image and

the edit prompt in the modified version of InstructPix2Pix for cross-attention map regularization. We use the same approach of negatively regularizing the unrelated tokens to edit (e.g. padding tokens, <*start of text*>, etc.) which also offers the model more freedom in the edit application compared to the positive regularization of the related tokens.

```
system_message = """
You are a bot that needs to take the reference of the text under edit
    from an edit prompt or the object that affects it. Here are some
    examples
    'Replace the top book on the desk.' would transform into the
        following reference 'The top book on the desk.'
    'Add a plate on the wooden table.' would transform into the
        following reference 'The wooden desk.'
Please return just the transformed text as the reference and nothing
    more.
"""

messages = [
    {"role": "system", "content": system_message},
    {"role": "user", "content": edit_prompt},
]
```

FIGURE 4. The prompt used for extracting the object reference from edit prompt via in-context learning with Llama3-8B.

4. Results and discussion

This section is focused on presenting the experimental results of the presented approach along with the discussions that emerged from the visuals and analysis on the metrics.

4.1. **Results.** This section is focused on presenting the experimental results of the proposed methodology for improving the edits in the interior-design context and enhancing the edit localization through cross-attention map regularization.

4.1.1. *Fine-tuning InstructPix2Pix on generated dataset.* As shown in Table 2, the proposed approach improves the performance of the model considerably across different metrics computed as the cosine similarity between the features extracted using CLIP [20] and DINOv2 [18]. The various CLIP metrics presented in the table compute different types of similarities consisting of the similarity between the input and output images (CLIP_{im}), the similarity

between the edited image and its textual description (CLIP_{out}), and the similarity between the changes in the captions and the images (CLIP_{dir}), while DINO only computes the similarity between the initial and edited image. The model resulting from this experiment is publicly available on HuggingFace [30] and can be freely used for image editing.

	$\text{CLIP}_{\text{im}} \uparrow$	$\text{CLIP}_{\text{dir}} \uparrow$	$\text{CLIP}_{\text{out}} \uparrow$	DINO $\uparrow$
IP2P	84.25	0.025	26.16	87.67
IP2P-FT	**92.21**	**0.063**	**29.17**	**94.54**

TABLE 2. Comparisons between the metrics of the base InstructPix2Pix model and the fine-tuned one on the test set.

4.1.2. *Additional fine-tuning on the dataset with unchanged images.* After fine-tuning the model on the train data, an experiment of fine-tuning the model on the dataset with unchanged images was conducted. Doing this for more epochs results in a model that does not apply any modifications to the image, but fine-tuning for just one epoch does not alter the performance completely. Unfortunately, in most cases, the model still learns just not to change the original image at all, ignoring the edit instruction and underperforming in most of the cases. However, in some cases, even though their number is small, the output of this model is better than the previous one, this model also being publicly available [31].

4.1.3. *Enhancing Region of Interest detection using a Referring Expression Segmentation model.* The experiments conducted to incorporate ReLA's segmentation masks into the editing process via cross-attention map regularization did not show positive results up to this point. However, as seen in Figure 5, it enhances the localization of the edit region by forcing the model not to modify the unrelated background or objects. Nonetheless, the application of the edit is not done correctly in most cases by showing colors and shapes that are mostly random and off the edit prompt.

4.2. **Discussion.** The conducted experiments show improvements in instruction-based editing for interior design images which can be extrapolated to any context-specific case. However, there are a lot of observed particularities that occur during the analysis of the experimental results and this section aims to present them.

As shown and stated in Table 2, fine-tuning the base InstructPix2Pix model on the generated data with interior-design samples improves the model's ability to work in such an environment as shown in Figure 6. It can be seen that

the dataset not only offers the ground truth for the output image but also a lot of knowledge that is assimilated by the model through supervised learning.

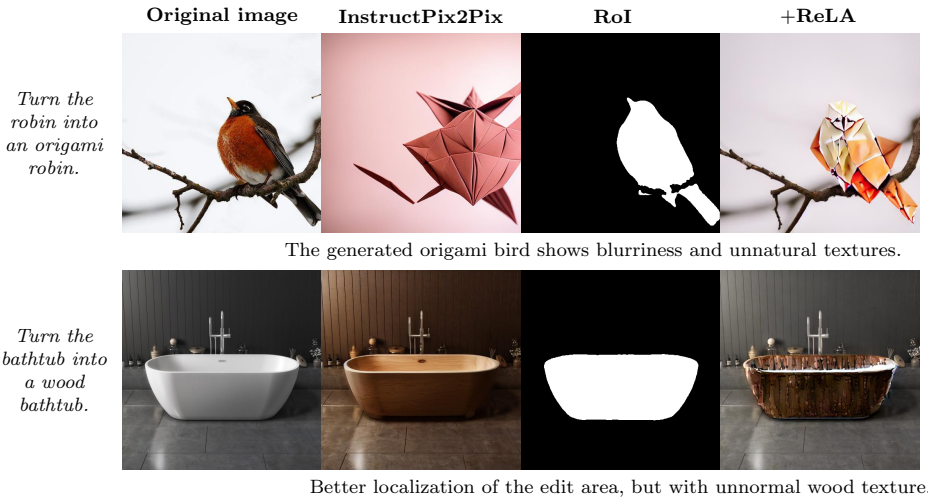


FIGURE 5. Examples of the images edited after integrating ReLA's segmentation mask for cross-attention map regularization.

Even though the new model's edits are more qualitative, there are still problems with the editing of unwanted parts like background or objects that are not referred to in the edit prompt. This can be seen in Figure 6 where, in the first image, one flower from the table disappears, the table color is changed from light grey to a slightly darker tone, and the lamp disappears. Furthermore, in the second image, the table top is changed correctly, but the color of the floor becomes more cherry.

4.2.1. *Dataset generation.* To eliminate such cases and better improve the performance, a more qualitative dataset needs to be created. First of all, the current dataset is not very diverse in the context and words used which is due to the way the prompts are generated and its reduced volume. Moreover, generating the initial captions can also be done by starting from interior design images available on the Internet and using an image-to-text model that describes the given image. After this step, the same idea introduced by Brooks et al. [3] can be applied by fine-tuning an LLM for the generation of caption pairs. However, this method has a limitation given by the number of publically available images for the context under discussion. This affects the scalability of the method for various contexts and the accuracy as the image-to-text model would also introduce noise to the dataset. For example, for the

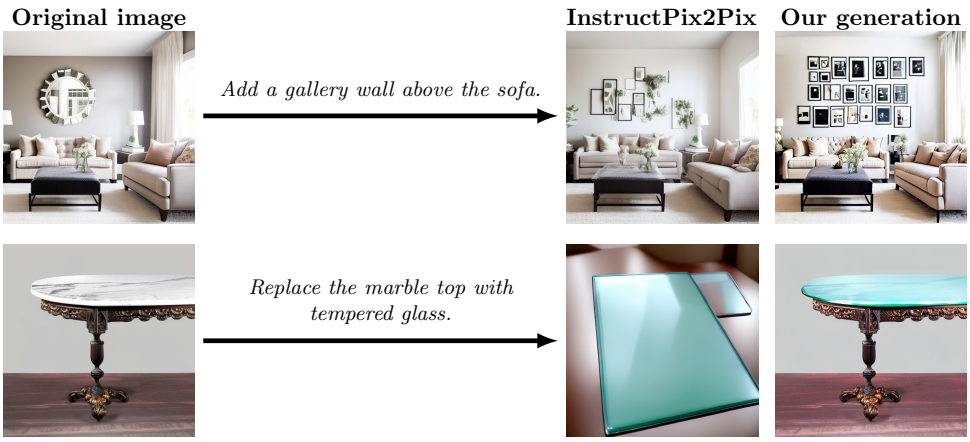


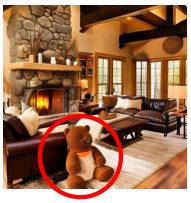
FIGURE 6. The new model's edits are a lot more qualitative than the InstructPix2Pix's ones for interior design.

interior design case, one such dataset is the Interior Design IKEA dataset [26] which has only 6,000 images.

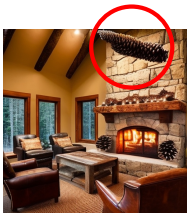
On the other hand, the introduced noise also comes from the level of image generation but in two different aspects. Firstly, the diffusion model used, here Stable Diffusion (SD), does not have a good understanding of interior design scenes as shown in Figure 7 and also fails to correctly follow the given prompt. Furthermore, there are also cases when it introduces more objects than presented in the textual description. Secondly, the noise is also introduced during the generation with Prompt-to-Prompt followed by CLIP filtering because, in some cases, the images do not differ only by the resulting actions expressed in the edit prompt as seen in Figure 8.

Comparisons to related work are limited due to the recent publishing date of Emu Edit [24] and LIME [25]. However, as stated before, all the solutions addressed in Section 2 require at least a large volume of initial captions that are used to build the dataset for supervised training. Compared to these, we propose and show the effectiveness of a new approach to generate a context-specific dataset for this task with no previously available information. Hence, in contrast to previous work in this area, the presented approach increases the scalability and the amount of data that can be generated by not relying on any available information.

Furthermore, compared to the approach introduced by Enis et al. [25], using ReLA for computing the Region of Interest applies a regularization effect only if the referred object is present in the input image. However, the LIME approach gives a more general approach by computing the mask for all-purpose



A rustic living room with a stone fireplace, leather sofas, a wooden coffee table, and a bear skin rug on the floor.



A rustic living room with a stone fireplace, leather armchairs, and a pine coffee table with a bowl of pinecones as a centerpiece.

FIGURE 7. Generated images that show the limited knowledge of Stable Diffusion in interior design.

tasks, even for creating masks with initially non-existent objects as minimally, but not explicitly expressed in the paper in just one example. Nonetheless, using ReLA in the Remove and Replace tasks would offer a greater benefit due to its better performance in text-based segmentation tasks, but with the overhead of using an additional language model for extracting the region reference out of the edit prompt. Furthermore, this implies a growth in the time needed to compute the edit because, in LIME, the segmentation map is computed using the internal features already existent in InstructPix2Pix, while we propose a method that uses two additional networks.

Remove the floor-to-ceiling windows and replace them with a large artwork.

Change the glass top to a wooden top.

FIGURE 8. Examples of generated samples that do not correctly follow the edit instruction.

5. Conclusions and Future Work

This paper introduced `LangDes`, a new approach for improving the performance of the instruction-based image editing task in the interior design setting. Afterward, we proposed a promising approach for improving the future performance on the instruction-based image editing task, followed by experimental results and the associated discussions in Section 4.2.

The conducted experiments in the interior design setting showed overwhelming results and confirmed positive answers to the research questions RQ1-RQ3 formulated in Section 1. So, we proved that the proposed method for generating context-specific with no previous data stays valid, and we showed its efficiency in improving the InstructPix2Pix performance in the interior-design context. Afterward, we experimented with the integration of the text-based segmentation model ReLA in the editing pipeline to improve the edit localization. However, as an answer to research question RQ3, the current experimental results only prove the localization improvement, but the application of the edit is still not under control, with the model returning images with random textures within the RoI.

One first direction for future work would be to increase the quality and diversity of the generated dataset. To increase the latter, the initial textual data needs to be more diverse. A solution for this would be to combine the output of multiple LLMs such as Llama2-70B [27], Gemini [2] or Mistral 8x7B [13]. On the other hand, task-room-specific agents can be created using in-context learning, but with the disadvantage of a large number of possible combinations that will increase the time required for the conducted experiments. Furthermore, to increase the quality and consistency of the image pairs, different text-to-image models like Imagen [22], Muse [5] or an interior design fine-tuned version of Stable Diffusion open-sourced at [32] can be used, as well as applying a more comprehensive filtering pipeline as the one presented in Emu Edit [24].

Another interesting topic is the complexity of the edit prompt. Even though the presented work focuses on prompts with single edits of single objects, a possible direction for experiments can be targeting more complex instructions such as the ones involving multiple actions. This can be both seen as extra evaluation of the model resulted from in presented work, as well as an extension at the level of data generation for creating such samples.

As a last future work direction, we may refer to *enhancing region of interest detection* using various RES models, along with investigating the cause of incorrectly applying the edit despite correctly localizing the targeted area. To improve the editing of parts of the objects, a combination of GRES and Multi-Granularity Referring Expression Segmentation (MRES) [28] can be

used. Compared to, GRES, MRES, introduced by Wang et al. [28], supports expressions for segmenting part-level regions of the target objects within a model called UniRES, but with no support for a good performance with multiple objects at the same time. Having these two models, an additional decisional network for detecting which type of segmentation model should be used for computing the mask before the edit.

References

[1] Alex Reuneker. Lexical Diversity Measurements. `https://www.reuneker.nl/files/ld/`, 2017. Accessed: 2024-01-15.

[2] Rohan Anil, Sebastian Borgeaud, and et al. Gemini: A Family of Highly Capable Multimodal Models. *CoRR*, abs/2312.11805, 2023.

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Canada, 2023*, pages 18392–18402. IEEE, 2023.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Subbiah, and et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[5] Huiwen Chang, Han Zhang, and et al. Muse: Text-To-Image Generation via Masked Generative Transformers. In *International Conference on Machine Learning, ICML 2023, Honolulu, Hawaii, USA*, volume 202, pages 4055–4075. PMLR, 2023.

[6] Xiaoliang Dai, Ji Hou, and et al. Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack. *CoRR*, abs/2309.15807, 2023.

[7] Michael Daller. Guiraud's index. 2010.

[8] Daniel Dugast. *La Statistique Lexicale*. SLATKINE, 1980.

[9] G. Udny Yule. The Statistical Study of Literary Vocabulary. *Cambridge University Press*, 1944.

[10] 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(4):141:1–141:13, 2022.

[11] Amir Hertz, Ron Mokady, and et al. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations, 2023*. OpenReview.net, 2023.

[12] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598:1–14, 2022.

[13] Albert Q. Jiang, Alexandre Sablayrolles, and et al. Mixtral of Experts. *CoRR*, abs/2401.04088, 2024.

[14] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: generalized referring expression segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23592–23601. IEEE, 2023.

[15] Philip M. McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392, 2010.

[16] Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date. `https://ai.meta.com/blog/meta-llama-3/`, 2024. Accessed: 2024-05-10.

[17] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[18] Maxime Oquab, Timothée Darcet, and et al. DINOv2: Learning Robust Visual Features without Supervision. *CoRR*, abs/2304.07193, 2023.

[19] Koutilya PNVR, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. Ld-znet: A latent diffusion approach for text-based image segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4134–4145. IEEE, 2023.

[20] Alec Radford, Jong Wook Kim, and et al. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021,*, volume 139, pages 8748–8763. PMLR, 2021.

[21] Robin Rombach, Andreas Blattmann, and et al. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022*, pages 10674–10685. IEEE, 2022.

[22] Chitwan Saharia, William Chan, Saxena, and et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc., 2022.

[23] Christoph Schuhmann, Romain Beaumont, and et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022.

[24] Shelly Sheynin, Adam Polyak, and et al. Emu Edit: Precise Image Editing via Recognition and Generation Tasks. *CoRR*, abs/2311.10089, 2023.

[25] Enis Simsar, Alessio Tonioni, Yongqin Xian, Thomas Hofmann, and Federico Tombari. LIME: localized image editing via attention regularization in diffusion models. *CoRR*, abs/2312.09256, 2023.

[26] Ivona Tautkute, Aleksandra Mozejko, and et al. What Looks Good with my Sofa: Multimodal Search Engine for Interior Design. *CoRR*, abs/1707.06907, 2017.

[27] Hugo Touvron, Louis Martin, Kevin Stone, and et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288, 2023.

[28] Wenxuan Wang, Tongtian Yue, and et al. Unveiling Parts Beyond Objects: Towards Finer-Granularity Referring Expression Segmentation. *CoRR*, abs/2312.08007, 2023.

[29] Victor-Eugen Zarzu. Dataset for interior design. `https://huggingface.co/datasets/victorzarzu/interior-design-edit-captions`, 2024.

[30] Victor-Eugen Zarzu. Fine-tuned InstructPix2Pix model. `https://huggingface.co/victorzarzu/ip2p-interior-design-ft`, 2024.

[31] Victor-Eugen Zarzu. Fine-tuned InstructPix2Pix model on the dataset with unchanged images. `https://huggingface.co/victorzarzu/ip2p-interior-design-ft-unchanged-one-epoch`, 2024.

[32] Victor-Eugen Zarzu. Interior design fine-tuned version of Stable Diffusion. `https://huggingface.co/stablediffusionapi/interiordesignsuperm`, 2024.

[33] Victor-Eugen Zarzu. Testing data. `https://huggingface.co/datasets/victorzarzu/interior-design-prompt-editing-dataset-test`, 2024.

[34] Victor-Eugen Zarzu. Training data. `https://huggingface.co/datasets/victorzarzu/interior-design-prompt-editing-dataset-train`, 2024.

Babeș-Bolyai University, Faculty of Mathematics and Computer Science, 1 Mihail Kogălniceanu, Cluj-Napoca 400084, Romania

Email address: `victor.zarzu@stud.ubbcluj.ro`