Prompt-based image editing for interior-design images

Victor Zarzu
Babeș-Bolyai University

WeADL 2024 Workshop

The workshop is organized by the Machine Learning research group (www.cs.ubbcluj.ro/ml) and the Romanian Meteorological Administration (https://www.meteoromania.ro/)

Machine Learning Research Group
MLyRE

METEO ROMANIA

Prompt-based image editing (1)

Task description

Given an original image and a text prompt describing the desired edit, the task is to generate a new image that reflects the edit applied to the original one.

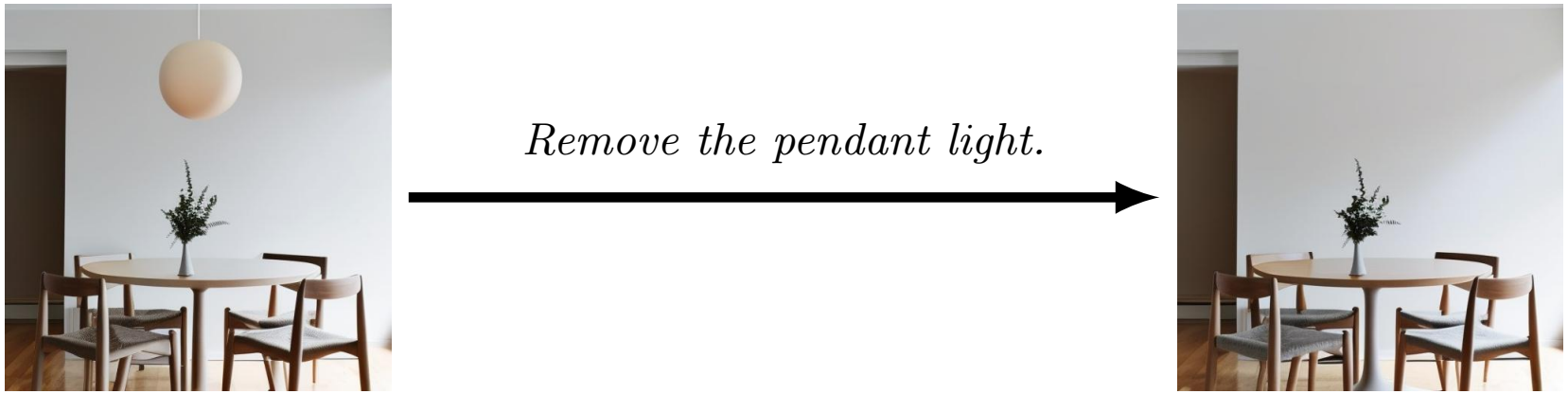


Figure: Example of an instruction-based image edit in the interior design setting.

Prompt-based image editing (2)

Problem Importance

- Various possible edit versions for the same prompt.
- Fast high-quality image edits with little human effort.
- Reduce the average edit time of an image.

- **Applications**
  - Increase the quality of old images.
  - Remove unwanted details of already taken photos.
  - Automate and ease the image edit process and increase the number of users that can accomplish it.

Background - Diffusion models (1)

- Generative models that can produce highly qualitative images/videos by using a multistep framework that removes noise at each timestep.
- The training consists of two parts that runs over T timesteps as depicted in Figure (2)
  - The forward or diffusion process.
  - Reverse diffusion process.

Background - Diffusion models (2)

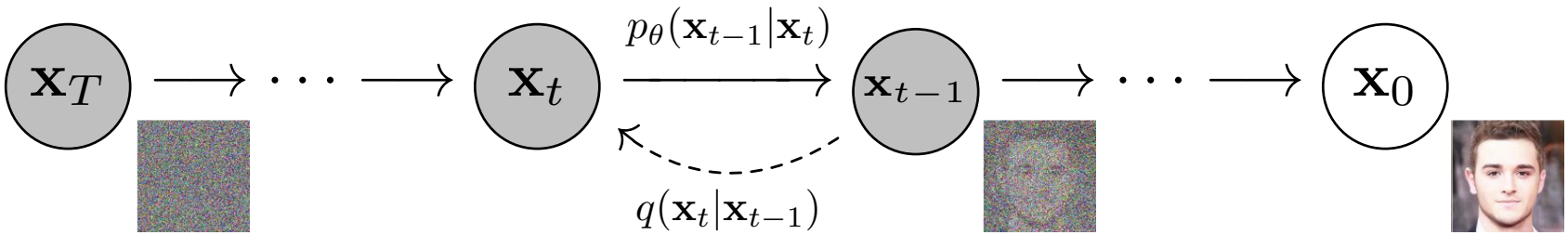


Figure: The diffusion and reverse diffusion processes. Source: [HJA20]

Background - Diffusion models (3)

- They use a U-Net architecture.
- The information passed is extended with knowledge about
  - The current timestep through a positional sinusoidal encoding.
  - The prompt through an encoding that can be computed using CLIP for example.

Background - Latent diffusion models

- Diffusion models where the whole diffusion process is done in the latent space of an autoencoder (usually a Variational one).
- **Advantages**
  - High-resolution results.
  - Preserve details.
  - Computationally efficient.

Background - Classifier-free guidance

- Increases the quality of the diffusion model's generation but reduces its diversity.
- Jointly training a conditional diffusion model $p_\theta(z|c)$ and an unconditional one $p_\theta(z)$.
- Sampling is done using Formula (1), where ω is a parameter that controls the guidance.

$$\tilde{\epsilon}_\theta(z_t, c) = (1 + \omega)\epsilon_\theta(z_t, c) - \omega\epsilon_\theta(z_t) \tag{1}$$

Background - Similarity

- Encode the images and texts and compute their similarity using cosine similarity between the resulting vectors.
- Encoders
  - CLIP [RKH+21].
  - DINOv2 [ODM+23].

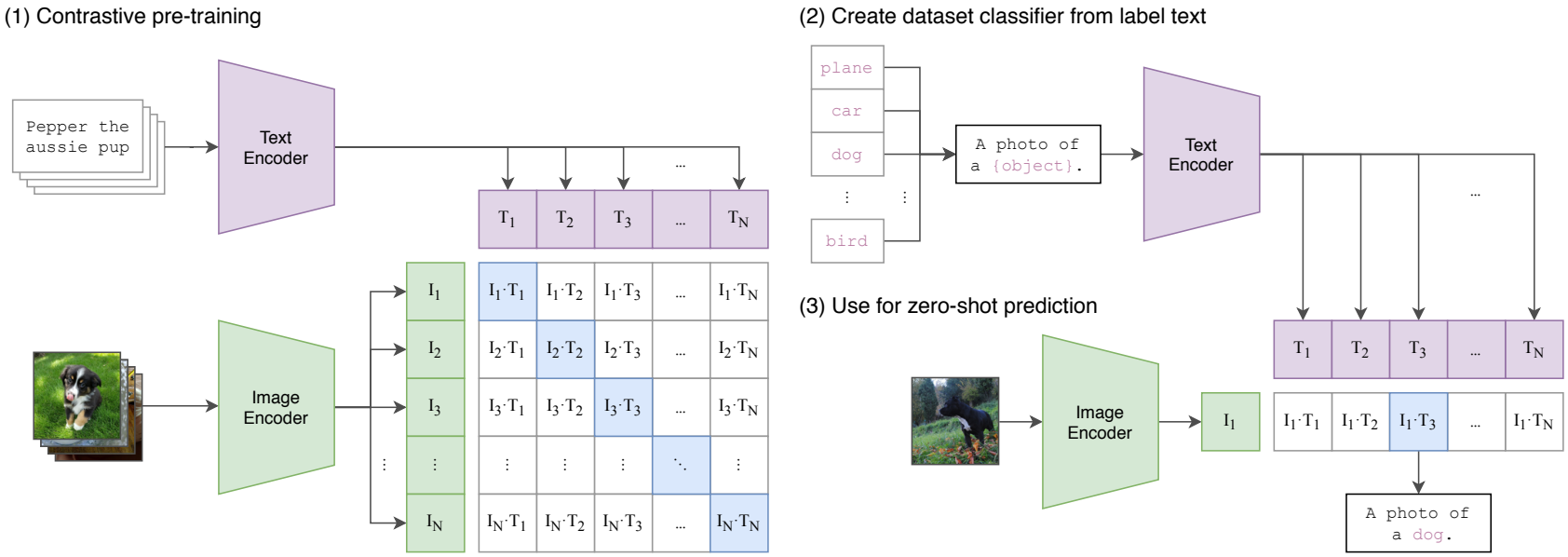


Figure: CLIP architecture. Source [RKH+21]

Prompt-to-Prompt

- Method for generating two similar images based on an initial prompt $\mathcal{P}$ and another edited version of it $\mathcal{P}^*$.
- Injects cross-attention maps from the generation of the initial image $\mathcal{I}$ during the generation of the second one $\mathcal{I}^*$.
- In the case of **word swapping** (e.g. $\mathcal{P}$ = "a velvet chair" to $\mathcal{P}^*$ = "a velvet sofa"), Formula (2) is used during diffusion

$$Edit(M_t, M_t^*, t) := \begin{cases} M_t, & \text{if } t < \mathcal{T} \\ M_t^*, & \text{otherwise,} \end{cases} \tag{2}$$

InstructPix2Pix (1)

- First model that edits existent images by a given prompt.
- Trained on synthetic data.
- Textual data generated by a fine-tuned GPT-3 [BMR+20].
- Image pairs generated using Prompt-to-Prompt followed by CLIP-based filtering.

InstructPix2Pix (2)

- Same architecture as Stable Diffusion [RBL+22].
- Additional layers added to incorporate the initial image as a conditioning.
- Trained using classifier-free guidance with two conditioning to minimize the diffusion objective showcased in Formula (3).

$$L = \mathbb{E}_{\varepsilon(x),\varepsilon(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, \varepsilon(c_I), c_T)||_2^2\right] \quad (3)$$

LIME: Localized Image Editing via Attention Regularization in Diffusion Models

- Region of Interest determined by leveraging the intermediate features of InstructPix2Pix.
- Aims to reduce the effect of unrelated tokens in the edit.
- Modifies the resulting dot product QK^T of the cross-attention layers to a regularized value $R(QK^T, M)$ as shown in Formula (4).

$$R(QK^T, M) = \begin{cases} QK^T_{ijt} - \alpha, & \text{if } M_{ij} = 1 \text{ and } t \in S \\ QK^T_{ijt}, & \text{otherwise} \end{cases} \tag{4}$$

InstructPix2Pix in the interior-design setting

Original image

Replace the marble top with tempered glass.

InstructPix2Pix

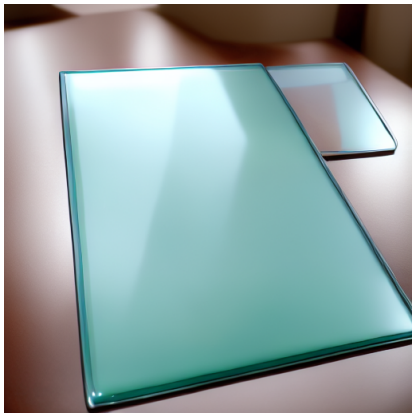


Figure: InstructPix2Pix [BHE23] has limited knowledge in the interior design setting.

A new approach for improving the performance in the interior-design setting

- Generate a context-specific dataset with no previously available data in two steps:
  - Generate the textual data.
  - Generate image pairs based on captions.
- Each instance is a tuple consisting of the initial image, an edit prompt and the edited image.
- Fine-tune the base InstructPix2Pix on the generated dataset.

Textual data generation (1)

- Generate all 3 elements of the tuple.
- Create room-specific agents via in-context learning.
- Instances with hierarchical difficulty: one object caption followed by captions for rooms.

	MTLD [MJ10] ↑	Dugast's U^2 [Dan80] ↑	Guiraud's Index [Dal10] ↑	Yule's K [G. 44] ↓
GPT-3.5	28.13	12.83	3.87	356.49
GPT-4	**32.72**	**13.73**	**4.52**	**278.80**

Table: Comparison of diversity in the textual data generated by GPT models.

Textual data generation (2)

```
completion = client.chat.completions.create(
    model="gpt-4-1106-preview",
    messages=[
        {"role": "user", "content": '''
          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 briefest 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
          For the Remove operation make sure to specify in the initial
               description the feature that you want to remove. For example, if
               you want to remove the drawers of a desk, make sure to specify
               that the desk has drawers in the initial description.
          Additionally, for the Add operation make sure to specify that the
               feature that you want to add does not exist in the initial
               description. For example, if you want to add drawers to a desk,
               make sure that you specify that the desk is one without drawers
               in the initial description.
          {"input": "An elegant dining room with a long table with no candle
               that seats eight, a statement chandelier, and plush velvet chairs
               .", "edit": "Add a candle on the table.", "output": "An elegant
               dining room with a long table with a candle on it that seats
               eight, a statement chandelier, and plush velvet chairs."}
        '''}
    ]
)
```

Image pair generation (1)

- Using Prompt-to-Prompt based on the two generated captions.
- Generating:
  - 30 pairs of images for single objects.
  - 50 pairs of images for rooms.
- Followed by CLIP filtering.
- Publicly available on HuggingFace for **train** and **test**.

Image pair generation (2)

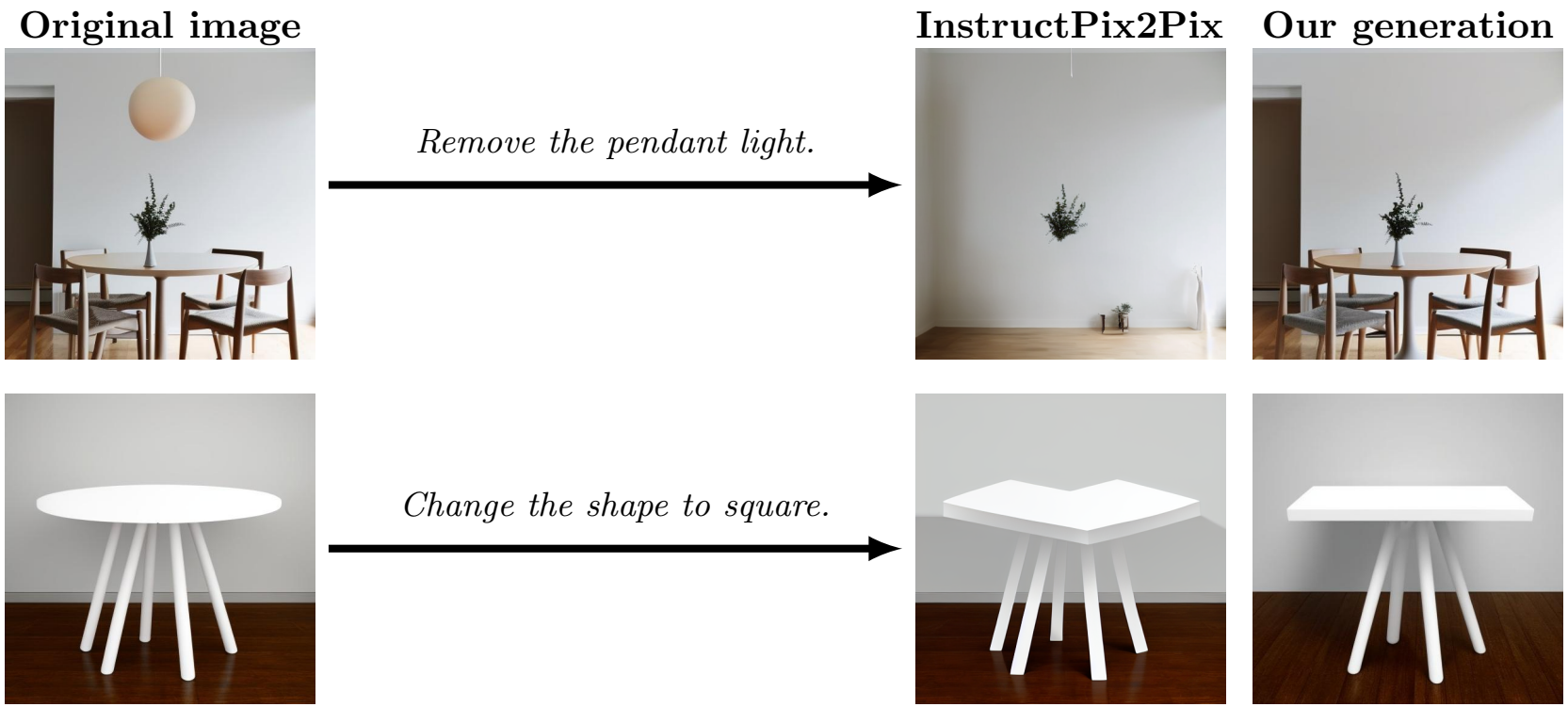


Figure: Comparison between the generated data and InstructPix2Pix's performance on it.

Augmenting the generated dataset

- This problem was not treated before.
- Should force the model to correctly identify the objects in the image.

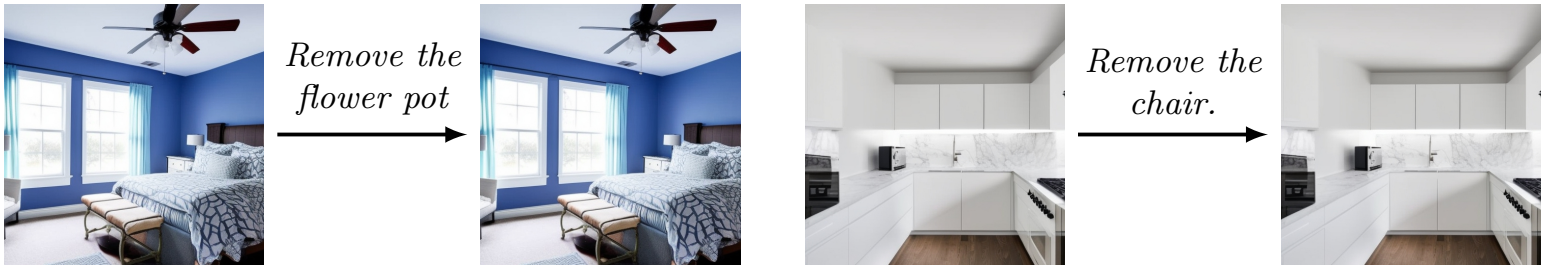


Figure: Data augmentation with samples containing no change in the output image

Fine-tune InstructPix2Pix on the generated data

- Modified images' size from 512×512 to 256×256.
- Trained for 300 epochs with float16 precision.
- Introduced in the training set.

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

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

New model's results

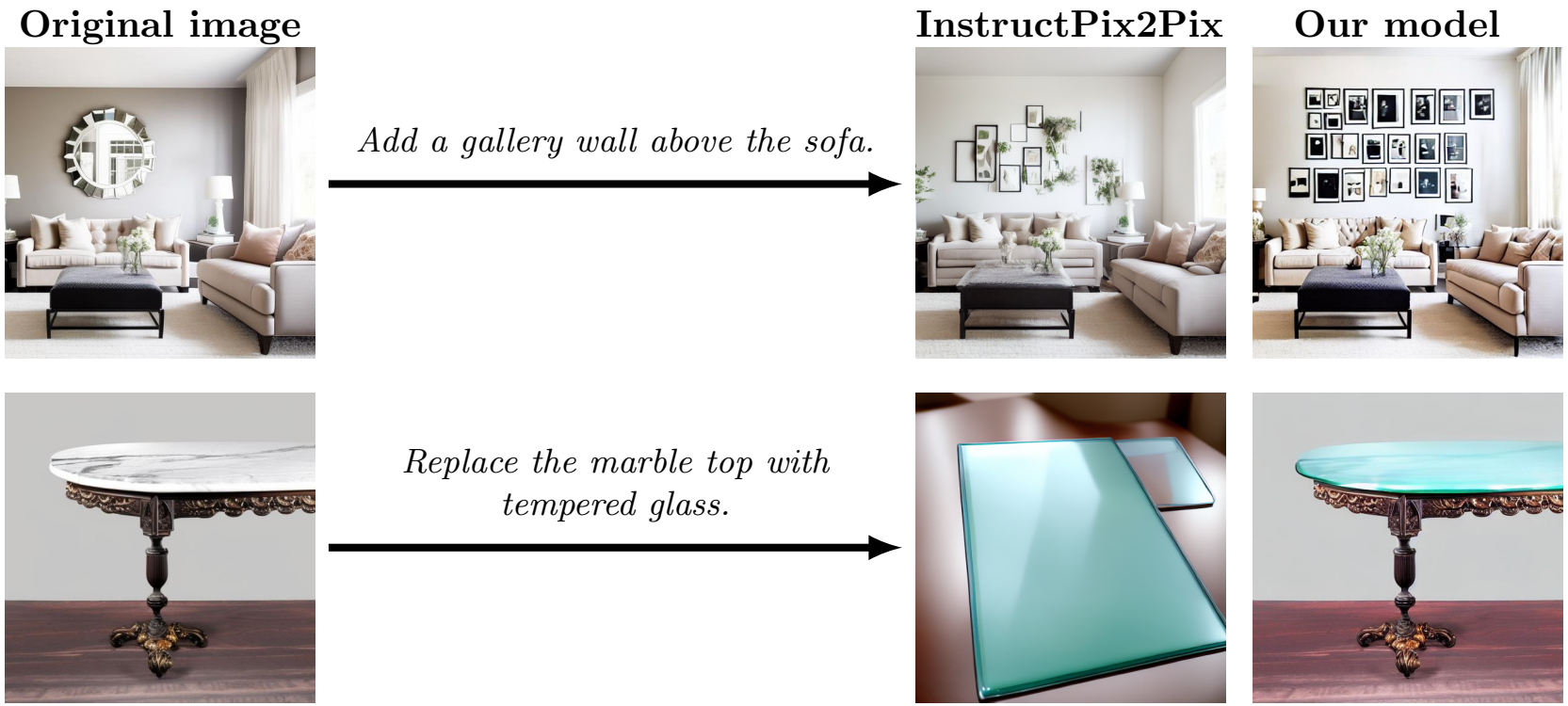


Figure: The new model's edits are a lot more qualitative than the InstructPix2Pix's ones for interior design.

Fine-tune on the unchanged data

- Training for more epochs results in a method that does not edit any image.
- Running for only one epoch results in a model that in some cases has better results than the previous one, but it still does not edit the images in most cases.

Referring Expression Segmentation

- Computing the segmentation mask of an object in an image based on a given expression.
- Generalized Referring Expression Segmentation [LDJ23]
  - Compute the mask for multiple objects.
  - Controlled support when the object is not in the image => empty mask.

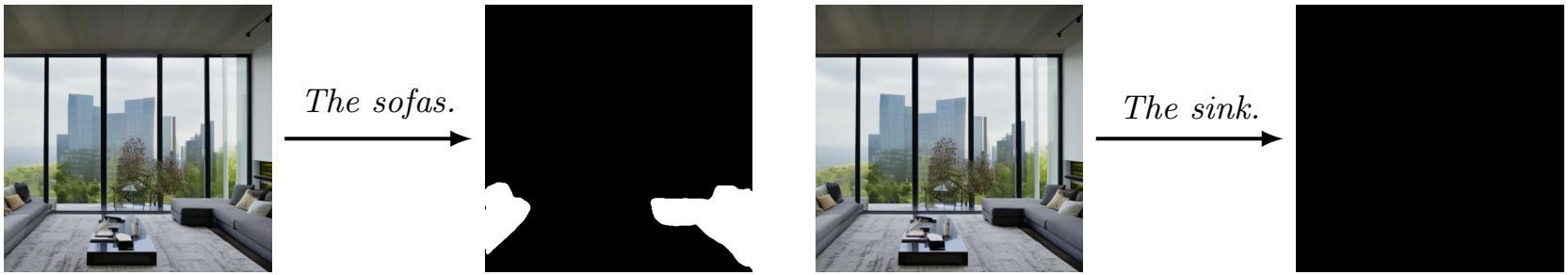


Figure: Examples for GRES computed by ReLA model.

Improve the edit localization (1)

- Compute the region of interest using ReLA.
- Negatively regularizing the cross-attention maps of the unrelated tokens.
- Create an LLM agent via in-context learning for extracting the object(s) reference out of the edit prompt.

Improve the edit localization (2)

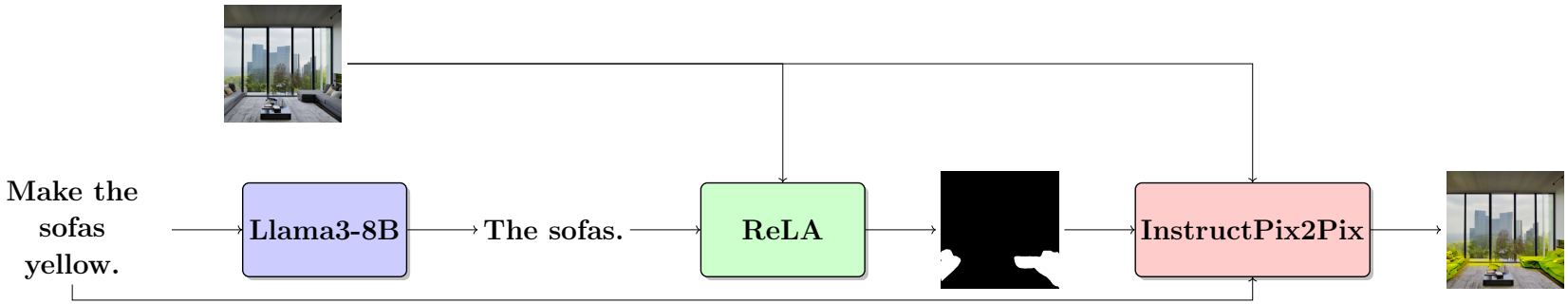
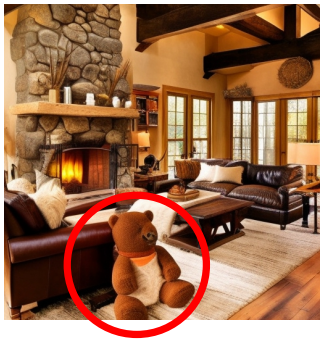


Figure: The pipeline for computing the edit through cross-attention regularization using ReLA's segmentation mask.

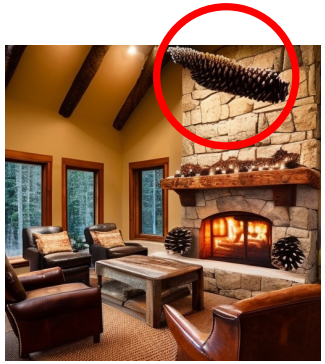
Results

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

Limited knowledge in interior design (1)



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

Limited knowledge in interior design (2)

- **Solution** - generate the images with different models like
  - Muse [CZB+23].
  - Imagen [SCS+22].
  - Interior-design fine-tuned Stable Diffusion published on HuggingFace **here**.

Edit prompts not followed correctly (1)

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

Change the glass top to a wooden top.

Figure: Examples of generated samples that do not correctly follow the edit instruction.

Edit prompts not followed correctly (2)

- **Solution** - a generation and filtering pipeline enhanced with approaches proposed in Emu Edit [SPS+23]:
  - binary injecting the masks of the objects under edit in the Prompt-to-Prompt generation as in Formula (5).
  - integrating image detectors that validate the success of different tasks in the resulting image.

$$x_t \cdot m + (1 - m) \cdot y_t \tag{5}$$

Increase diversity in the dataset

- Generate the textual data with different models like Llama3 [Met24], Gemini [ABW+23] or Mistral 8x7B [JSR+24].
- Create task-room-specific agents with in-context learning.
- Generate images on the same pair with the previously mentioned text-to-image models.

Thank you!

Questions?

Bibliography I


Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al.

Bibliography II

Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.


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, Vancouver, BC, Canada, June 17-24, 2023*, pages 18392–18402. IEEE, 2023.


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack

Bibliography III

Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.
Language models are few-shot learners.
In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

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.
In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors,

Bibliography IV

International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, pages 4055–4075. PMLR, 2023.


Michael Daller.
Guiraud's index.
2010.


Daniel Dugast.
La Statistique Lexicale.
SLATKINE, 1980.


G. Udny Yule.
The Statistical Study of Literary Vocabulary.
Cambridge University Press, 1944.

Bibliography V


Jonathan Ho, Ajay Jain, and Pieter Abbeel.
Denoising diffusion probabilistic models.
In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.


Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed.
Mixtral of experts.

Bibliography VI

CoRR, abs/2401.04088, 2024.


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.


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.

Bibliography VII


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.


Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski.
Dinov2: Learning robust visual features without supervision.
CoRR, abs/2304.07193, 2023.

Bibliography VIII

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer.
High-resolution image synthesis with latent diffusion models.
In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
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, 18-24 July 2021, Virtual Event*, volume 139 of

Bibliography IX

Proceedings of Machine Learning Research, pages 8748–8763. PMLR, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar 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.
In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc., 2022.

Bibliography X


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.