

Instruct-Image Editing with Cross Attention Remapping

Ketan Ramaneti

Nikhil Kandukuri

Rikki Hung

kramanet@andrew.cmu.edu

nkanduku@andrew.cmu.edu

rikkih@andrew.cmu.edu

Abstract/Introduction

Recently, instruction-guided image editing has become the focus of image diffusion models. The initial work in instruction-guided image editing has been done via Google DreamBooth, and later extended by the authors of InstructPix2Pix (IP2P). IP2P has achieved soTA instruction-based editing by combining two pretrained models: a language model (GPT3) and a text-to-image diffusion model (Stable Diffusion). Given an image and a text instruction prompt, GPT3 generates the edited caption, which is passed into the stable diffusion to generate the output with cross attention mechanisms. However, IP2P shows signs of overediting, which is attributed to its attention maps.

Cross-attention techniques were initially presented by authors of Prompt2Prompt and included by IP2P, referring to attention maps in its model pipeline that maps pixels to specific words in a text instruction. However, these attention maps are not localized enough, and thus can lead to overediting of the image and edits made to unrelated regions. Recent works have examined different approaches to identify regions of interest, showing success in limiting cross-attention by image segmentation techniques or min-max scaling and interpolation. Considering IP2P as baseline, we experiment on mitigating over-editing with changing input to be noisy image rather than concatenating it as additional input. Our initial approach aimed at refining the attention maps of IP2P for localized edits, but the base code was tricky to analyze given the deadlines' timeline constraints

Related Papers

Instruction-based Image Editing Models

Prompt2Prompt introduced the task of instruction-based editing and formulated cross-attention control. The latter approach became a critical mechanism enabling the model to simultaneously attends to both the textual prompts and the image, understanding the relationship between them and allowing the edits are aligned with user intentions. Later, InstructPix2Pix (IP2P) extended the instruction-based editing by removing the need for paired input/output prompt at inference time by introducing GPT3 into the pipeline. In the process of generating the edited image, cross attention maps are generated at each denoising step for every token in the input edit instruction prompt.

Augment Edits via Augmented Attention

Dense Attention proposes augmenting text-to-image generation with additional layout guidance, achieving soTA for generation with rich details. Via analyzing the relationship between layout guidance and attention maps, the authors introduce a novel attention modulation mechanism that guides attention focusing adaptively according to layout. Grounded Attention refocusing propose "attention refocusing" in addition to spatial layouts, dynamically adjusting attention during the synthesis process and disentangling relationships between textual descriptions and visual features.

Localized attention techniques

Focus on Your Instruction (Fol) authors follows IP2P pipeline and examines further the effects of multi-step instruction editing via directing each sub-instruction towards their own corresponding region of interest. First, Fol undergoes mask extraction for sub-instruction by identifying the keyword, extracting and blurring its cross-attention map, then enhance the map's contrast via interactive min-max scaling. With extracted region of interest maps, new modified cross-attention maps for the whole IP2P diffusion process are produced by combining masked attention with prompts and masked-out attention with null prompts.

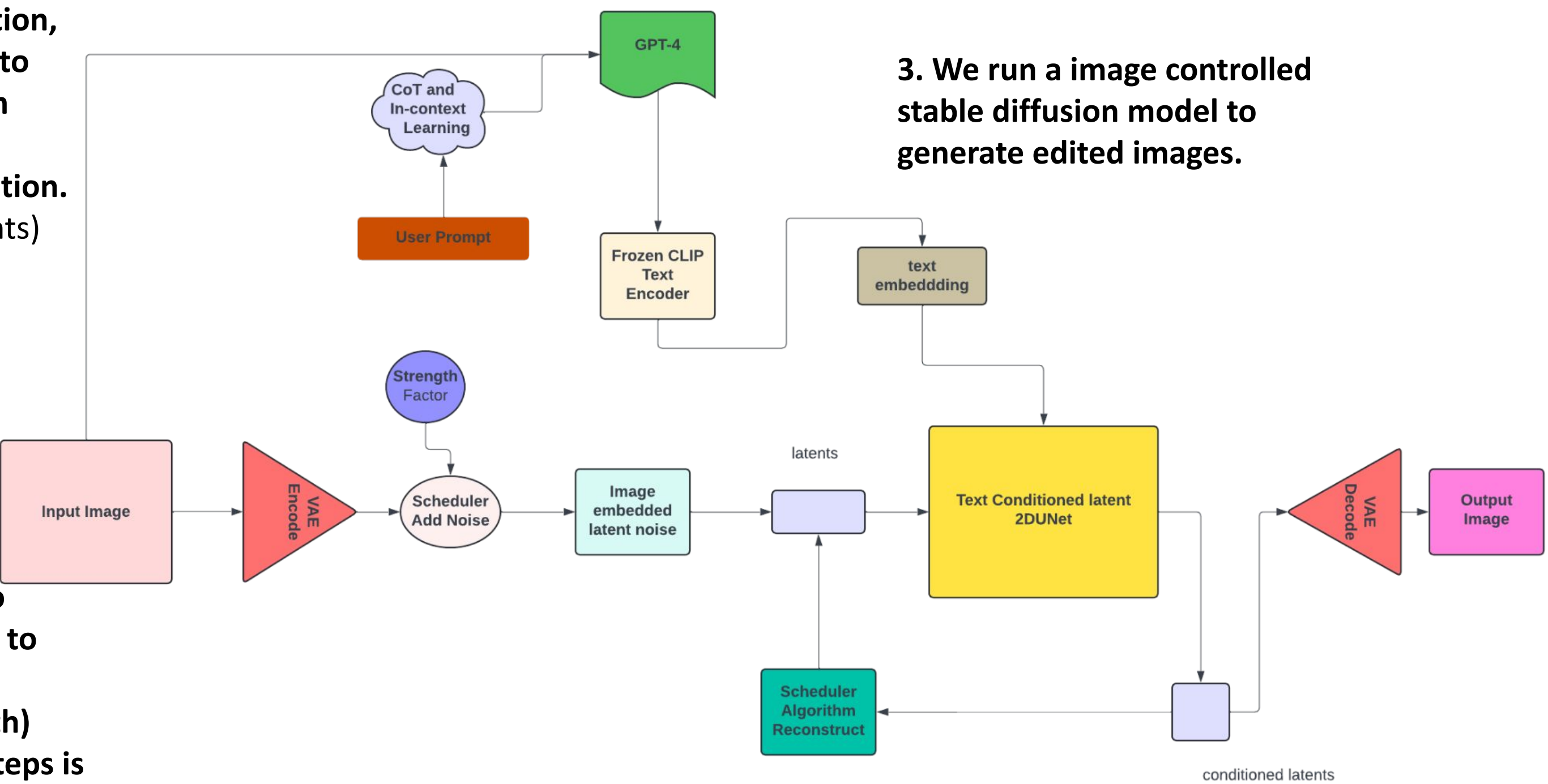
Localized Image Editing (LIME) investigates an alternate approach of identifying region of interests by segmentation. By extracting multi-resolution cross-attention maps across the U-Net architecture, LIME applies resizing, bilinear interpolation, normalization, and clustering on features to generate segmentation masks for regions of interest.

Dataset/Task/Method

We manually curate 50 images to constitute our dataset from renown image editing datasets. These sources include the datasets used by state-of-the-art models performing image editing. We have combined images from InstructPix2Pix, DreamBooth, MagicBrush Dataset, Imagic Dataset and few custom images. Our Task involves performing experiments on these images with the edit instructions and analyze.

1. Given user edit instruction, image, we send them to GPT to rephrase it as an image generation task rather than edit instruction. (Examples in experiments)

2. We run a scheduler to convert the input image to noise latents based on a hyperparameter(strength) num of scheduler timesteps is strength*num_diffuion steps



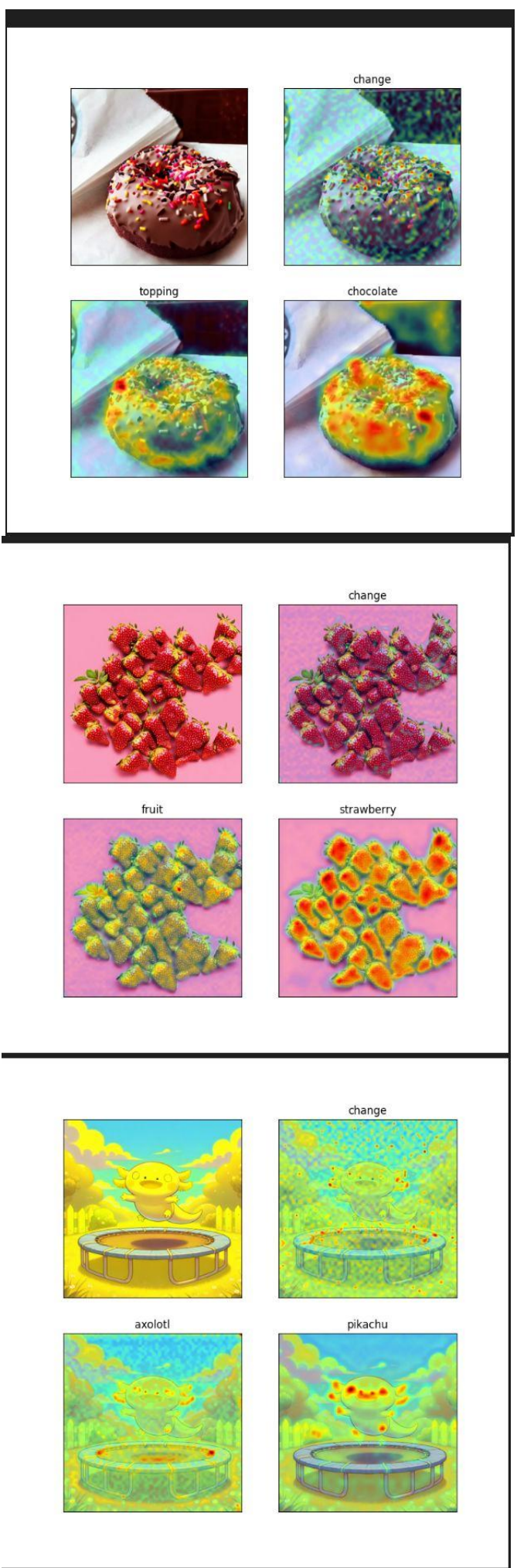
3. We run a image controlled stable diffusion model to generate edited images.

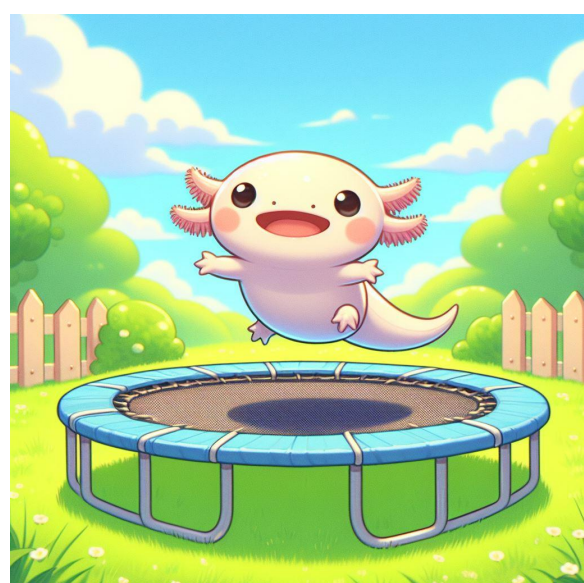




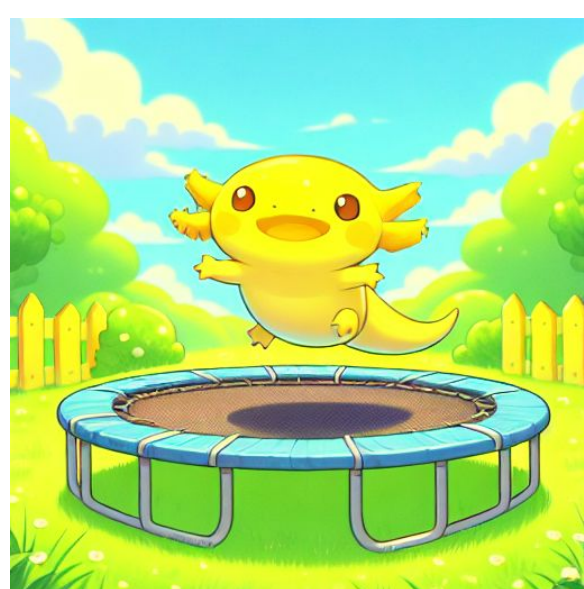

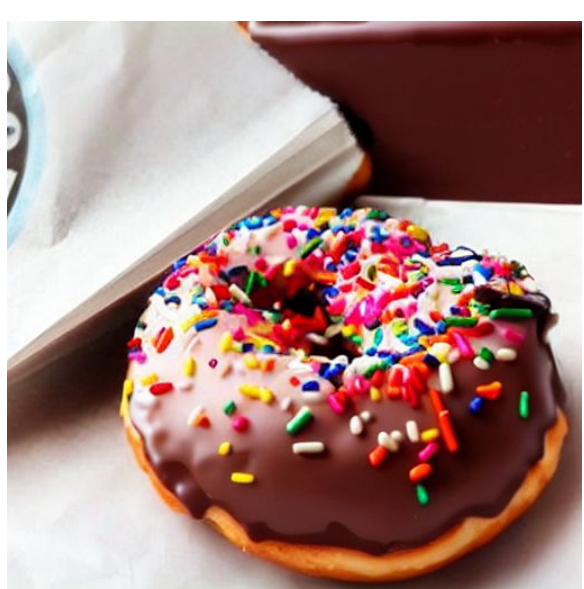


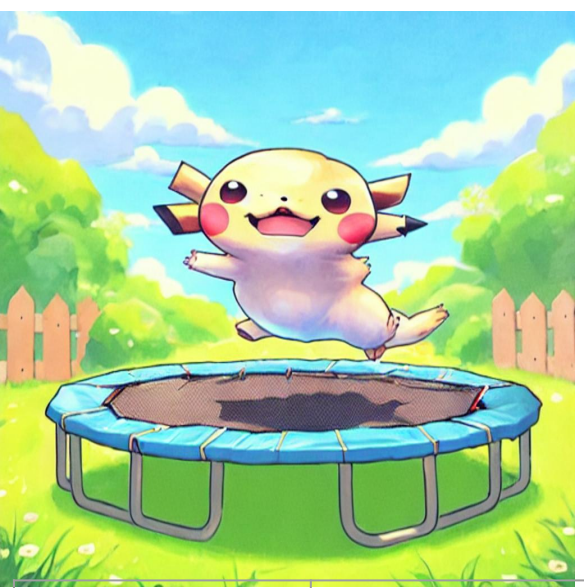




Experiments

2. Qualitative Generations vs Baseline

- We generated images with both baseline and our approach, and also experiments with varying the guidance scales.
- We choose 150 as our diffusion steps, default (best) guidance scale for InstructPix2Pix (img: 1.5, text: 7.5) and choose guidance scale (img: 0.4~0.55, text: 12~15) for our approach)

Table shows the input images, output images from both models, along with their Clip-I(left) and DinoV2(right) scores.



	Input		IP2P baseline		Our Approach					
										
										
	94.63%	99.38%	80.91%	66.21%	95.54%	98.04%	88.96%	95.07%	0.7651	0.8155
										
	90.93%	96.91%	95.43%	92.68%	91.80%	93.24%	91.77%	92.38%	96.83%	94.86%

References

<https://arxiv.org/abs/2210.04885> - What the DAAM: Interpreting Stable Diffusion Using Cross Attention
<https://arxiv.org/abs/2312.10113> - Focus on Your Instruction: Fine-grained and Multi-instruction Image Editing by Attention Modulation
<https://arxiv.org/abs/2211.09800> - InstructPix2Pix: Learning to Follow Image Editing Instructions
<https://openreview.net/forum?id=CDixzkzeyb> - Prompt-to-Prompt Image Editing with Cross-Attention Control
<https://ieeexplore.ieee.org/document/10377914> - Dense Text-to-Image Generation with Attention Modulation
<https://arxiv.org/abs/2306.05427> - Grounded Text-to-Image Synthesis with Attention Refocusing
<https://arxiv.org/abs/2103.00020> - Learning Transferable Visual Models From Natural Language Supervision
<https://arxiv.org/abs/2112.10752> - High-Resolution Image Synthesis with Latent Diffusion Models
<https://ieeexplore.ieee.org/document/10204880> - DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation
<https://arxiv.org/abs/2312.09256> - LIME: Localized Image Editing via Attention Regularization in Diffusion Models

Conclusion

We observe that using our proposed method we are able to achieve more localized edits which was one of the major drawbacks of current SOTA models. From qualitative analysis, with our proposed methodology which works on image-controlled generation, we are able to edit only the areas which require edits and keep the rest of the image almost unaltered. This is reflected in the quantitative scores as well, where our approach has slightly higher similarity scores for both metrics.

3. Quantitative Scores

We calculate the average Clip-I and DinoV2 scores for 10 of our examples, as presented in the table.

Both scores are metrics that compares the similarity of the output image to the input image. A higher score indicates that fewer edits are performed.

	Clip-I	DinoV2
Baseline	89.45%	89.60%
Our Approach	91.79%	90.50%

Edit prompts for baseline

1. change axolotl into pikachu
2. change peach into strawberry
3. change the toppint of the donut to chocolate
4. change the mayonnaise dip into ketchup
5. change the sad expression into a happy smile

Edit prompts for our approach, GPT refined

1. Pikachu on a trampoline
2. Axolotl holding a strawberry
3. Donut with a chocolate topping
4. Burger and ketchup dip
5. Happy smiling dog