

iEdit: Localised Text-guided Image Editing with Weak Supervision

Supplementary Material

In this Supplementary Material, we offer additional insights into our proposed method and elaborate on the experiments outlined in the main paper. We begin with an overview of the experimental setting in Section 4.1. Following that, in Section B, we delve into the limitations of our proposed method. Section D provides further results from additional ablation studies, while Section C offers detailed information on the dataset construction method and the resulting dataset. Finally, in Section E, we present supplementary qualitative results, comparing our method to state-of-the-art approaches.

A. Experimental Setting

We utilise LDMs [33] pre-trained on LAION-5B [37] with the Stable Diffusion (SD) checkpoint v1.4⁴. Fine-tuning of iEdit involves approximately 10,000 steps on 2 16GB NVIDIA Tesla V100 GPUs, with a resolution of 384×384 . The batch size is set to 1, and the learning rate is 2×10^{-4} . To optimise fine-tuning within computational constraints, we alternate updating the input and middle layers of the UNet. Following [7, 43], the classifier-free guidance scale is set to 7.5. Inference, generating four possible editing results per image, takes approximately 10 seconds.

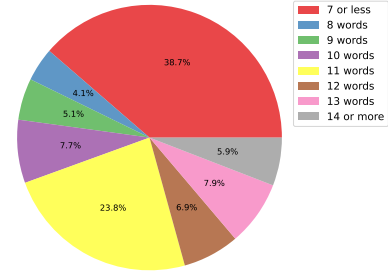
B. Limitations

Our approach relies on off-the-shelf methods to enhance effectiveness. Specifically, we utilise CLIPSeg for segmentation masks and BLIP for generating captions in a controllable manner. However, both tools have imperfections in generating ideal triplets of samples (first image, second image, and edit text). Occasionally, this leads to visually distinct content between the two input images, which poses challenges to the editing task. InstructPix2Pix [4] proposes an alternative method, by generating images instead of retrieving them, leveraging pre-trained DMs in a cyclic manner. However, as discussed in Section 4, this approach also exhibits weaknesses.

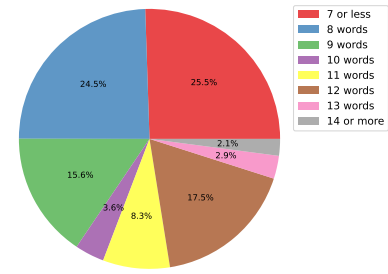
To ensure accessibility and feasibility in training, we have limited ourselves to low computational resources. While this choice accommodates low-memory and few GPU environments, unlocking higher performance may necessitate optimising all parameters simultaneously, albeit at an increased financial cost and carbon footprint.

Evaluating image editing methods poses challenges due to the absence of ad-hoc metrics and a standardised evaluation set. Human evaluation, though valuable, is costly

⁴<https://github.com/CompVis/stable-diffusion>



(a) LAION



(b) Our dataset

Figure S1. Overall caption for both figures

and subjective. We plan to explore more robust evaluation methods in the future.

C. Paired Dataset Construction

In Figure S2, we present a comparison of the approaches used for constructing paired datasets. The construction of LAION-edit-200K is detailed in Section 3.1, while LAION-caption-200K is briefly outlined in Section 4.4. Our observation indicates that the pairs and edit prompts we obtained closely align with the image editing triplets described in Section 3.1. For example, in the first sample, LAION-caption-200K retrieves an image very similar to the source image, sharing the same semantics. Furthermore, the prompt does not explicitly outline any differences. In contrast, our proposed method retrieves an image where the main object changes, reflected in the edit prompt, e.g. “a steamroller”.

In Figure S3 and Figure S4, we illustrate the most frequently used nouns and adjectives in the data ob-

























Original Image	LAION-caption-200K	LAION-edit-200K (ours)	Original Image	LAION-caption-200K	LAION-edit-200K (ours)
					
<i>"learn how to draw a rocket ship"</i>	<i>"Rocket pictures drawn with a pencil"</i>	<i>"a hand drawing a steamroller ship with colored pencils"</i>	<i>"jason isbell's reunions reaches no. 1 on the billboard top country albums chart"</i>	<i>"jason isbell & the 400 unit"</i>	<i>"a man playing a piano in front of a microphone"</i>
					
<i>"ask the pro's: fishing from the shore in seaward, ak"</i>	<i>"activity fishing horizon horizon over water leisure activity men nature one person outdoors real people rock rock - object scenics - nature sea silhouette sky solid standing water"</i>	<i>"a woman standing on a cliff while fishing"</i>	<i>"riviera 235 enclosed fly-bridge"</i>	<i>"fairline targa 58 7572"</i>	<i>"a white sailing vessel is in the water"</i>
					
<i>"cranberry sauce margaritas with rosemary sugar: a holiday party with stages // stirandstrain.com"</i>	<i>"this delicious holiday cranberry mocktail is infused with rosemary and cranberry syrup. topped with a fizzy lime soda! the perfect drink for holiday parties."</i>	<i>"a couple of glasses filled with purple liquid"</i>	<i>"banyan tree..."</i>	<i>"palermo, sicily, italy. botanical garden. ficus also called magnolioides - myvideoimage.com"</i>	<i>"a group of lupine that are next to each other"</i>
					
<i>"reclaimed solid wood pie safe kitchen pantry by griffinfurniture. black bedroom furniture sets. home design ideas"</i>	<i>"301 moved permanently. black bedroom furniture sets. home design ideas"</i>	<i>"a blue clothes closet with a wooden top"</i>	<i>"rsbp small square christmas card pack - christmas gathering"</i>	<i>"rotkehlchen im winter"</i>	<i>"a painting of amphibian sitting on a tree branch"</i>

Figure S2. Comparison of paired dataset construction approaches.

Ablation Settings		Scores			
Losses	Fine-tuning Dataset	CLIPScore (%) \uparrow	FID \downarrow	SSIM- M (%)	SSIM- \bar{M} (%) \uparrow
$\mathcal{L}_{global} + \mathcal{L}_{mask}$	LAION-edit-200K	65.85	156	79.42	58.42
$\mathcal{L}_{global} + \mathcal{L}_{mask} + \mathcal{L}_{loc} + \mathcal{L}_{perc}$	LAION-edit-200K	66.09	146	79.31	59.54
$\mathcal{L}_{global} + \mathcal{L}_{mask} + \mathcal{L}_{loc} + \mathcal{L}_{perc}$ +Masked Inference	LAION-edit-200K	66.97	128	79.65	77.99

Table S1. Additional ablation study of iEdit.

Total number of captions	200475
Number of unique adjectives	34920
Number of unique nouns	93277
Average number of words per prompt	11.11

(a) The original LAION captions

Total number of captions	198591
Number of unique adjectives	2919
Number of unique nouns	10476
Average number of words per prompt	9.84

(b) The paired dataset constructed by the approach proposed in Sec. 3.1.

Table S2. Overall statistics of the datasets.

tained for the LAION-caption-200K and LAION-edit-200K. Analysing LAION-caption-200K captions reveals nouns such as ‘I’, ‘ideas’, and ‘sale’, which are less likely to occur frequently in an edit prompt. Further exploration of the least frequent nouns and adjectives uncovers numerous URLs, foreign words, random numbers, and emojis in the LAION dataset, which are uninformative. In contrast, our dataset features words like ‘quinoa’, ‘muzzler’, and ‘aconite’, indicating a cleaner and more relevant composition. This highlights the noise present in LAION captions, which may not be ideal for forming effective edit prompts.

Further statistics for both datasets are provided in Table S2b and Table S2a. Additionally, the distribution of edit prompt lengths is illustrated in Figure S1b and Figure S1a. Notably, while the average number of words per prompt is higher for the LAION dataset, nearly 40% of them consist of less than 7 words, indicating a lack of detail. This observation is corroborated by the samples in Figure S2, such as “banyan tree”. Given these statistics and our findings that our fine-tuned pre-trained LDMs converge with less than 200K samples, our dataset comprises 200K pairs, but it can be easily expanded.

D. Additional Ablation Study

In Table S1, we provide an additional ablation setting to Table 2 in the main paper. The results show that using only \mathcal{L}_{mask} improves the quality of the generated images and provides better background preservation and the addition of \mathcal{L}_{loc} and \mathcal{L}_{perc} further boosts its performance in all metrics.

E. Additional Qualitative Results

We present additional examples comparing edits performed by our method with state-of-the-art text-guided image editing methods—SDEdit, DALL-E 2 [32], DiffEdit [7], InstructPix2Pix [4]—on images generated by the LDM [33] in Figure S3 and on real images in Figure S4. We consistently observe results in line with the qualitative and quantitative findings presented in the main paper. Notably, SDEdit [28] exhibits shortcomings in faithfully representing the input image (‘a zebra’, ‘fried eggs’) or adhering to the target prompt (‘a sapphire crown’, ‘an orange frog’). DALL-E 2, primarily designed as an inpainting method, excels in preserving the inverse mask area but often lacks fidelity to the style and shape of the input image (‘a school bus’, ‘a winter tree’) and occasionally struggles with seamless integration into the rest of the image (‘a pink car’). DiffEdit frequently yields unsuccessful translations, often stemming from inaccurate mask detection. For example, in ‘a strawberry cake’, the alteration occurs on the plates rather than the cake, and ‘in a bowl of oranges,’ only some of the fruits transform into oranges. Even with accurate mask detection, it may result in failures (‘a winter tree’, ‘a flying bird’). InstructPix2Pix inherits weaknesses such as affecting the entire image (‘a zebra’, a glass with a funny print’, ‘a shark’), struggling with multiple changes (‘a white teddy bear wearing blue’, ‘a doughnut with raspberry and white chocolate sauce’), looking artificial or mismatching the style of the input image (‘a bag with strawberry print’, ‘an astronaut riding a bicycle’), or failing to achieve the target translation (‘a white wedding cake’, ‘a yellow rose’). In contrast, our method consistently demonstrates higher fidelity to both the edit prompt and the input image.

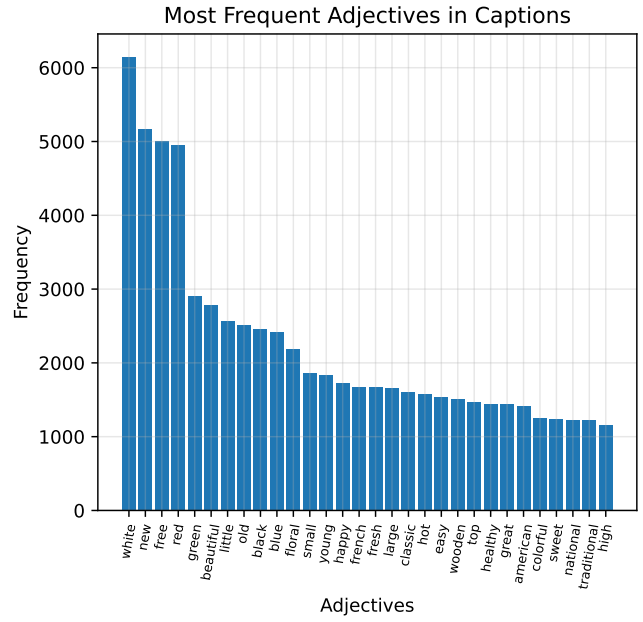
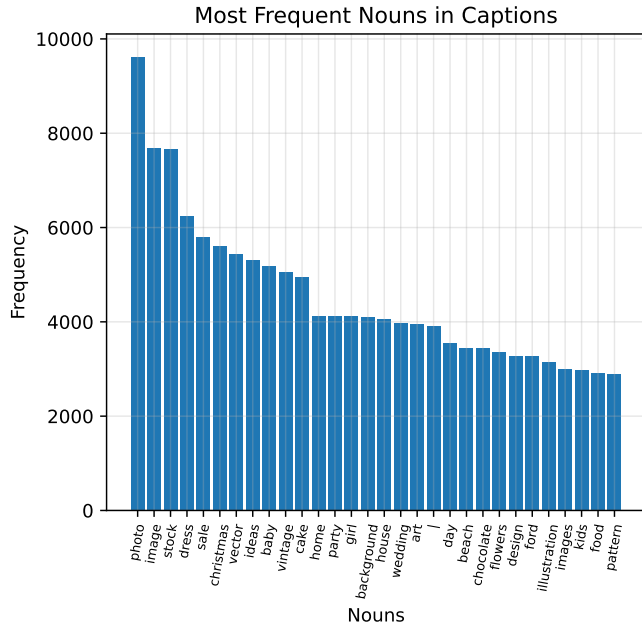


Figure S3. Most frequently used Nouns and adjectives in the original LAION captions of the constructed dataset.

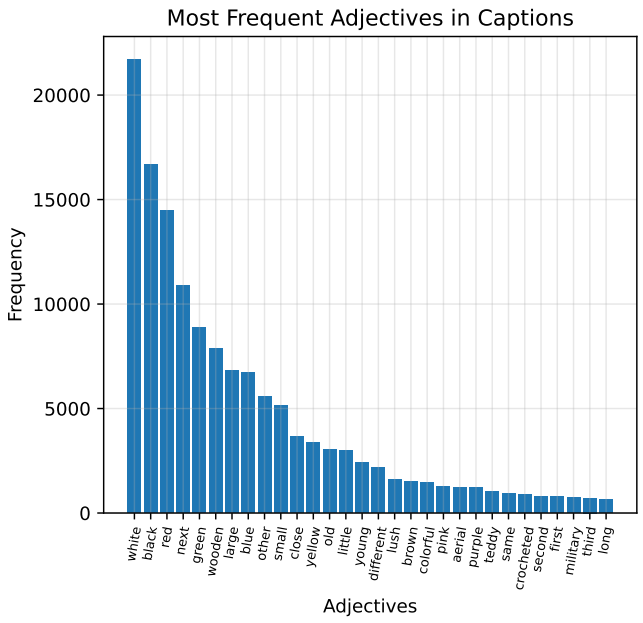
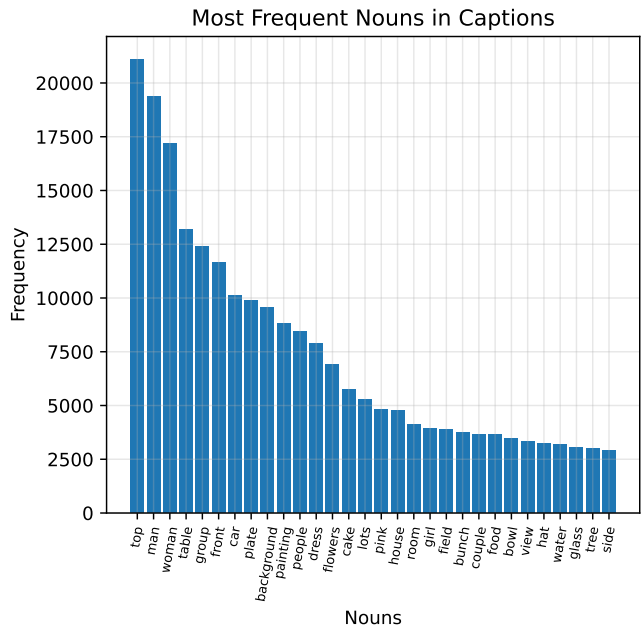

































































Figure S4. Most frequently used Nouns and adjectives in the edit prompts of the constructed dataset.

Input Image		SDEdit	DALL-E 2	DiffEdit	InstructPix2Pix	iEdit (ours)	iEdit-M (ours)
	a strawberry cake						
	a zebra						
	a bus						
	a pink car						
	a sports car						
	a laughing cat						
	a screaming cat						
	a mouse						
	a boat on the sea						

Continued on next page

Table S2 – Continued from previous page

Input Image		SDEdit	DALL-E 2	DiffEdit	InstructPix2Pix	iEdit (ours)	iEdit-M (ours)
	a glass with funny prints						
	a flying bird						
	a ball with a smiley face						
	an astronaut riding a bicycle						
	an astronaut running on the street						
	a crown with sapphire						
	a white daisy						
	a yellow rose						
	a cupcake with a candle						

Continued on next page

Table S2 – Continued from previous page












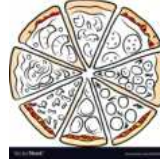












































Input Image		SDEdit	DALL-E 2	DiffEdit	InstructPix2Pix	iEdit (ours)	iEdit-M (ours)
	a victorian lamp						

Table S3. Comparison of our method to state-of-the-art on images generated by the LDM [33].

Input Image		SDEdit	DALL-E 2	DiffEdit	InstructPix2Pix	iEdit (ours)	iEdit-M (ours)
	a pizza						
	a wooden house						
	a zebra on snow						
	rubber ducks						
	a white wedding cake						
	a bronze horse						
	a bag with strawberry						

Continued on next page

Table S3 – Continued from previous page

Input Image		SDEdit	DALL-E 2	DiffEdit	InstructPix2Pix	iEdit (ours)	iEdit-M (ours)
	a bowl of apples						
	a glass bowl of fruits						
	a bowl of oranges						
	a glass vase of tulips						
	a winter tree						
	a shark						
	an orange frog						
	a frog on a lilly pad						
	a wolf						

Continued on next page

Input Image		SDEdit	DALL-E 2	DiffEdit	InstructPix2Pix	iEdit (ours)	iEdit-M (ours)
	two pigs						
	two black and white cows						
	a dog						
	a white teddy bear wearing blue						
	a doughnut with white chocolate and raspberry sauce						
	fried eggs						
	a single bed						
	a high-speed train						

Table S4. Comparison of our method to state-of-the-art on real images.