LOCATEdit: Graph Laplacian Optimized Cross Attention for Localized Text-Guided Image Editing

Achint Soni[1] Meet Soni[2] Sirisha Rambhatla[1]
[1]University of Waterloo [2]Stony Brook University
{a2soni, sirisha.rambhatla}@uwaterloo.ca, meet.soni@stonybrook.edu

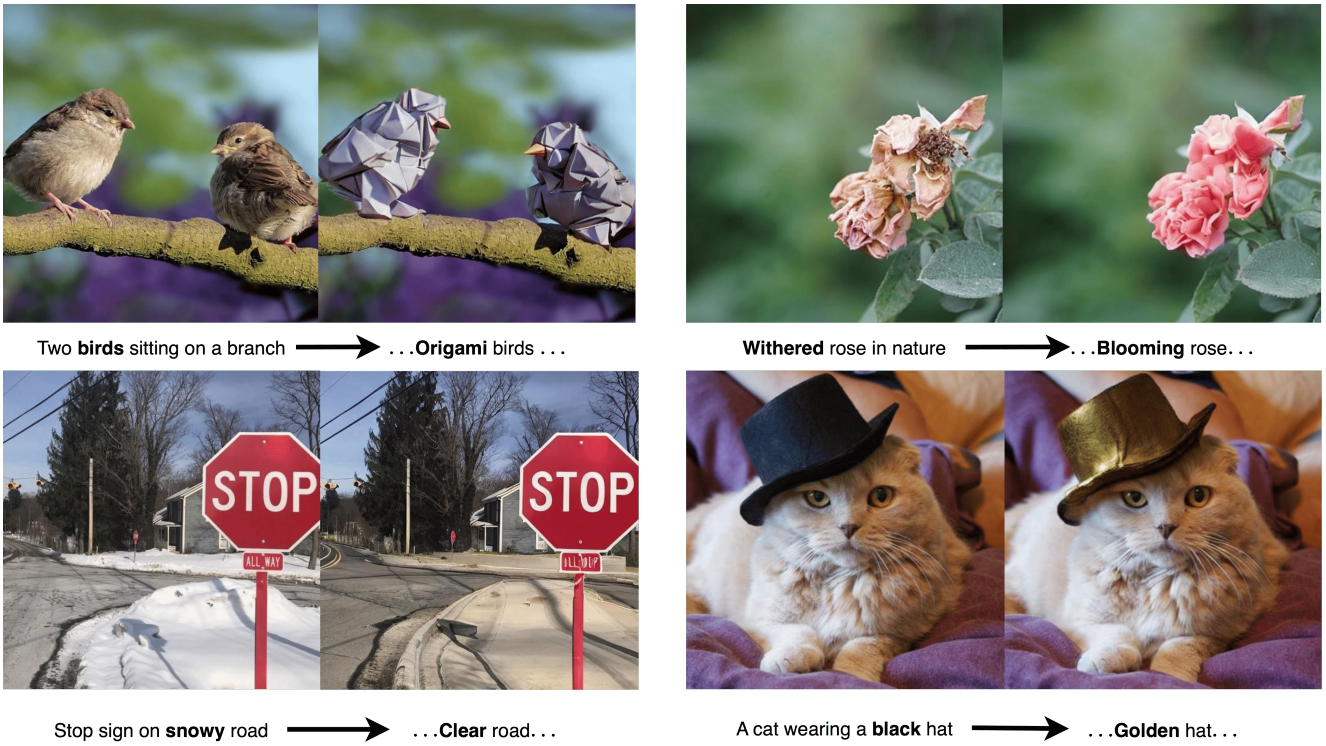

Figure 1. Our LOCATEdit demonstrates strong performance on various complex image editing tasks.

Abstract

Text-guided image editing aims to modify specific regions of an image according to natural language instructions while maintaining the general structure and the background fidelity. Existing methods utilize masks derived from cross-attention maps generated from diffusion models to identify the target regions for modification. However, since cross-attention mechanisms focus on semantic relevance, they struggle to maintain the image integrity. As a result, these methods often lack spatial consistency, leading to editing artifacts and distortions. In this work, we address these limitations and introduce ***LOCATEdit****, which enhances cross-attention maps through a graph-based approach utilizing self-attention-derived patch relationships to maintain smooth, coherent attention across image regions, ensuring that alterations are limited to the designated items while retaining the surrounding structure. LOCATEdit consistently and substantially outperforms existing baselines on PIE-Bench, demonstrating its state-of-the-art performance and effectiveness on various editing tasks. Code can be found on* https://github.com/LOCATEdit/LOCATEdit/

1. Introduction

Diffusion models have become popular for image generation, yet practical applications demand precise control for editing. Text-guided editing techniques [4, 43, 49] have emerged as powerful tools to facilitate such modifications across domains, from digital art [8, 12, 17, 38] to medical imaging [25, 42], enabling more intuitive image manipulation through natural language prompts. However, prompt-driven editing is often imprecise [2, 6, 15].

To attain precise control in text-guided image editing, recent studies use masks derived from cross-attention maps; however, inaccuracies in these maps can result in edits spilling over unintended regions, causing problems such as object identity loss [18, 35, 44] and background drift [23, 49]. Because of this, techniques that depend exclusively on cross-attention could make global changes when only localized modifications are required [9, 15]. These problems highlight the necessity for a method that precisely identifies editing areas without jeopardizing the integrity of the overall image.

Recent methods have demonstrated improved mask accuracy through the utilization of cross- and self-attention masks, while simultaneously adopting the dual branch editing paradigm [43, 49]. Additionally, they incorporate target image embeddings as auxiliary guidance derived from source image embeddings and the editing information contained in source-target prompt pairs. Despite these, challenges such as unintended spills continue to be a problem, which can be seen in Figure 2. Our key observation is that naively combining cross-attention and self-attention results in significant information loss. Consequently, we propose to induce spatial consistency and precise identification of regions to be edited via a graph-based approach. Given graphs $\mathcal{G}_{\text{src}}$ and $\mathcal{G}_{\text{tgt}}$ for the source and target branch, respectively, each of these graphs is constructed using the respective Cross and Self-Attention, hence CASA graphs, encoding cross-attention maps as nodes and self-attention relationships as weighted edges. With this abstraction, these graphs intrinsically depict the structure of the image, thereby connecting local and global contexts, while also maintaining the semantic relevance.

We explicitly enforce graph structure by proposing a graph Laplacian regularizer on $\mathcal{G}_{src}$ and $\mathcal{G}_{tgt}$ to impose spatial consistency, motivated by the effectiveness of Laplacian regularization in image denoising and mesh editing [27, 37]. Prior works on segmentation and spatial regularization [41, 51] also demonstrate that this Laplacian constraint effectively maintains object boundaries and preserves local detail, thus *disentangling* the areas of interest from unrelated regions. Furthermore, Belkin and Niyogi [1] and Lim et al. [22] illustrate that this regularization also enhances the separation of semantic characteristics. By integrating a Laplacian smoothness factor into the diffusion process, LOCATEdit optimizes the attention values across interconnected patches without any additional training, hence reducing background drift and limiting global changes as can be seen in Figure 1. This ensures that modifications are confined to designated areas while preserving the overall structural integrity of the original image. Notably, this optimization admits a closed-form solution, hence eliminating the need for iterative refinement [43]. Overall, our contributions can be summarized as follows:

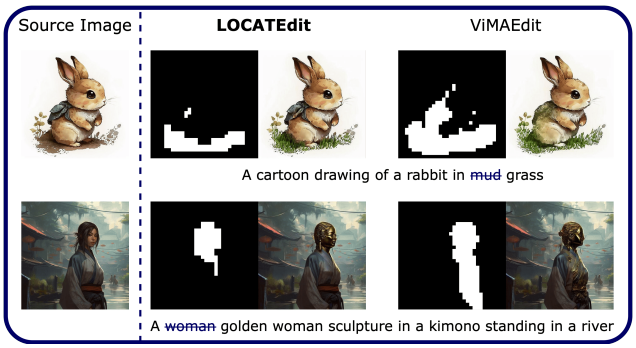

Figure 2. Example of over-editing caused due to imprecise masks.

- **CASA Graph:** We introduce **LOCATEdit**, a method which encapsulates word-to-pixel relevance through pixel-to-pixel relationships by modeling attention maps with CASA graph.
- **Improved spatial consistency:** By optimizing masks through graph Laplacian regularization on the CASA graph, we maintain object structure and confine changes to intended regions, hence minimizing distortions.
- **Disentangled and faithful editing:** Leveraging Laplacian smoothing, LOCATEdit achieves precise semantic modifications while preserving the original image context, ensuring disentangled editing.

2. Related Work

Recent advances in image editing have leveraged a range of conditioning modalities—including text, reference images, and segmentation maps—to drive semantic, structural, and stylistic modifications [13]. In this work, we focus specifically on text-guided image editing with an emphasis on preserving the original content and ensuring effective foreground-background disentanglement.

2.1. Text-guided Image Editing

Early methods exploited the power of CLIP [30] to align images and text. For example, [16] fine-tuned diffusion models during reverse diffusion using a CLIP loss to adjust image attributes, though these approaches were limited to global changes and often suffered from degraded image quality. Later works such as DiffuseIT [19] and StyleDiffusion [46] improved performance by introducing seman-

tic or style disentanglement losses; however, they are computationally expensive and typically confined to specific style modifications. More recent frameworks like InstructPix2Pix [3] preserve source content using text instructions, yet require carefully curated instruction-image pair datasets and supervised training. Additionally, methods that manipulate text embeddings for disentangled editing have been explored [47], though they often yield only marginal improvements over earlier approaches.

Collectively, these studies underscore both the promise and limitations of text-guided editing, motivating our work on refining attention maps to achieve spatially consistent and localized modifications.

2.2. Training Free Image Editing

Recent advances in text-to-image synthesis [7, 26, 31, 33, 34] have enabled high-quality photorealistic image generation from text prompts. Building on these advances, several studies have proposed dual-branch, training-free approaches that leverage rich feature and attention maps from pre-trained diffusion models for image editing. These methods exploit signals from the source image's diffusion process to drive content modification, obviating the need for additional model training while achieving remarkable success in altering image content.

Notably, PRedITOR [32] generates a target CLIP embedding via a diffusion prior model but struggles with fine detail and background consistency. Other methods enhance structural control: P2P [9] replaces cross-attention maps to maintain spatial alignment, and PnP [40] injects spatial features and self-attention maps into decoder layers. Approaches like MasaCtrl [4] preserve structure through mutual self-attention, while editing-area grounding techniques and attention regularization losses are employed in DPL [49] and refined further in ViMAEdit [43]. Despite these advances, challenges in achieving precise localization and consistent edits persist, motivating our work.

2.3. Graph Laplacian

In optimization and semi-supervised learning, Laplacian regularization promotes smooth variation along a graph, similar to how Conditional Random Fields (CRFs) refine segmentation by enforcing spatial and color consistency [20]. Unlike CRFs, Laplacian smoothing is fully differentiable and easily integrated into neural networks. Its effectiveness has been demonstrated in tasks such as image matting [21], where the matting Laplacian preserves edges while interpolating unknown regions, and in action localization [28], where it refines class activation maps for more coherent predictions.

3. Background

3.1. Diffusion models

Diffusion models [11, 26, 36] constitute a class of generative approaches that operate through two complementary processes—forward and backward diffusion. In the forward diffusion process, starting from an original clean sample $\mathbf{z}_0$ (drawn from the data distribution), Gaussian noise is iteratively added at each timestep $t = 1, 2, \ldots, T$. Specifically, one obtains

$$\mathbf{z}_t = \sqrt{\alpha_t}\,\mathbf{z}_0 \;+\; \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T,$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an independent Gaussian noise term injected at timestep t. The sequence $\{\alpha_t\}_{t=1}^{T}$ governs the noise variance at each stage, ensuring that after T diffusion steps, $\mathbf{z}_T$ is approximately distributed as a standard Gaussian.

The backward diffusion process reverses this corruption procedure by progressively denoising the noisy sample $\mathbf{z}_T$ into a cleaner sample $\mathbf{z}_{T-1}$, then $\mathbf{z}_{T-2}$, and so forth, converging to a final clean reconstruction $\mathbf{z}_0$. To accomplish this, one samples $\mathbf{z}_{t-1}$ from a conditional distribution over $\mathbf{z}_t$, typically parameterized by a learnable denoising function. Formally, the update rule may be expressed as

$$\mathbf{z}_{t-1} = \boldsymbol{\mu}_t\big(\mathbf{z}_t, \theta\big) + \sigma_t\,\tilde{\boldsymbol{\epsilon}}_t, \quad t = T, \ldots, 1,$$

where $\tilde{\boldsymbol{\epsilon}}_t$ is a random Gaussian noise , $\boldsymbol{\mu}_t$ and σ_t represents the mean and variance of distribution that $\mathbf{z_{t-1}}$ can be sampled from, and θ encapsulates the learned parameters. In the DDIM formulation [36], one often employs a deterministic variant by modifying the variance schedule, making the sampling process more efficient while maintaining high sample quality.

A pivotal component in modern diffusion models is the noise prediction network $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$. Rather than predicting $\mathbf{z}_0$ or $\mathbf{z}_{t-1}$ directly, the network estimates the noise present in the corrupted sample $\mathbf{z}_t$. Once trained, this noise predictor effectively guides the reverse diffusion steps to iteratively remove the injected Gaussian noise.

3.2. Attention mechanism

In practice, the noise prediction model is frequently instantiated by a U-Net architecture, chosen for its efficacy in pixel-level prediction tasks. Each U-Net block typically consists of (i) a residual convolutional sub-block that refines the spatial representation of the intermediate feature maps, and (ii) a self-attention sub-block that captures long-range patch-to-patch dependencies. (iii) cross-attention sub-block that aligns the image to textual information

In the mechanism, feature tensors are first projected into three distinct embeddings—queries Q, keys K, and values

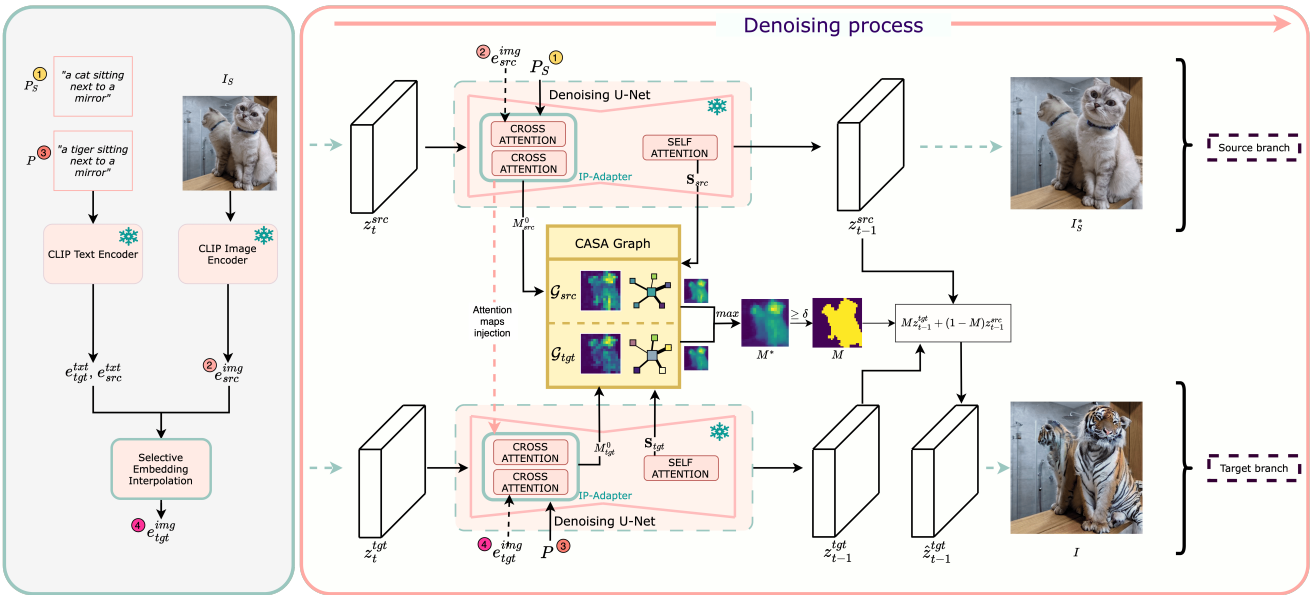

Figure 3. **Overview of our text-guided image editing pipeline.** LOCATEdit refines cross-attention maps with graph Laplacian regularization for spatial consistency, uses an IP-Adapter for additional guidance, and employs selective pruning on text embeddings to suppress noise, ensuring the edited image preserves key structural details.

V. Attention is computed as

$$\text{Attention}(Q, K, V) = \text{Softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)\, V,$$

where d is the dimensionality of the query/key vectors, In both self-attention and cross-attention layers, Q is projected from spatial features. In self-attention, K and V also come from spatial features, whereas in cross-attention, they are projected from textual embeddings. These projections use learned metrics that are optimized during training.

4. LOCATEdit

In this section, we present LOCATEdit for precise, localized text-guided image editing that refines the cross-attention maps. Our approach integrates two complementary modules. First, we utilize the CASA graph to impose spatial coherence and ensure that edits are restricted to the designated areas. Second, building upon previous work [43], we integrate an image embedding-enhanced denoising process augmented by a selective pruning operator applied to the text embedding offsets. This operator eliminates minor semantic variations, therefore minimizing unwanted changes and avoiding unnecessary editing of non-target regions. Together, these modules allow LOCATEdit to maintain the structural integrity of the original image while precisely implementing the desired edits.

4.1. Dual-Branch Editing Paradigm

Our pipeline employs a dual-branch design in which a source branch reconstructs the original image and a target branch generates the edited output. To maintain structural consistency, both branches start from the same initial noise $\mathbf{z}_T$ and share intermediate latent variables. Crucially, we inject the cross-attention maps from the source branch into the target branch [9] to maintain the spatial structure. Formally, if Q^{src} and K^{src} are the query and key embeddings from the source branch and V^{tgt} denotes the value embeddings from the target branch, then the target cross-attention is computed as

$$\text{Attention}\big(Q^{\text{src}}, K^{\text{src}}, V^{\text{tgt}}\big) = \text{Softmax}\Big(\frac{Q^{\text{src}}(K^{\text{src}})^\top}{\sqrt{d}}\Big)\, V^{\text{tgt}}.$$

4.2. Selective Embedding Interpolation

Following previous work [43], we employ an IP-Adapter [50] to provide explicit guidance for target image generation. After extracting the source image embedding e^{img}_{src} and the CLIP-based text embeddings e^{txt}_{src} and e^{txt}_{tgt} corresponding to the source and target prompts respectively, the conventional target image embedding is computed as

$$e^{img}_{\text{tgt}} = e^{img}_{\text{src}} + \Big(e^{txt}_{\text{src}} - e^{txt}_{\text{tgt}}\Big). \tag{1}$$

This embedding is then processed by the IP-Adapter, which projects it into a latent feature space that is integrated into

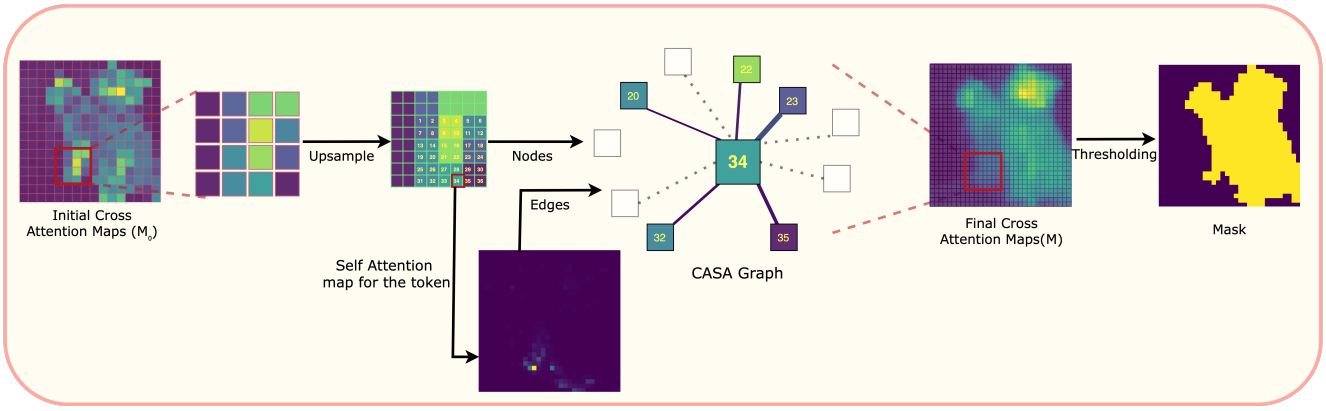

Figure 4. **CASA (Cross and Self-Attention) Graph Construction workflow.** The initial cross-attention maps are upsampled to form a patch-level adjacency graph, then Laplacian regularization enforces spatial consistency. Thresholding the refined maps yields final, more robust attention masks.

the diffusion model's cross-attention mechanism. Specifically, given the query Q (derived from the noisy latent), the IP-Adapter produces additional key and value features K^{IP} and V^{IP} from the projected target embedding. These are then combined with the original key K and value V features to form the final cross-attention:

$$Z = \text{Attention}(Q, K, V) + \lambda \text{Attention}(Q, K^{\text{IP}}, V^{\text{IP}})$$

thereby incorporating semantic guidance into the diffusion process without requiring any additional training.

A limitation of directly using the difference $e^T_{\text{src}} - e^T_{\text{tgt}}$ in Equation (1) is that low-magnitude components, inherent in the entangled nature of CLIP text embeddings [24], can lead to unintended edits. To mitigate this, we introduce a selective pruning operator $\mathcal{H}$ that thresholds the text difference, retaining only the dominant semantic offsets. Formally, we replace Equation (1) with

$$e^I_{\text{tgt}} = e^I_{\text{src}} + \mathcal{H}\Big(e^T_{\text{src}} - e^T_{\text{tgt}}\Big), \tag{2}$$

where $\mathcal{H} : \mathbb{R}^d \to \mathbb{R}^d$ is defined elementwise as

$$\big[\mathcal{H}(\mathbf{y})\big]_i = \begin{cases} y_i, & \text{if } |y_i| \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Here, d is the embedding dimension and τ is determined via a percentile threshold on the absolute values of the difference. This selective pruning ensures that only significant semantic shifts contribute to the target image embedding, thereby reducing the risk of global edits and preserving the structural consistency of non-target regions. The pruned embedding is then processed through the IP-Adapter as described above, ensuring that the final diffusion process is both semantically guided and robust to minor, spurious variations.

4.3. Formulating CASA Graph

While the IP-Adapter provides explicit semantic guidance, the cross-attention maps extracted during denoising may still contain spills that lead to unintended edits. To address this, we refine these attention maps by modeling them as a CASA graph, where each node represents an image patch and the edges capture patch-to-patch relationships obtained from self-attention as can be seen in Figure 4. The graph Laplacian regularization enforces a smoothness constraint across the CASA graph, penalizing abrupt differences in attention between strongly connected patches. In effect, this smoothing suppresses isolated high responses that can cause over-editing, ensuring that only spatially coherent regions receive significant modifications. By harmonizing the attention values over connected patches, LOCATEdit robustly confines edits to the intended regions and preserves the overall spatial consistency.

Formally, within each U-Net block, each prompt word is linked to a cross-attention map; however, only the cross-attention maps related to the blend word(s) are necessary. Following previous studies [5, 9, 49], we compute the average of the cross-attention maps obtained from multiple U-Net blocks to get initial maps. We obtain initial attention maps for both the source and target branches, denoted as $M^{\text{src}}_0 \in \mathbb{R}^{r\times r}$ and $M^{\text{tgt}}_0 \in \mathbb{R}^{r\times r}$. These masks are then upsampled to a higher resolution of $R \times R$ (where $R = \gamma r$ and $\gamma > 1$) to capture fine spatial details, and subsequently flattened to yield the initial saliency maps $\mathbf{m}^{\text{src}}_0, \mathbf{m}^{\text{tgt}}_0 \in \mathbb{R}^{R^2}$.

To prioritize high-confidence regions, we compute a weight for each patch by applying the sigmoid function $\sigma(\cdot)$ to the corresponding element of $\mathbf{m}^{\text{src}}_0$ and then squaring the output. Squaring the sigmoid output emphasizes larger values while further suppressing lower ones, thereby enhancing the reliability of high-confidence regions. These weights are assembled into a diagonal confidence matrix

with a scaling factor α:

$$\mathbf{\Lambda}^{\text{src}} = \text{diag}\Big(\sigma\Big(\alpha \mathbf{m}_0[1]\Big)^2, \ldots, \sigma\Big(\alpha \mathbf{m}_0[R^2]\Big)^2\Big). \quad (4)$$

and similarly for $\mathbf{\Lambda}^{\text{tgt}}$.

Next, we extract self-attention maps $\mathbf{S}^{\text{src}} \in \mathbb{R}^{R^2 \times R^2}$ and $\mathbf{S}^{\text{tgt}} \in \mathbb{R}^{R^2 \times R^2}$ for the source and target branches, respectively. To ensure mutual relationships are treated uniformly and to guarantee the convexity of the optimization, we symmetrize both the maps as

$$\mathbf{S}_{\text{sym}} = \frac{1}{2}\Big(\mathbf{S} + \mathbf{S}^\top\Big). \quad (5)$$

Now, for each branch we construct CASA graph $\mathcal{G} = (V, E)$ where each node $v_i \in V$ corresponds to a patch in the flattened saliency map $\mathbf{m}_0$. The edge weight between nodes v_i and v_j is given by the symmetrized self-attention map $\mathbf{S}_{\text{sym}}$. This graph structure, with nodes representing the initial saliency values and edges capturing inter-patch relationships, serves as the foundation for the CASA graph.

4.4. Graph Laplacian Regularization

After initializing CASA graphs $\mathcal{G}_{\text{src}}$ and $\mathcal{G}_{\text{tgt}}$ for both branches, we optimize for the value of their nodes using graph Laplacian optimization.

Formally, graph Laplacian is defined by:

$$\mathbf{L} = \mathbf{D} - \mathbf{S}_{\text{sym}}.$$

where $\mathbf{D}$ is a degree matrix for $\mathbf{S}_{\text{sym}}$, which is computed as

$$\mathbf{D}(i,i) = \sum_{j=1}^{R^2} \mathbf{S}_{\text{sym}}(i,j), \quad \mathbf{D}(i,j) = 0 \quad \text{for } i \neq j,$$

Lemma 1. *The graph Laplacian $\mathbf{L} \in \mathbb{R}^{R^2 \times R^2}$ is positive semidefinite.*

Detailed proof is provided in Appendix 8.

We then optimize the initial saliency maps $\mathbf{m}_0^{\text{src}}$ and $\mathbf{m}_0^{\text{tgt}}$ for both branches through the following convex optimization problem:

Theorem 1. *Let $\mathbf{m}_0 \in \mathbb{R}^{R^2}$ be the initial saliency map, and let $\mathbf{\Lambda}$ and $\mathbf{L}$ be defined as above. The optimal saliency map $\mathbf{m}^* \in \mathbb{R}^{R^2}$ is the unique minimizer of*

$$J(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_0)^\top \mathbf{\Lambda} (\mathbf{m} - \mathbf{m}_0) + \lambda\, \mathbf{m}^\top L\, \mathbf{m},$$

with the solution

$$\mathbf{m}^* = (\mathbf{\Lambda} + \lambda\, L)^{-1}\, \mathbf{\Lambda}\, \mathbf{m}_0.$$

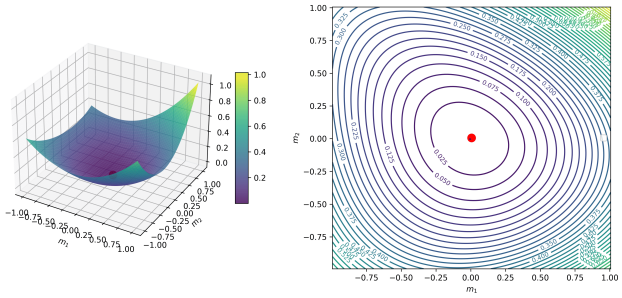

Figure 5. Illustration of the convex objective $J(\mathbf{m})$ in a 2D slice of the higher-dimensional space. The single global minimum, marked in red, highlights the function's convex nature.

Detailed proof is provided in Appendix 9.

The refined saliency maps $\mathbf{m}^{*\text{src}}$ and $\mathbf{m}^{*\text{tgt}}$ are then reshaped back to M^{*src} and M^{*tgt}, which are then used to obtain M^* by taking the element-wise maximum of the two maps:

$$M^* = \max\{M^*_{\text{src}}, M^*_{\text{tgt}}\},$$

Thresholding M^* with δ gives the final spatial mask M. An optimized CASA graph enforces a smooth, spatially consistent mask that preserves high-confidence regions and mitigates over-editing in less reliable areas.

Moreover, to maintain background consistency and prevent unintended changes outside the editing region, this optimized mask is used to replace the target branch's latent representation at each denoising step:

$$\hat{\mathbf{z}}^{\text{tgt}}_{t-1} = M \odot \mathbf{z}^{\text{tgt}}_{t-1} \; + \; (1 - M) \odot \mathbf{z}^{\text{src}}_{t-1}, \quad t = T, \ldots, 1.$$

Here, $\odot$ denotes Hadamard product. This step ensures that the background and non-editable regions of the source image remain unchanged throughout the iterative denoising process.

5. Experiments

5.1. Dataset and Evaluation metrics

We follow recent work [15, 43, 48] and evaluate our approach using the PIE-Bench dataset [15], which is currently the only established benchmark designed for prompt-based image editing. PIE-Bench contains 700 images categorized into ten different editing tasks, with each image accompanied by a source prompt, a target prompt, blend words (i.e., terms that specify the required edits), and an editing mask. Although only the source prompt, target prompt, and blend words are necessary for performing prompt-based editing, the editing mask is employed to gauge how well the method preserves the background.

To thoroughly assess our models, we adopt the evaluation strategy described in [15], focusing on three main criteria: 1) *Structure consistency*, measured by the difference

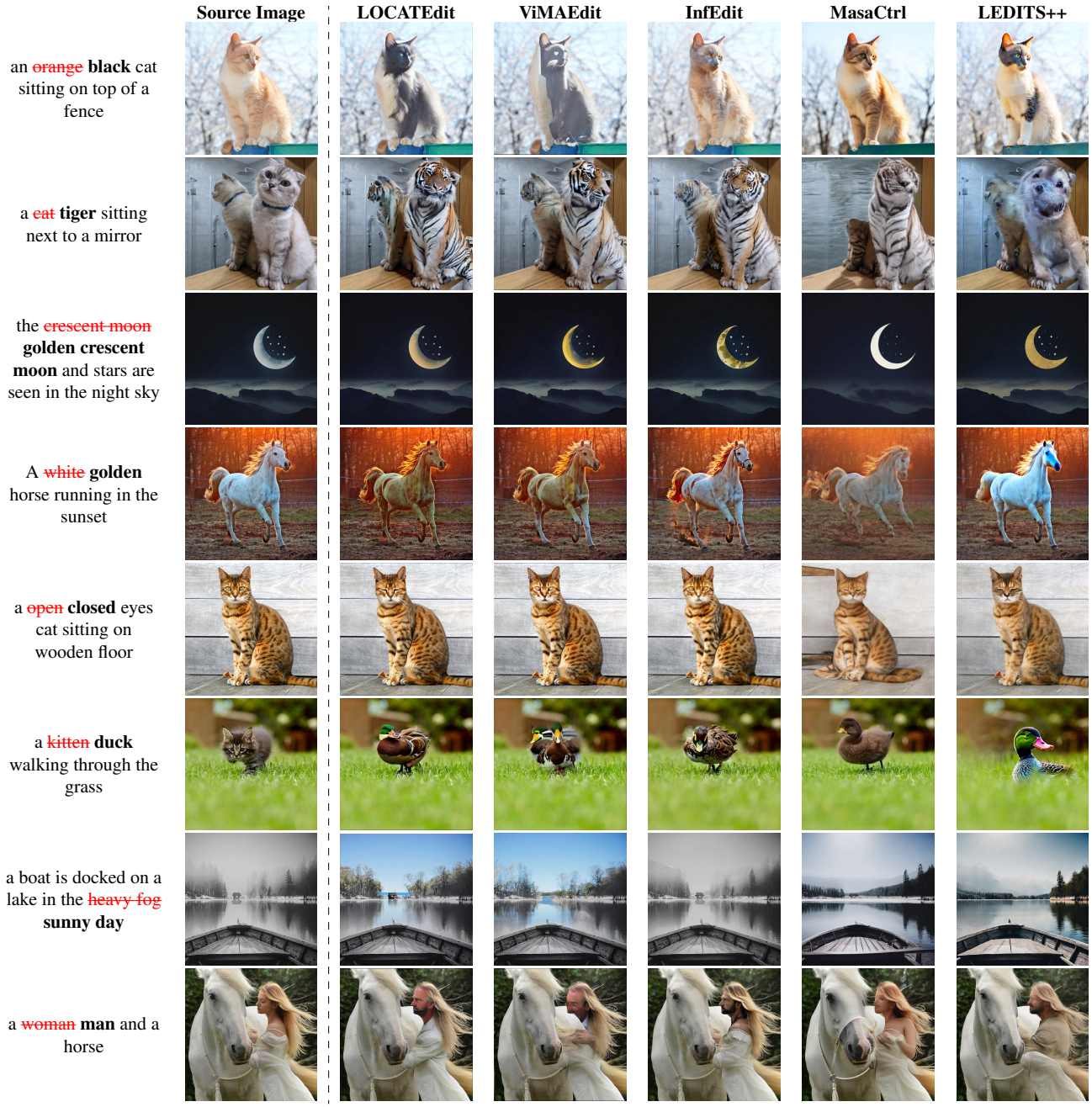

Table 1. Qualitative comparisons with competing text-guided editing methods. LOCATEdit yields more localized edits while preserving overall structure, outperforming baselines in both fidelity and consistency.

in DINO self-similarity maps [39], 2) *Background preservation*, evaluated via PSNR, LPIPS [52], MSE, and SSIM [45], and 3) *Target prompt–image alignment*, determined by CLIP similarity [10].

5.2. Comparison with existing methods

LOCATEdit consistently yields superior spatial consistency and semantic alignment compared to state-of-the-art text-guided image editing methods as can be seen in Table 2. Unlike P2P-Zero [29] and PnP-based techniques [15, 39], which tend to induce global modifications and suffer from spatial inconsistencies, LOCATEdit confines edits to intended regions, thereby preserving the source image's structure. Mask-guided methods such as ViMAEdit [43] improve localization but can still introduce artifacts in non-target areas. Our graph Laplacian regularization refines

Method		Editing	Structure	Background Preservation				CLIP Similarity	
Inverse	**Sampling (steps)**		**Distance** $\downarrow \times 10^3$	**PSNR** $\uparrow$	**LPIPS** $\downarrow \times 10^3$	**MSE** $\downarrow \times 10^4$	**SSIM** $\uparrow \times 10^2$	**Whole** $\uparrow$	**Edited** $\uparrow$
VI	DDCM(12)	InfEdit	13.78	28.51	47.58	32.09	85.66	25.03	22.22
VI	DDIM(50)	ViMAEdit	12.65	28.27	44.67	30.29	85.65	25.91	22.96
PnP-I	DDIM(50)	P2P-Zero	51.13	21.23	143.87	135.00	77.23	23.36	21.03
		MasaCtrl	24.47	22.78	87.38	79.91	81.36	24.42	21.38
		PnP	24.29	22.64	106.06	80.45	79.68	25.41	22.62
		P2P	<u>11.64</u>	27.19	54.44	33.15	84.71	25.03	22.13
		ViMAEdit	11.90	28.75	43.07	28.85	85.95	25.43	22.40
		LOCATEdit (Ours)	13.19	**29.20**	<u>41.60</u>	<u>26.90</u>	**86.53**	<u>25.96</u>	**23.02**
EF	DPM-Solver++(20)	LEDITS++	23.15	24.67	80.79	118.56	81.55	25.01	22.09
		P2P	14.52	27.05	50.72	37.48	84.97	25.36	22.43
		ViMAEdit	14.16	28.12	45.62	33.56	85.61	25.51	22.56
		LOCATEdit (Ours)	**8.71**	<u>29.16</u>	**39.31**	**24.01**	<u>86.52</u>	**26.07**	22.43

Table 2. Comparison of different methods based on structure, background preservation, and CLIP similarity metrics.

Method	Structure	Background Preservation				CLIP Similarity	
	Distance $\downarrow \times 10^3$	**PSNR** $\uparrow$	**LPIPS** $\downarrow \times 10^3$	**MSE** $\downarrow \times 10^4$	**SSIM** $\uparrow \times 10^2$	**Whole** $\uparrow$	**Edited** $\uparrow$
LOCATEdit	13.19	29.20	41.60	26.90	86.53	**25.96**	**23.02**
w/o diagonal weighting matrix	8.68	**29.59**	**38.16**	**23.17**	**86.83**	25.33	22.34
w/o symmetric self-attention	**8.59**	29.42	38.75	24.09	85.66	25.29	22.26
w/o α-based control	8.86	29.26	38.96	24.53	86.60	25.31	22.28
with high α	20.37	24.27	87.37	60.82	82.22	26.58	23.22

Table 3. Comparison of different methods based on structure, background preservation, and CLIP similarity metrics.

cross-attention maps by enforcing smooth, coherent patch-to-patch relationships, addressing these issues directly.

Moreover, while approaches like Edit-Friendly Inversion [14] and InfEdit with Virtual Inversion [48] achieve better semantic alignment, they often struggle to disentangle editable regions from the preserved background. In contrast, our method robustly separates these regions, ensuring that modifications are both precise and localized, which can be seen in qualitative comparision we provide in Table 1.

Overall, our experiments demonstrate that our method not only enhances the fidelity of the edited regions but also maintains the overall structural integrity of the source image. By leveraging CASA graph-based attention refinement, our approach outperforms existing techniques across multiple metrics, underscoring the importance of spatially consistent and disentangled editing for practical text-controlled image editing applications.

5.3. Ablation Study

To demonstrate the effectiveness of our model, we provide results for three different model ablations: 1) **w/o diagonal weighting matrix**: we use uniform L^2 penalty as the first term of Equation 6, 2) **w/o symmetric self-attention**: We do not parameterize the similarity matrix $\mathbf{S}$ which is essential for the Laplacian to be positive semidefinite, and 3) **w/o α-based control**: We keep $\alpha = 1$ in Equation 4.

Table 3 shows that each of our contribution outperforms the baselines in terms of Structure and Background preservation. We also observe that while combining different techniques together results in a slightly worser results in Structure and Background Similarity metrics, we are able to achieve state-of-the-art CLIP Similarity. It is to be noted that even the worser results are better than all the baseline methods reported in Table 2. Finally, when we were tuning the α parameter, we observed that a higher value of α edits images with way better CLIP similarity but significantly worsens the results for other metrics. This is to be expected because a high α results in "hard thresholding" where it makes a clear distinction between areas that should be trusted and those that should be adjusted, but it also leads to abrupt transitions.

6. Conclusion

In this paper, we introduced a text-controlled image editing framework LOCATEdit that refines cross-attention masks using graph Laplacian regularization. It leverages self-attention-derived patch relationships to enforce spatial consistency and localized, disentangled modifications while preserving the structural integrity of the source image. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods in semantic alignment and background fidelity. By confining edits to intended regions, our technique avoids unwanted alterations and maintains overall coherence. Future work will extend this framework to non-symmetric regularization and more complex editing scenarios, further enhancing controllable image generation.

References

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6):1373–1396, 2003. 2

[2] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8861–8870, 2024. 2

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18392–18402, 2023. 3

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22560–22570, 2023. 2, 3

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG), 42(4):1–10, 2023. 5

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427, 2022. 2

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022. 3

[8] Yifan Gao, Jinpeng Lin, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. Textpainter: Multimodal text image generation with visual-harmony and text-comprehension for poster design. In Proceedings of the 31st ACM International Conference on Multimedia, pages 7236–7246, 2023. 2

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022. 2, 3, 4, 5

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021. 7

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020. 3

[12] Nisha Huang, Weiming Dong, Yuxin Zhang, Fan Tang, Ronghui Li, Chongyang Ma, Xiu Li, and Changsheng Xu. Creativesynth: Creative blending and synthesis of visual arts based on multimodal diffusion. arXiv preprint arXiv:2401.14066, 2024. 2

[13] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. arXiv preprint arXiv:2402.17525, 2024. 2

[14] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12469–12478, 2024. 8

[15] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In The Twelfth International Conference on Learning Representations, 2024. 2, 6, 7

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2426–2435, 2022. 2

[17] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. In Proceedings of the 28th international conference on intelligent user interfaces, pages 919–933, 2023. 2

[18] Eunseo Koh, Sangeek Hyun, MinKyu Lee, Jiwoo Chung, and Jae-Pil Heo. Structure-preserving text-based editing for few-step diffusion models. 2

[19] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264, 2022. 2

[20] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Icml, page 3. Williamstown, MA, 2001. 3

[21] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. IEEE transactions on pattern analysis and machine intelligence, 30(2):228–242, 2007. 3

[22] Jungbin Lim, Jihwan Kim, Yonghyeon Lee, Cheongjae Jang, and Frank C Park. Graph geometry-preserving autoencoders. In Forty-first International Conference on Machine Learning, 2024. 2

[23] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7817–7826, 2024. 2

[24] Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16410–16419, 2022. 5

[25] Abdullah al Nomaan Nafi, Md Alamgir Hossain, Rakib Hossain Rifat, Md Mahabub Uz Zaman, Md Manjurul Ahsan, and Shivakumar Raman. Diffusion-based approaches in medical image generation and analysis. arXiv preprint arXiv:2412.16860, 2024. 2

[26] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and

Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021. 3

[27] Jiahao Pang and Gene Cheung. Graph laplacian regularization for image denoising: Analysis in the continuous domain. IEEE Transactions on Image Processing, 26(4):1770–1785, 2017. 2

[28] Jungin Park, Jiyoung Lee, Sangryul Jeon, Seungryong Kim, and Kwanghoon Sohn. Graph regularization network with semantic affinity for weakly-supervised temporal action localization. In 2019 IEEE International conference on image processing (ICIP), pages 3701–3705. IEEE, 2019. 3

[29] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In ACM SIGGRAPH 2023 Conference Proceedings, pages 1–11, 2023. 7

[30] 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 Proceedings of the 38th International Conference on Machine Learning, pages 8748–8763. PMLR, 2021. 2

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1 (2):3, 2022. 3

[32] Hareesh Ravi, Sachin Kelkar, Midhun Harikumar, and Ajinkya Kale. Preditor: Text guided image editing with diffusion prior. arXiv preprint arXiv:2302.07979, 2023. 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022. 3

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35:36479–36494, 2022. 3

[35] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9370–9379, 2024. 2

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020. 3

[37] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. Laplacian surface editing. In Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing, pages 175–184, 2004. 2

[38] Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: create your artwork with versatile control on text-to-image generation. In European Conference on Computer Vision, pages 92–109. Springer, 2024. 2

[39] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10748–10757, 2022. 7

[40] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1921–1930, 2023. 3

[41] Shinji Uchinoura and Takio Kurita. Graph laplacian regularization based on the differences of neighboring pixels for conditional convolutions for instance segmentation. In 2022 26th International Conference on Pattern Recognition (ICPR), pages 3611–3617. IEEE, 2022. 2

[42] Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3d meddiffusion: A 3d medical diffusion model for controllable and high-quality medical image generation. arXiv preprint arXiv:2412.13059, 2024. 2

[43] Kejie Wang, Xuemeng Song, Meng Liu, Jin Yuan, and Weili Guan. Vision-guided and mask-enhanced adaptive denoising for prompt-based image editing. arXiv preprint arXiv:2410.10496, 2024. 2, 3, 4, 6, 7

[44] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path. arXiv preprint arXiv:2303.16765, 2023. 2

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004. 7

[46] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 7677–7689, 2023. 2

[47] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1900–1910, 2023. 3

[48] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9452–9461, 2024. 6, 8

[49] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. Advances in Neural Information Processing Systems, 36:26291–26303, 2023. 2, 3, 5

[50] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023. 4

[51] Jin Zeng, Jiahao Pang, Wenxiu Sun, and Gene Cheung. Deep graph laplacian regularization for robust denoising of

real images. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops, pages 0–0, 2019. 2

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 7

LOCATEdit: Graph Laplacian Optimized Cross Attention for Localized Text-Guided Image Editing

Supplementary Material

7. Broader Impact

Our work advances the precision of text-guided image editing by ensuring that modifications are both spatially consistent and semantically faithful. This improvement has the potential to benefit a wide range of applications—from enhancing creative workflows in digital art and advertising to supporting critical tasks in medical imaging and scientific visualization—by reducing the need for extensive manual post-processing. At the same time, the increased reliability of automated editing tools underscores the importance of establishing robust ethical guidelines for their use, particularly in contexts where the authenticity of visual information is paramount. By delivering a method that better preserves the structural integrity of the source images, our approach paves the way for more trustworthy and accessible image editing solutions that can democratize creative technologies and support various high-stakes applications.

8. Proof of Lemma 1

Proof. To prove that $\mathbf{L}$ is PSD, we must show that for any $\mathbf{x} \in \mathbb{R}^n$, the quadratic form $\mathbf{x}^\top \mathbf{L}\mathbf{x}$ is nonnegative:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \mathbf{x}^\top (\mathbf{D} - \mathbf{S}_{\text{sym}})\mathbf{x}.$$

Expanding this expression, we have:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \mathbf{x}^\top \mathbf{D}\mathbf{x} - \mathbf{x}^\top \mathbf{S}_{\text{sym}}\mathbf{x}.$$

The degree matrix $\mathbf{D}$ is diagonal, with entries $\mathbf{D}(i,i) = \sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j)$. Therefore:

$$\mathbf{x}^\top \mathbf{D}\mathbf{x} = \sum_{i=1}^{n} \mathbf{D}(i,i)x_i^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j) \right) x_i^2.$$

The second term, $\mathbf{x}^\top \mathbf{S}_{\text{sym}}\mathbf{x}$, is given by:

$$\mathbf{x}^\top \mathbf{S}_{\text{sym}}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j)x_i x_j.$$

Substituting these into the quadratic form, we get:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j)x_i^2 \right) - \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j)x_i x_j.$$

Reorganizing terms:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j) \left(x_i^2 + x_j^2 - 2x_i x_j \right).$$

This simplifies to:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j)(x_i - x_j)^2.$$

Since $\mathbf{S}_{\text{sym}}(i,j) \geq 0$ (by definition of the symmetrized self-attention matrix) and $(x_i - x_j)^2 \geq 0$, every term in the summation is nonnegative. Therefore:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Thus, $\mathbf{L}$ is positive semidefinite. □

9. Proof of Theorem 1

9.1. Optimization Problem

We consider the following optimization problem:

$$\min_{x \in \mathbb{R}^{R^2}} J(x), \tag{6}$$

where the objective function is defined as

$$J(x) = (x - x^{(0)})^\top \mathbf{\Lambda}(x - x^{(0)}) + \lambda\, x^\top L\, x.$$

Here, the fidelity term $(x - x^{(0)})^\top \mathbf{\Lambda}(x - x^{(0)})$ penalizes deviations from the initial mask $x^{(0)}$ with stronger penalties in regions of higher confidence (as encoded by the diagonal weight matrix $\mathbf{\Lambda}$). The smoothness term $\lambda\, x^\top L\, x$ promotes a spatially coherent solution by enforcing that the mask varies smoothly across similar patches, as determined by the self-attention structure. The hyperparameter $\lambda > 0$ balances the trade-off between fidelity and smoothness.

9.2. Existence and Uniqueness of the Solution

To obtain the refined mask, we solve the minimization problem in Equation (6). The first term is strictly convex since $\mathbf{\Lambda}$ is positive definite, and the second term is convex because L is positive semidefinite. Thus, the overall objective $J(x)$ is strictly convex and has a unique minimizer.

Taking the gradient with respect to x yields:

$$\nabla J(x) = 2\, \mathbf{\Lambda}(x - x^{(0)}) + 2\lambda\, L\, x. \tag{7}$$

Setting $\nabla J(x) = 0$ gives:

$$\mathbf{\Lambda}(x - x^{(0)}) + \lambda\, L\, x = 0. \tag{8}$$

Rearranging, we obtain:

$$(\mathbf{\Lambda} + \lambda\, L)\, x = \mathbf{\Lambda}\, x^{(0)}. \tag{9}$$

Since $\mathbf{\Lambda} + \lambda L$ is positive definite, it is invertible, and the unique solution is

$$x^* = (\mathbf{\Lambda} + \lambda L)^{-1} \mathbf{\Lambda} x^{(0)}.$$

The positive semidefiniteness of L ensures the convexity of the regularization term, thereby guaranteeing the existence and uniqueness of the solution.

9.3. Additional Qualitative Results

This section presents qualitative results for refined masks achieved through graph Laplacian regularization and compares the editing outcomes with existing image editing methods.

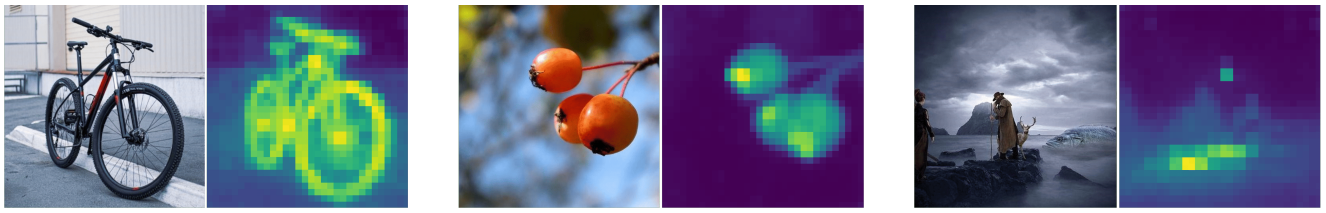

Figure 6. Refined masks after Graph Laplacian Regularization

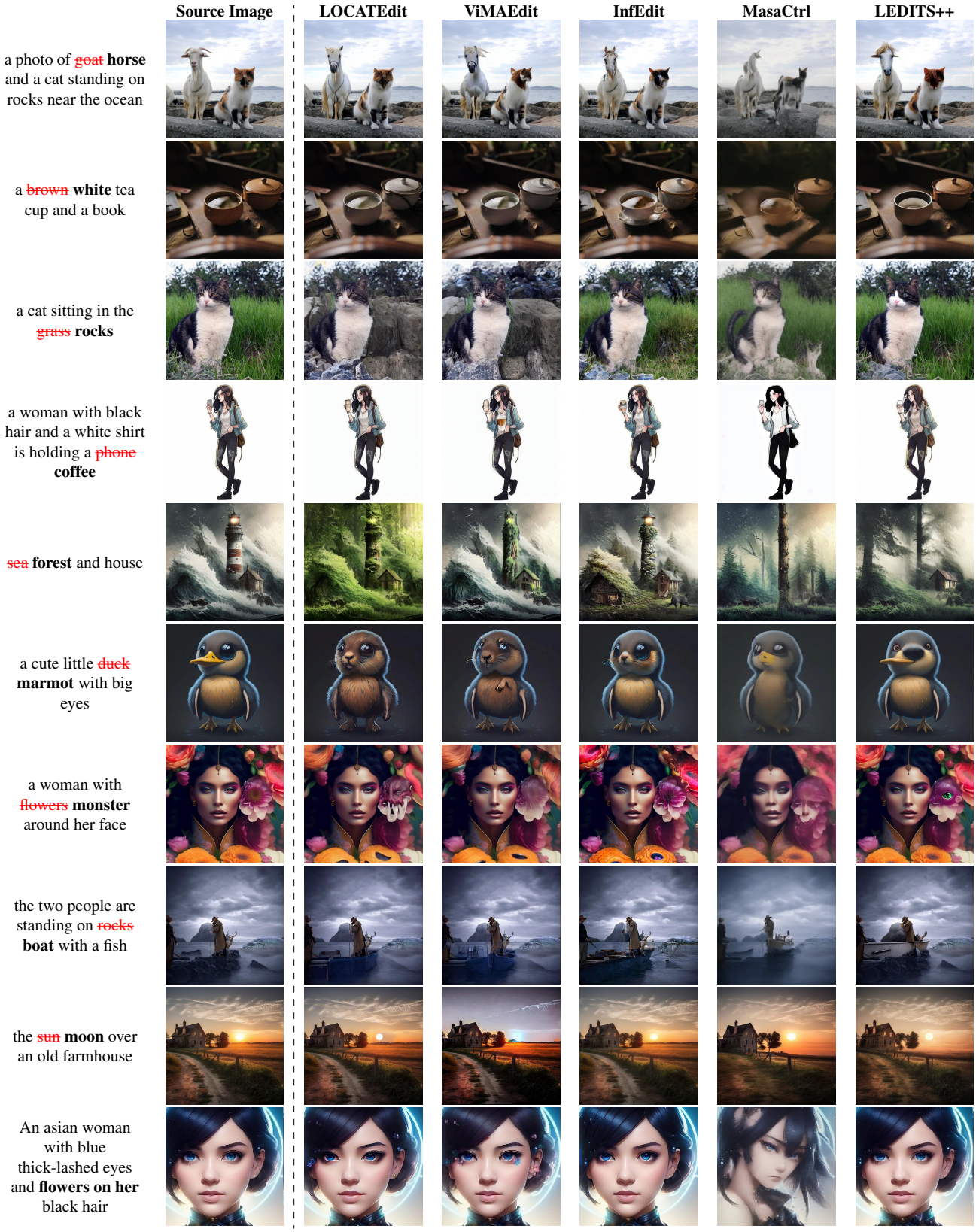

Table 4. Additional qualitative results on PIE-Bench