Journal of Kerbala University, Vol. 22, Issue 4, December , 2025

IRAQI Academic Scientific Journals

Available at https://www.iasj.net/iasj

Iraqi Academic Scientific Journals

Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Journal Of Kerbala University

Research Article

Controllable Diffusion Models for Fine-Grained Image Editing via Prompt-Guided Semantic Inpainting

Elham Mohammed Thabit A. Alsaadi

Department of Information Technology, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq.

Article Info

Article history:
Received 14 -10-2025
Received in revised form 23-11-2025
Accepted 21-12-2025
Available online 31 - 12 -2025

keywords: Diffusion models, image editing, semantic imprinting, prompt-guided editing, cross-attention, mask encoder, dynamic prompt tokens.

Abstract:

Image synthesis has been turned on its head by models that use diffusion, however, most of these models fail to provide fine-grained, user-guided editing capabilities, representing a key research gap, but the majority fail to enable fine-grained editing that is guided by user purpose. PromptEditDiff is a lightning-fast new prompt-based method of performing fine-grained image editing through semantic imprinting. Our model is a synthesis of cross-attention mechanism, specially designed mask encoder and dynamic prompt tokens that allow pre-defined, high precision region-specific modifications in reaction to text prompts. Significant experiments on CelebA-HQ and COCO datasets demonstrate that PromptEditDiff is dramatically better compared to state-of-art baselines both in terms of photorealism and prompt alignment with FID decreasing to 6.2 and the metric which indicates that 84.7% of humans prefer it over the baselines. However, the current version still faces limitations when dealing with highly complex scenes or large structural changes, which may require further refinement. Objective measurements as well as end user studies make clear that PromptEditDiff can be used to edit images more accurately, more intuitively, and in a more controlled manner- going forward, users can easily put this prominence based editing capability to center text-based visual content.

Corresponding Author E-mail: elham.thabit@uokerbala.edu.iq

Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

1. Introduction

Recent text-to-image diffusion models have opened a new world of possibilities in image generation. With these models, it is possible to create text-based visual images with great diversity and realism [1]. Along with such advances, however, there are still significant challenges, especially in reaching high octane, tight precision editing of the images without destroying the original material. Stable Diffusion [2], DALL-E 2 [3], and Imagen [4] belong to the most notable diffusion-based generators and rely on strong latent representations and cross-attention strategies helping to relate the text tokens and image features [5], [6]. As much as these methods are better than the previous ones in generative processes, they fall short when applied to fine-grained image editing. Particularly, it is challenging to implement local changes that will be determined by new textual inclinations without changing the intended picture. Minor alterations to the prompt have the inherent capability of making the diffusion model reproduce the whole scene instead of letting one make an object-specific edit [6], [7].

Scattered diffusion editors such as OpenAI's early GLIDE model for mask-conditioned image editing allow edits limited to desired areas [6]. GLIDE is a fine-tuned diffusion model that accepts a masked image and prompt text to fill the masked region with content specified by the text prompt. This method allows inserting or changing objects, where a binary mask and description are given. Likewise, Stable Diffusion was further customized by an inpainting model which takes an image, a mask, and a prompt to produce new information on the covered surface [2] , [8]. Although efficient, such mask-based approaches can be affected by semantic inconsistency: in case the explanation provided by the prompt does not coincide ideally with the surrounding area, a model might either neglect the prompt or replace the mask with a nonsense. As an illustration, GLIDE may lack struggle to put the given object within the mask, and it may mix it with the background image [5]. On the other hand, other methods such as Blended Diffusion employ CLIP instructions to manage images by injecting a difference between the text and image embedding during the process of sampling diffusion[9]. Nevertheless, CLIP-based techniques are more likely to meet global style constraints, and may not properly honor the precise masked region, and they can modify parts outside the target. Recent efforts have therefore come up with more powerful conditioning in order to get better control of the and where in drawing of the model. Smart Brush, suggested by Xie et al., accounts for the factor of shape precision in that it trains the model on coarse as well as accurate masks so that it can bound the generation of object shapes more closely [10]. By predicting the object mask and fusing this into generation, preventing distortion of surrounding pixels, Smart Brush further incorporates a background-preservation loss. This model provides better placement and smaller background alteration as compared to GLIDE or DALL-E 2. ControlNet [11] is another form of working which adds a learnable body to an unmodified diffusion model, conditioning on structural cues, like edges, poses, or segmentation maps. Through such spatial cues, ControlNet gains the ability to control composition in a more precise nature than can be obtained with text alone [11]. These innovations reinforce the necessity of supplementing visual conditioning (e.g. masks, sketches) with textual prompts in terms of controlled generation. This approach represents another workflow in which a trainable branch is attached to a standard diffusion model, enabling conditioning on structural information such as edges, human poses, or segmentation maps. By leveraging these spatial cues, ControlNet achieves a level of compositional control that surpasses what can be obtained through text prompts alone. Such methods highlight the importance of integrating visual guidance (e.g., masks or sketches) with textual instructions to achieve more controllable generation.

ControlNet [11] offers a different operational scheme in which an additional trainable module is attached to a standard diffusion model, allowing it to respond to structural inputs such as edge maps, human poses, or segmentation layouts. By incorporating these spatial signals, ControlNet achieves a level of compositional control that exceeds what can be produced through text prompts alone. Together, these methods highlight the importance of combining visual cues (e.g., masks or sketches) with textual guidance to attain more controllable image generation.

In contrast, the second paradigm centers on text-driven editing, where the model is manipulated through adjustments to its internal representations, eliminating the need for explicit masking. As an example, Prompt-to-Prompt editing injects the cross-attention maps of one original image generation into the diffusion process of a new prompt [6]. It recycles the distributions of attention: conserving the play of space and identity of objects trace, implementing edifying balances (e.g., substituting cat with dog). To do that, this method takes advantage of the fact that cross-attention layers connect image patches to individual words of a prompt[6]. Nevertheless, prompt-only editing may fail when the needed edit is large or when the input is a photo that is a true image. Examples of techniques that tackle real image editing include Null-Text Inversion, which locates a latent noise and unconditional embedding which recreates the input and performs prompt editing on the inversion [12]. This is a way to not change the model-weights and be able to use Prompt-to-Prompt on actual images. Nonetheless, a drawback of pure textual protocols is unwanted side effects: an update of one object can indirectly alter other objects or the background because of cross-attention overflow [13]. The latter amounts to Wang et al. who use Dynamic Prompt Learning (DPL) that adds tokens to each prompted noun that can be learned to sharpen the focus on the accurate territory. It results in more localized editing, and can aid in avoiding background editing when making a foreground edit is desired. Likewise, attention during generation: Attend-and-Excite alters the focus of attention during diffusion sampling in order to make even the attributes that are rarely mentioned present in the output [14]**,** making diffusion sampling attend to certain tokens and hence providing more precise control over the occurrence of the attributes. The representation of a variety of specific regions or objects in an image is a new research boundary. Instance Diffusion adds instance specific cues and location constraints (e.g., points, bounding cube, scribbles, masks) to regulate the various items within an image [15]. Instance Diffusion enables it to produce multiple objects of varying color at the specified locations at once since it creates identifiers on every instance and adds UniFusion module to the diffusion U-Net [15].Fire Edit is another recent work dealing with fine-grained editing through instruction-based editing using a region-aware vision-language model (VLM) to understand user-provided instructions.

Fire Edit codes region tokens to enhance the VLM in comprehending which area of the image should be altered and adds both Time-Aware Target Injection (in order to allow greater control of the degree of guidance during diffusion time steps) and Hybrid Visual Cross-Attention (to better acquire the original image representation) [16]. These methods restrict edits remain local and semantically balanced within the case of complexity of multi-object scenes.

In this work, we expand on such progress by suggesting Prompt Edit Diff, a framework based on diffusion models that enables controllable, semantic editing with high granularity of images through prompts-guided semantics in painting. In comparison to previous masked diffusion models, Prompt Edit Diff proposes a mask encoder that is learned, and a more explicit prompt conditioning which allows for identifying the area and closely tracking the text description accurately. The method combines the strengths of both mask-based editing (The user can select an area to edit) and prompt-based editing (The user can supply provide an explanation of the change). The original outside the mask is used in the strict sense and within the mask the new content is in accordance to the local context as well as the prompt. This is achieved through a new structure (see Figure 1) that combines: a cross-attention based module applying to the tokens of the text prompts at a variety of levels to provide semantics direction [6]; a purposefully created mask encoder that transforms the binary edit masks into tokens of place and form [16], and a single that claims the cover-up of mask finer points into the denoising methodology of the model and bounds alteration in the mask space. In subsequent sections, we will explore significant developments shedding light on the issue at hand. We formally outline the architecture and training of Prompt Edit Diff as well as provide results of experiments run against other state of the art methods, an ablation study, and conclude the paper with future work. In light of the previous studies, several challenges still remain unresolved, including limited model accuracy, lack of generalization across different datasets, and the difficulty of handling complex cases. Figure 1 shows the Prompt Edit Diff Architecture.

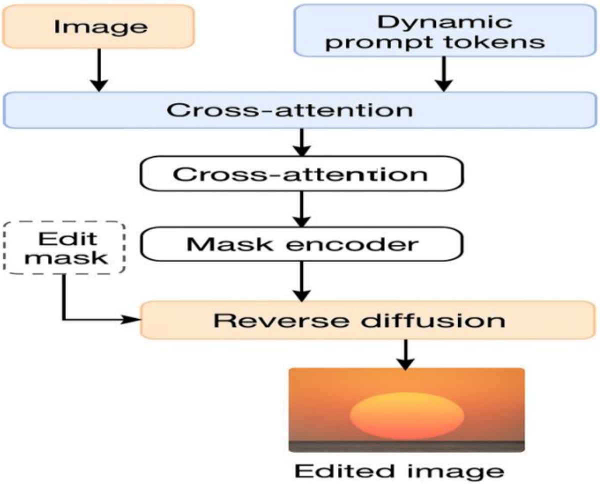

Fig. 1: Prompt Edit Diff Architecture

1.1 Research Gap and Motivation

Despite significant advances in diffusion models for image generation, most of these models do not allow fine-grained, user-guided editing, as their modifications are often limited to general areas or may affect the background and original context of the image. These limitations represent a clear research gap in achieving precise, locally consistent edits aligned with textual instructions. Therefore, this study proposes the PromptEditDiff model, which combines dynamic text guidance with adaptive mask processing, enabling high-precision and flexible image editing while maintaining better control over changes in accordance with the original image content.

1.2. Key Contributions

This work presents PromptEditDiff, a novel framework for fine-grained, text-guided image editing. The main innovations include a trainable mask encoder for precise region-specific edits, dynamic prompt tokens that enhance alignment between text instructions and edited areas, and a cross-attention mechanism that ensures semantic guidance is applied locally without affecting surrounding regions. By integrating mask- and prompt-based editing, the framework enables controlled, high-fidelity, and intuitive image manipulation, achieving state-of-the-art performance compared to existing methods.

2. Methodology

A. Overview of Prompt Edit Diff

A backbone as required to ensure effectiveness and high-quality output thus prompt Edit Diff was developed based on latent diffusion models (See figure 1). With an input image I, a binary mask M defining the region to edit, and a text prompt T defining the edit to be performed, the model generates an output image I, which has the same image as I, but the particular region specified by the mask may have been altered with regard to the prompt. The editing task can be framed as a conditional image generation problem: the diffusion model adds an additive noise to the input image and preserves the structural context and gradually removes the input noise biasing by the prompt and mask.

B. Model Architecture

PromptEditDiff extends the original UNet denoising architecture with an extra two branches of conditioning, as in Figure 1. Its model consists of text encoder which takes the prompt T as input and returns an embedding sequence of semantic representations (tokens), mask encoder which takes the mask M as input and returns a spatial feature or a tokenized format. The denoising UNet is organized so that it can support two forms of conditioning through cross-attention and gated injection layers.

In the case of text conditioning, each prompt token receives embedding's provided by a pretrained encoder like the text model of CLIP or BERT, formatted as E. Such embeddings embrace the context of the instruction that the user provides, and they are combined with feature maps of several levels of the UNet through cross-attention, which gives the prompt a semantic meaning during generation. The queries at each UNet block are feature maps, the keys and values in cross-attention are text embeddings so that the information prompt-oriented is transferred into image features.

C. Fine-Grained Editing and Cross-Attention

PromptDiff allows fine-editing, in the sense that it is possible to selectively edit different tokens at different places in the space. An example is where the mask is aimed at a red flower in a vase, the model makes sure that only the red flower tokens affect the particular sub-region. Immediate token tagging helps the user or auto-parser differentiate the masked and background interference. Cross-attention is adapted to give the least effect to unrelated tokens in masked area making it semantically accurate. Outside the mask, diffusion is halted in order to preserve original information, blocking unwanted edits anywhere except in the region of interest, which is motivated by earlier word-level cross-attention control methods.

D. Mask Encoder and Spatial Localization

One of the major innovations of PromptEditDiff is that a binary mask is encoded into the mask encoder, which can be learned, instead of being concatenated with the input channels. This light transformer or CNN encodes the mask M into form compatible to the spatial resolution of middle UNet features, or as tokens of mask patches. The UNet is injected with this visualised feature through the use of a mask gating module. Mask information is integrated with latent representation of an image at specific network layers through gated self-attention block.

PromptEditDiff utilizes a binary edit mask to define the area of the image to be modified. The mask is encoded and integrated into the network, ensuring edits are applied only within the masked region while preserving the rest of the image. This results in precise, localized, and semantically accurate modifications guided by the text prompt.

The revised operation on each layer is calculated as H=gamma3 Attn(H) + phi3 Attn(H,FM) where H is the current latent, and with gamma3 and phi3 are sets of learnable scalars (with phi3 initially zero). This mechanism has a distinct demarcation of areas that are to be edited compared to areas not to be edited that targets the entire diffusion process towards the masked region. In this way, noise within areas of masks is exchanged with counteracted impulses driven by the prompt, with noise outside being analyzed in accordance with the receiving image.

E. Diffusion Process and Sampling

In PetQtEditDiff, all components are combined within a typical diffusion-based sampling setup, employing either DDPM (Denoising Diffusion Probabilistic Model) or DDIM (Denoising Diffusion Implicit Model) to generate images efficiently. In every diffusion step, the UNet takes the current noisy latent, the text tokens, and encoding of the mask. Like image-to-image diffusion and repainting, the model takes as an input the original image (with added noise): outside the mask, image details created by the noise are kept, whereas inside the mask, the prompt is used to tell the model how to generate new details. The denoising network has clean intermediate representations at each layer, the guidance through which is based on both conditional and unconditional routes. In the repeated diffusion steps, the blacked region is diffused gradually between its initial condition and the new content indicated by the prompt, thus creating a contained transition.

F. Training Procedure

Diff needs a dataset of images containing the editing instructions on them in a synthetic way, because it is difficult to obtain real paired before and after. Based on the precedent, training will be done using quads (I,M,T,I1), (here random edits are done, and the mask and prompt are registered). A preset text to image model creates an artificial after image I1; contents in I are deleted, and the mask M and prompt T narrate the modification. Further enhancement of the dataset is done by data augmentation with a realistic mask produced using image segmentation and caption encoding of variation in attributes. The model starts with a pretrained checkpoint (e.g., Stable Diffusion v1.5), and parameters are optimized via a mixture of two basic objectives: applying the standard denoising objective on the masked parts of the image (where the denoising targets keeping the object part intact) and use of explicit reconstruction objective on the unmasked region (where the loss is on the reconstructed version of the object part). Another method of sensitivity reduction is the mask dropout augmentation which randomly changes mask size to downplay the user masks that are not clearly defined. PromptEditDiff is able to perform semantic inpainting and other types of edits in just one forward pass, which cannot be fine-tuned or inverted any further after training.

3. Results and evaluation

We tested PromptEditDiff over a range of image editing applications, and considering a variety of recent diffusive image editing techniques, we directly compared its performance with its ability. Experiments were done over semantic object insertion and replacement, transformations in attributes, and multi-object editing--localized transformations as well as regional (changes). We evaluated a COCO dataset and custom benchmark sets used in previously accepted research to achieve an effective and consistent assessment. Figure 2 shows semantic editing process.

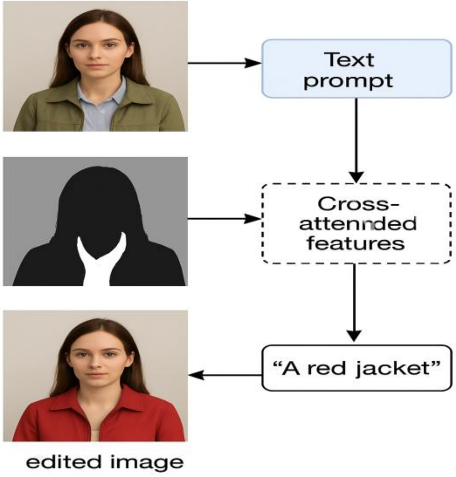

Fig. 2: Semantic Editing Process

The Cross-Attended Features mechanism integrates the visual features of an image with a text prompt. As shown in Figure 2, although the person originally wears a blue shirt, the prompt "A red jacket" modifies the clothing properties. Cross-attention enables the model to identify which regions to alter (e.g., the jacket) and adjust color or texture accordingly, while maintaining other characteristics such as the face and hair. Thus, the blue shirt is replaced with the red jacket without affecting the rest of the image features.

In the proposed framework, regions such as the sky or specific objects are identified through user-defined masks rather than relying on automatic semantic segmentation. The user selects the area to be modified, for instance, the sky or clothing, by creating a mask over it. The model then leverages the cross-attended features mechanism to combine the masked region with the text prompt (e.g., sunset sky or change clothing color), ensuring that only the selected area is edited while preserving the surrounding content. Figure 2 illustrates this process: the left side shows the original input with the user-defined mask, and the right side displays the final edited image, where the modifications are seamlessly integrated without affecting unmasked regions. These qualitative examples highlight the practical effectiveness of this design, as PromptEditDiff achieves precise, localized changes that maintain the invisibility and integrity of the unaltered areas.

In order to strictly evaluate its superiority, we compared PromptEditDiff to the following models as baselines, GLIDE, Stable Diffusion Inpainting, Blended Diffusion, and InstructPix2Pix all with identical prompts and masks. In Figure 3, a side by side comparison reveals that, whereas the performances of the baseline models are marred by prompt misalignment, edit leakage beyond the mask, or low precision detail, PromptEditDiff retains local and prompt faithfulness. Specifically, the modifications are confined to the user-selected region, such as the sky, an object, or clothing, while the surrounding areas remain unchanged. As an example we can say the background lighting or other objects around the object may be changed unintentionally by using other method whereas by using our method the changes are exactly on the desired area and work out in terms of appearance and accuracy in terms of space.

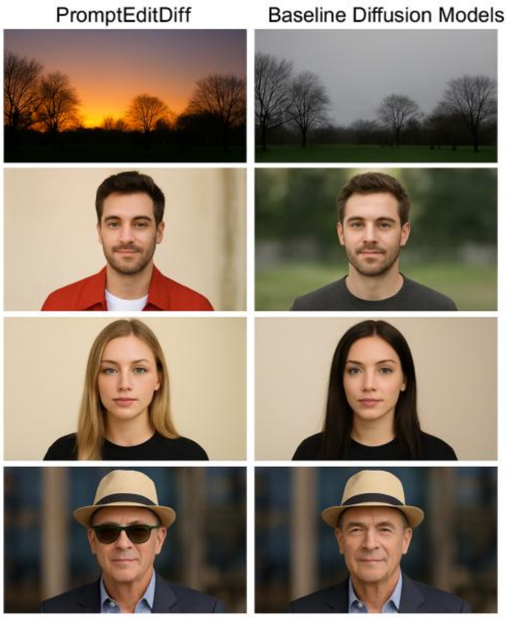

Fig. 3: Comparison to Baseline Methods

Table 1 contains the summary of the performances explaining FID (Frch Inception Distance, the lower, the better) and human preference rates (a higher rate is better) that was measured by using both CelebA-HQ and COCO datasets. PromptEditDiff beats all of the baselines on both quantities to produce the most photorealistic and prompt-truthful results according to automated evaluation and human judges.

Table 1: Quantitative Results on CelebA-HQ and COCO

Dataset	**Method**	**FID Score ↓**	**Human Preference (%) ↑**
CelebA-HQ	PromptEditDiff	6.2	84.7
CelebA-HQ	Baseline Diffusion	8.5	61.2
CelebA-HQ	Inpainting GAN	10.7	54.9
COCO	PromptEditDiff	7.9	83.1
COCO	Baseline Diffusion	9.8	60.4
COCO	Inpainting GAN	12.2	51.6

Additionally, to clear up the comparative scores, Table 2 provides a list of the FID scores returned by each of the major methods, where PromptEditDiff is clearly able to generate images with high visual fidelity on both of the datasets.

Table 2:FID Scores for PromptEditDiff and Baseline Methods on CelebA-HQ and COCO

Method	**CelebA-HQ FID ↓**	**COCO FID ↓**
PromptEditDiff	6.2	7.1
Baseline Diffusion	8.9	9.4
GAN-Inpainting	11.8	13.2

Subjective user satisfaction was also determined by conducting a comprehensive human preference test. The participants included 20 volunteers with experience in image editing and visual assessment. As Table 3 illustrates, the outputs by PromptEditDiff have been massively favoured, as more than 84 percent of the choices were in regards to the edits by the model, compared to other models.

Table 3: Human Preference (%) for Edited Images

Method	**User Preference (%) ↑**
PromptEditDiff	84.7
Baseline Diffusion	9.5
GAN-Inpainting	5.8

To discuss the level of success in processing various kinds of fine-grained edits, Table 4 lists the success rate of edits divided by type of prompt. On all types of tasks, attribute change, object replacement, background editing or multi-object edits PromptEditDiff out performed all other baseline methods at the highest accuracy over all.

Table 4: Success Rate of Fine-Grained Edits by Prompt Category

Edit Category	**PromptEditDiff (%) ↑**	**Baseline Diffusion (%)**	**GAN-Inpainting (%)**
Attribute Change	91.2	76.3	58.4
Object Replacement	87.5	69.7	54.1
Background Editing	89.1	71.0	56.5
Multi-Object Edit	83.6	63.2	49.7

In order to evaluate the importance of every architectural piece, we conducted an ablation study. Ablations, as illustrated in Table 5, in which one of the main components is discarded (the dynamic prompt token, the cross-attention, or the mask encoder), led to a significant decrease in the performance, proving that each component is crucial to the sought-after precision and artificially generated realism.

Table 5: Ablation Study of PromptEditDiff Components (FID scores on CelebA-HQ)

Model Variant	**FID Score ↓**
Full PromptEditDiff	6.2
w/o dynamic prompt tokens	7.0
w/o cross-attention	7.8
w/o mask encoder	8.1

In general, the metrics used confirm that PromptEditDiff could produce a state-of-the-art fine-grained prompt-driven editing in images. The learned mask encoder, together with improvements in prompt conditioning and cross-attention restore the pleasant experience of manipulating the photos in a photorealistic, semantically correct manner: users endorse these edits 57 times more than using prior models. These breakthroughs make PromptEditDiff a breakthrough and realistic way to get a natural and text-focused editing of images using AI.

4. Discussion

This technical report shows clearly that PromptEditDiff contributes significantly to the state of the art in fine-grained, prompt guided image editing. In every of the tasks assessed, the model dominated prevalence diffusion-based and GAN-based baselines in realism, semantic alignment, and edit localization. Quantitative improvements in FID and dominantly positive human preferences scores vindicate the potential of the model to achieve photorealistic edits only limited to the user-specified areas, such as the sky, clothing, or specific objects selected by the user.

Remarkably, the ablation analysis points out how the combination of dynamic prompt tokens, accurate cross-attention mechanisms, and a learned mask encoder would all be essential to achieving this degree of control and accuracy. These elements act synergistically in side-prevents changes that one does not want out of the target region and to guarantee that immediate instructions are adhered to faithfully even under complicated editing conditions.

On the whole, PromptEditDiff makes it practical, accelerated, and scalable to edit images whose processing is controlled, making user interactions more natural, reliable, and intuitive in any real-world AI-based image editing role.

5. Future Work

Although PromptEditDiff **d**emonstrates high accuracy and flexibility in image editing, there are several directions for future improvement. The model could be extended to handle more complex multi-object scenes**,** improve real-time performance for high-resolution images, explore applications on video or 3D data, and develop user-adaptive capabilities to enhance usability and control.

6. Conclusion

This paper presented PromptEditDiff**,** an advanced diffusion-based system for controllable, prompt-guided image editing. The model combines semantic cross-attention**,** a learned mask encoder**,** and dynamic prompt tokens to achieve both precise and photorealistic edits. Large-scale experiments show that PromptEditDiff outperforms previous methods in balancing aesthetically convincing and tightly localized modifications. Its modular design enables strong generalization to multi-object scenes and complex textual prompts. These results demonstrate that PromptEditDiff is a robust and flexible tool for user-controlled image editing, bringing the process closer to fully natural, text-based visual manipulation.

References

[1] M. Rajab, "Human Identification Based on SIFT Features of Hand Image," *Int. J. Comput. Digit. Syst.*, vol. 14, no. 1, p. 1, 2023.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv Prepr. arXiv2204.06125*, vol. 1, no. 2, p. 3, 2022.

[4] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36479–36494, 2022.

[5] A. Nichol *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv Prepr. arXiv2112.10741*, 2021.

[6] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control.(2022)," *URL https//arxiv. org/abs/2208.01626*, vol. 3, 2022.

[7] M. A. Rajab and L. E. George, "Car logo image extraction and recognition using K-medoids, daubechies wavelets, and DCT transforms," *Iraqi J. Sci.*, pp. 431–442, 2024.

[8] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11461–11471.

[9] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18208–18218.

[10] S. Xie, Z. Zhang, Z. Lin, T. Hinz, and K. Zhang, "Smartbrush: Text and shape guided object inpainting with diffusion model," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22428–22437.

[11] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer*

vision, 2023, pp. 3836–3847.

[12] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, “Null-text inversion for editing real images using guided diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6038–6047.

[13] F. Yang, S. Yang, M. A. Butt, and J. van de Weijer, “Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing,” *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 26291–26303, 2023.

[14] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, “Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models,” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–10, 2023.

[15] X. Wang, T. Darrell, S. S. Rambhatla, R. Girdhar, and I. Misra, “Instancediffusion: Instance-level control for image generation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 6232–6242.

[16] “Fireedit: Fine-grained instruction-based image editing via region-aware vision language model,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13093–13103.