

# REGION IN CONTEXT: TEXT-CONDITIONED IMAGE EDITING WITH HUMAN- LIKE SEMANTIC REASONING

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Recent research has made significant progress in localizing and editing image regions based on text. However, most approaches treat these regions in isolation, relying solely on local cues without accounting for how each part contributes to the overall visual and semantic composition. This often results in inconsistent edits, unnatural transitions, or loss of coherence across the image. In this work, we propose Region in Context, a novel framework for text-conditioned image editing that performs multilevel semantic alignment between vision and language, inspired by the human ability to reason about edits in relation to the whole scene. Our method encourages each region to understand its role within the global image context, enabling precise and harmonized changes. At its core, the framework introduces a dual-level guidance mechanism: regions are represented with full-image context and aligned with detailed region-level descriptions, while the entire image is simultaneously matched to a comprehensive scene-level description generated by a large vision-language model. These descriptions serve as explicit verbal references of the intended content, guiding both local modifications and global structure. Experiments show that it produces more coherent and instruction-aligned results. Code is available at: <https://anonymous.4open.science/r/Region-in-Context-F7F1/README.md>

## 1 INTRODUCTION

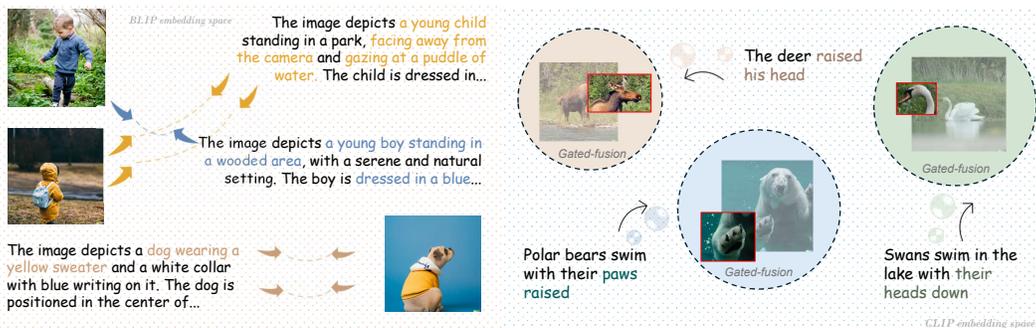
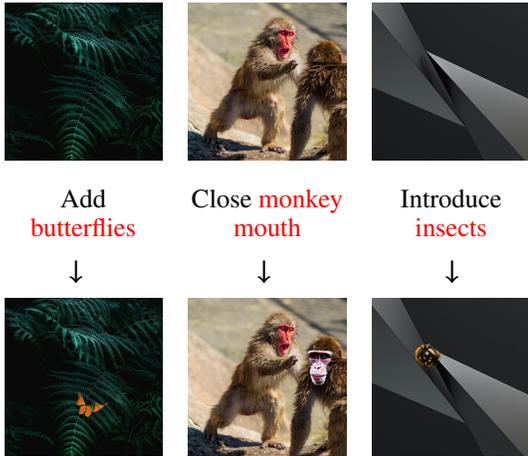


Figure 1: Illustration of our vision-language embedding optimization strategy. At the global level (left), the model aligns image and full target description embeddings. At the local level (right), it aligns the region in the full image to make the region “see itself in context”, with their corresponding region descriptions using contrastive learning.

Just as solving a puzzle requires understanding the context of the entire image to determine where each piece belongs, many human visual tasks involve reasoning about the whole before acting on the parts. We do not place a puzzle piece based solely on its shape or color in isolation; instead, we look at the full picture to infer its correct position. This process reflects a broader principle in visual understanding: local decisions are guided by global context. Whether assem-

054 bling a puzzle or arranging elements in a design, humans rely on a mental model of the com-  
 055 plete scene to ensure that each part contributes meaningfully and harmoniously to the whole. This  
 056 same principle applies to the task of image editing. When modifying a specific region within an  
 057 image—such as changing the color of an object or altering its appearance—humans rarely con-  
 058 sider that region in isolation. Instead, we take into account the broader scene to ensure that  
 059 the edit fits naturally, both visually and semantically, within the overall composition as Fig. 1.



061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076 Figure 2: Illustration of editing results using the  
 077 segmentation mask from ZONE Li et al. (2024),  
 078 which first segments the editing region, then ed-  
 079 its and blends it in the original image. Although  
 080 the method includes a blending step during post-  
 081 processing, the edited images still exhibit unnatu-  
 082 ral appearances.

083 (2024). This close bond has inspired growing interest in integrating language understanding into vi-  
 084 sual generation systems, particularly for tasks like text-conditioned image editing. Fu et al. (2024);  
 085 Wu et al. (2024a). Broadly, existing approaches integrate LLMs in two main ways. First, several  
 086 works use LLMs to improve dataset construction by refining or generating detailed editing instruc-  
 087 tions and compositional prompts Brooks et al. (2023); Bai et al. (2025); Fu et al. (2024). Second,  
 088 a growing line of research embeds LLMs directly within the editing pipeline to act as reasoning  
 089 agents or feedback mechanisms. Notably, SLD Wu et al. (2024a) introduces an iterative feedback  
 090 loop where the LLM evaluates the consistency of the generated image with the input prompt and  
 091 issues correctional instructions to the diffusion model.

092 With these inspirations, in this paper, we propose a framework for image editing (Sec. 3.2) that  
 093 leverages language not only as an instruction medium but also as a semantic reference for global  
 094 guidance. Specifically, we introduce a dual-level alignment mechanism that encourages each edited  
 095 region to be understood in the context of the full image by introducing the gated region-context fu-  
 096 sion module (Sec. 3.2.2), while also aligning the entire edited result with a scene-level descrip-  
 097 tion (Sec. 3.2.1) by optimizing the vision-language embedding models, i.e., CLIP Radford et al. (2021)  
 098 and BLIP Li et al. (2022), with contrastive learning. These verbal references are automatically gen-  
 099 erated using large language models, i.e., Deepseek VL Lu et al. (2024), based on structured prompts  
 100 as described in Sec. 3.2.3 and serve as grounding anchors during the editing process. By compar-  
 101 ing the region-in-context with its corresponding region-level description and the full image with the  
 102 global description, our approach ensures that edits are both precise and semantically coherent. This  
 103 design reflects how humans evaluate edits—not in isolation, but through an understanding of the  
 104 broader visual and linguistic context.

061 However, many recent methods for text-  
 062 conditioned image editing focus primarily on  
 063 localizing and modifying target regions Liu  
 064 et al. (2024); Li et al. (2024), often relying  
 065 on isolated region embeddings or mask-based  
 066 attention. While these approaches enable tar-  
 067 geted control, they frequently lack mechanisms  
 068 to reason about the role of the region within the  
 069 full image. As shown in Fig. 2, even with post-  
 070 processing techniques such as blending Li et al.  
 071 (2024), the result may appear unnatural, like  
 072 a sticker pasted onto the image. Other works,  
 073 such as Guo & Lin (2024); Goel et al. (2024),  
 074 still fall short in achieving scene-level seman-  
 075 tic coherence. Moreover, approaches like Lin  
 076 et al. (2024) focus on learning localized atten-  
 077 tion masks, yet remain limited in modeling re-  
 078 lationships between the edited region and the  
 079 overall visual semantics. Vision and language  
 080 are fundamentally intertwined—humans often  
 081 interpret visual scenes through verbal descrip-  
 082 tions and, conversely, imagine visuals based  
 083 on textual cues Radford et al. (2021); Li et al.  
 084 (2022; 2023); Alayrac et al. (2022); Lu et al.

## 2 RELATED WORKS

Text-conditioned image editing with diffusion models has evolved through various strategies to balance edit quality, controllability, and user intent. Sampling-time or inversion-based methods, such as H-Spiegelglas et al. (2024); Lee et al. (2025); Guo & Lin (2024); Brack et al. (2024); Chen et al. (2024), avoid costly retraining by manipulating embeddings, noise schedules, or guidance strength during inference. However, their effectiveness heavily relies on the pretrained noise prediction model, which can limit editing flexibility and fidelity in complex scenarios. Fine-tuning-based methods such as Brooks et al. (2023); Zhang et al. (2024); Kawar et al. (2023) enable instruction-following edits by training on paired image–instruction data, supporting local editing, user interactions, and multi-instruction control. These approaches are particularly effective, as they directly adapt the model’s behavior during training, allowing for purposeful improvements through new data or customized learning strategies. To support evaluation and training grounded in real user intent, HumanEdit Bai et al. (2025) introduces a high-quality, human-rewarded dataset of image edits with natural instructions and diverse semantic transformations. Therefore, in this paper, we adopt this training approach to make the model reach our hypothesis.

A more specific purpose, such as region- and object-level editing, is addressed by models like Li et al. (2024); Liu et al. (2024); Goel et al. (2024); Lin et al. (2024), which enhance spatial control through semantic decomposition or multimodal supervision. Meanwhile, methods such as Nam et al. (2024); Hertz et al. (2022); Liu et al. (2024) explore cross-attention modulation and denoising mechanisms to improve edit faithfulness and structural consistency. More specifically, ZONE Li et al. (2024) proposes a zero-shot, segmentation-free approach that localizes instruction-relevant regions by leveraging CLIP-guided attention maps from a pretrained InstructPix2Pix Brooks et al. (2023) model. It further refines the region using Region-IoU and ensures smooth transitions via FFT-based edge blending, all without extra training or prompt tuning. While Lin et al. (2024); Goel et al. (2024) attempt to enhance regional focus through attention maps or using multimodal to extract more information to add to the model as an additional condition, they do not explicitly supervise whether the model has effectively learned to edit within the intended area. A more refined strategy is proposed in Liu et al. (2024), where regional conditions are injected into the noise prediction model and guided through a self-supervised training pipeline. However, this supervision is applied in isolation, using only binary masks to evaluate whether the edit occurred in the correct region—without considering how that region interacts with the overall image context.

To address this limitation, we propose a training framework that encourages the model to reason about each editable region in the context of the entire image, aligning it not only with spatial constraints but also with the intended semantic meaning of both the region and the global scene.

## 3 PROPOSALS

### 3.1 PRELIMINARIES

Diffusion models generate data by simulating a Markov chain of latent variables  $\mathbf{x}_1, \dots, \mathbf{x}_T$  starting from clean input and progressively adding Gaussian noise at each time step  $t$ , as in Eq. 1, where  $\beta_t$  is a small positive variance schedule. The full forward process is defined as Eq. 2 with  $q(\mathbf{x}_0)$  being the real data distribution. As shown in Eq. 3, the reverse process learns to denoise  $\mathbf{x}_t$  step-by-step back to  $\mathbf{x}_0$  using a neural network  $\epsilon_\theta$  to predict the noise, where  $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$  and  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ .

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

$$q(\mathbf{x}_{1:T} | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}), \quad (2)$$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1 - \beta_t}} \left( \mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}, \quad (3)$$

**Text-Conditioned Image Editing.** To enable text-conditioned generation or editing, the noise prediction network  $\epsilon_\theta$  is conditioned on a text prompt embedding  $\mathbf{e}_{\text{prompt}}$  as Eq. 4. In image editing, the input is typically an image  $\mathbf{x}$  and an instruction  $\mathbf{e}_{\text{text}}$  that specifies the desired modification. One

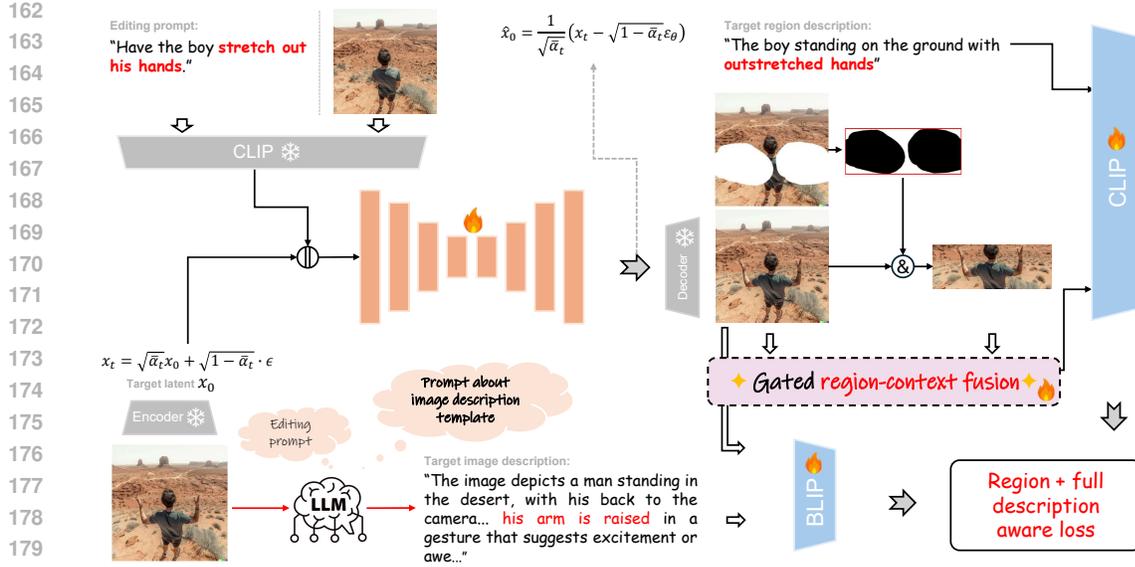


Figure 3: The proposed region-aware diffusion framework for text-guided image editing. The model aligns visual and textual representations at both region and scene levels using a gated region-context fusion module and contrastive supervision. A denoising process reconstructs the edited image from noisy input while ensuring semantic fidelity through region-level and global alignment losses.

common approach is to encode  $x$  into a latent representation  $z_T$  through a forward diffusion process and then perform denoising conditioned on  $y$  to obtain the edited image  $\hat{x}_0$ .

$$\epsilon_\theta(\mathbf{x}_t, t, \mathbf{e}_{\text{prompt}}) \approx \epsilon. \quad (4)$$

### 3.2 OUR FRAMEWORK

We propose a region-aware diffusion model training framework shown as Fig. 3 for text-guided image editing that jointly aligns image and text representations at both global and regional levels. Given a noisy latent input  $x_t$ , our model reconstructs the clean latent  $\hat{x}_0$  using the denoising equation as Eq. 5, where  $\epsilon_\theta$  is the noise prediction network conditioned on the text prompt  $c$ , and  $\bar{\alpha}_t$  is the cumulative product of the noise schedule up to timestep  $t$ .

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} (x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_\theta(x_t, t, c)), \quad (5)$$

To capture fine-grained semantics, both the full image description and the region-specific instruction are embedded and fused using a gated region-context mechanism. Let the predicted clean latent be  $\hat{x}_0 \in \mathbb{R}^{H \times W \times C}$ , and a bounding box defined as  $\mathbf{b} = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ , which is computed as the tightest box enclosing the region mask  $M \in \{0, 1\}^{H \times W}$ . The cropped region corresponding to  $\mathbf{b}$  is given by Eq. 6. The gated-fused region is computed in Eq. 7, where  $F$  denotes the gated region-context fusion module. To supervise region-level alignment, we compute the distance between the gated-fused region embedding  $f_r$  and the embedding of the target region description  $\mathbf{e}_{t_r}$ , encoded via an **optimized** shared CLIP embedding space encoder in Sec. 3.2.1. The region loss is defined as the distance between these two embeddings in Eq. 8. The global loss is the distance on the **optimized** BLIP embedding space between the full image  $\hat{x}_0$  and its corresponding target description as Eq. 9, where  $\mathbf{e}_f$  and  $\mathbf{e}_{t_f}$  is the embedding of full image  $\hat{x}_0$  and the target full image description  $t_f$ , respectively.

$$\hat{x}_0^{\mathbf{b}} = \hat{x}_0[y_{\min} : y_{\max}, x_{\min} : x_{\max}] \quad (6)$$

$$f_r = F(\hat{x}_0, \hat{x}_0^{\mathbf{b}}) \quad (7)$$

$$\mathcal{L}_{\text{region}} = 1 - \cos(f_r, \mathbf{e}_{t_r}) = 1 - \frac{f_r^\top \mathbf{e}_{t_r}}{\|f_r\|_2 \cdot \|\mathbf{e}_{t_r}\|_2} \quad (8)$$

$$\mathcal{L}_{\text{global}} = 1 - \cos(\mathbf{e}_f, \mathbf{e}_{t_f}) = 1 - \frac{\mathbf{e}_f^\top \mathbf{e}_{t_f}}{\|\mathbf{e}_f\|_2 \cdot \|\mathbf{e}_{t_f}\|_2} \quad (9)$$

The model is trained using a combination of region-aware loss, full description alignment loss, and denoising loss illustrated in Eq. 10. These objectives jointly minimize the distance between matching image-text pairs while pushing apart mismatched ones in the embedding space, and ensure consistency between the added noise and the predicted noise during the diffusion process.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{region}} + \mathcal{L}_{\text{global}} + \mathbb{E}_{x_0, t, \epsilon} \left[ \|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2 \right] \quad (10)$$

### 3.2.1 VISION-LANGUAGE EMBEDDING OPTIMIZATION

To promote semantic consistency between the generated visual content and its intended meaning, we optimize a vision-language embedding space where both image regions and scene-level representations are aligned with their corresponding target textual descriptions. This optimization is carried out independently for local (region-level) and global (scene-level) alignment using pretrained multimodal encoders.

For region-level alignment, we embed context-aware image patches  $f_r$  and their respective target region description  $t_r$  into a shared space using a modified CLIP-based model with gated cross-attention fusion. Similarly, for scene-level alignment, we use a pretrained BLIP model to embed the entire image  $\hat{x}_0$  and a detailed scene description  $t_f$ . While CLIP is effective for learning joint embeddings of image-text pairs, it has a limitation in the number of input tokens it can process (77 tokens), which constrains the level of detail that can be captured in textual descriptions. This makes CLIP more suitable for short, focused region descriptions. In contrast, BLIP is designed to handle longer and more descriptive text inputs (512 tokens), making it better suited for scene-level alignment, where a more comprehensive understanding of the global image context and detailed descriptions is required. In both cases, alignment is enforced using a symmetric contrastive loss defined as Eq. 11, encouraging the embeddings of matching image-text pairs to be closer than those of mismatched ones.

$$\mathcal{L}_{\text{optim}} = \frac{1}{2N} \sum_{i=1}^N \left[ -\log \frac{\exp(\mathbf{z}_i^\top \mathbf{t}_i / \tau)}{\sum_{j=1}^N \exp(\mathbf{z}_i^\top \mathbf{t}_j / \tau)} - \log \frac{\exp(\mathbf{t}_i^\top \mathbf{z}_i / \tau)}{\sum_{j=1}^N \exp(\mathbf{t}_i^\top \mathbf{z}_j / \tau)} \right] \quad (11)$$

Generally,  $\mathbf{z}_i$  denotes the visual embedding for the  $i$ -th sample. The contrastive loss over a batch of  $N$  image-text pairs, where  $\tau$  is a temperature parameter that scales the similarity scores. This formulation is applied independently to both region-level embeddings (using CLIP + fusion) and scene-level embeddings (using BLIP), ensuring that the model aligns visual and textual semantics at multiple levels of granularity.

### 3.2.2 GATED REGION IN CONTEXT

With the optimized embedding model with the vision-language embedding  $\mathbf{e}$ , to incorporate global context into localized region representations while maintaining semantic focus, we propose a *Gated Cross-Attention Fusion* module as shown in Fig. 4. To allow the feature of region  $\mathbf{e}_r$  to get understand its role within the broader scene  $\mathbf{e}_f$ , we model its feature  $\tilde{\mathbf{e}}_f$  (through Eq. 12) interaction with the full image using

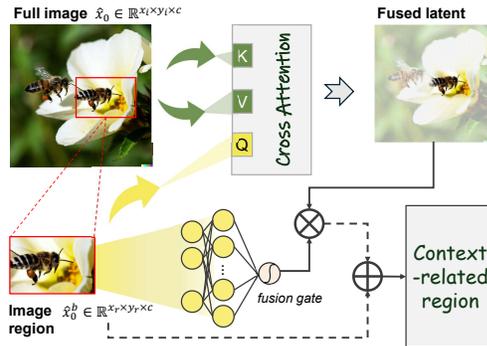


Figure 4: Gated Cross-Attention Fusion module. Region and full image embeddings are first transformed, then fused via multi-head cross-attention. A learnable gate regulates the influence of global context on the region, ensuring controlled and context-aware representation refinement.

multi-head cross-attention. In this setup in Eq. 13, the region embedding acts as the query, meaning it asks: *“Which part of the full image is relevant to me?”* The full image embedding provides both the keys and values, representing the available contextual information. Through this mechanism, the region learns to attend to the most semantically relevant aspects of the full image, enabling it to refine its representation based on the scene in which it appears.

$$\tilde{\mathbf{e}}_r = \text{CNN}_r(\mathbf{e}_r), \quad \tilde{\mathbf{e}}_f = \text{CNN}_f(\mathbf{e}_f) \quad (12)$$

$$\mathbf{h} = \text{MultiHeadAttn}(\mathbf{Q} = \tilde{\mathbf{e}}_r, \mathbf{K} = \tilde{\mathbf{e}}_f, \mathbf{V} = \tilde{\mathbf{e}}_f) \quad (13)$$

To avoid over-reliance on contextual signals, we introduce a gating mechanism as Eq. 14 that controls the influence of the attended global information. This formulation through Eq. 15 and 16 ensures that the region embedding is selectively enhanced by global image context, enabling coherent and context-aware modifications without compromising local semantic fidelity.

$$\mathbf{g} = \sigma(\mathbf{W}_g \tilde{\mathbf{e}}_r + \mathbf{b}_g) \quad (14)$$

$$\mathbf{z} = \tilde{\mathbf{e}}_r + \mathbf{g} \odot \mathbf{h} \quad (15)$$

$$f_r = \mathbf{W}_p \cdot \text{LayerNorm}(\mathbf{z}) + \mathbf{b}_p \quad (16)$$

### 3.2.3 GLOBAL LLM REFERENCE

To provide a rich semantic grounding for scene-level image alignment, we generate a detailed textual description of each image using a large vision-language model (VLM). Specifically, we employ DeepSeek-VL Lu et al. (2024), a multimodal causal language model, to produce structured paragraph-style captions that capture fine-grained visual content from the full image. Given an image and its editing instruction, we construct a carefully designed prompt as Table 4 that guides the model to describe observable attributes—such as objects, textures, colors, spatial relationships, and region-specific edits—without inferring unobservable factors. This process yields dense scene-level descriptions that serve as verbal references for aligning generated images during training. These LLM-generated descriptions are then embedded using a pretrained BLIP encoder and used in the global contrastive loss (Eq. 9) to ensure that the semantic content of the generated image aligns with the intended overall meaning. This global reference complements the localized region-text alignment, allowing the model to reason about edits both precisely and contextually.

### IMPLEMENTATION DETAILS

In the image description task, the model LLM is queried via a chat-style API using multimodal input, where the image and prompt are encoded with DeepSeek-VL’s processor and passed through a pretrained DeepSeek-VL2-Tiny checkpoint Wu et al. (2024b).

We train our Region-CLIP model using the HumanEdit dataset Bai et al. (2025), which provides paired images, target region-level instructions, and binary segmentation masks. All images and masks are resized to a resolution of  $512 \times 512$ . We use a pretrained CLIP ViT-B/16 model Radford et al. (2021). Training is conducted in two phases. In *Phase 1*, the CLIP backbone is frozen and only the fusion module is optimized for 50 epochs using the AdamW optimizer with a learning rate of  $1 \times 10^{-4}$  and a batch size of 64. In *Phase 2*, both the CLIP encoder and the fusion module are fine-tuned jointly for 20 additional epochs with a learning rate of  $1 \times 10^{-5}$ . It takes 7.78 and 3.47 hours to train the fusion module and the whole process, respectively. With the global embedding model, BLIP (blip-itm-base-coco) Li et al. (2022) is fine-tuned with a batch size of 32 and train for 20 epochs using the AdamW optimizer with a learning rate of  $1 \times 10^{-5}$ . The total time for training BLIP is 3.5 hours.

In training the noise prediction model, the pretrained UNet is initialized from a prior checkpoint and optimized using a learning rate of  $1 \times 10^{-4}$  over 9000 steps. We use a resolution of  $256 \times 256$ , a batch size of 8, in approximately 16 training hours. The training loop uses the accelerate library Gugger et al. (2022) with XFormers’ memory-efficient attention. All models were trained on a single NVIDIA A100 GPU.

Table 1: Quantitative comparison of our proposal against other existing baselines. We evaluate performance across six metrics: CLIP-I, DINO, CLIP-T (semantic alignment), LPIPS (perceptual similarity), FID (distributional realism), and IS (image diversity and quality). Our training framework is applied to multiple pretrained diffusion backbones, showing consistent improvements in both semantic and perceptual metrics.

Model	CLIP-I $\uparrow$	DINO $\uparrow$	CLIP-T $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	IS $\uparrow$
InstructPix2Pix Brooks et al. (2023)	0.7531	0.8900	0.8390	0.5008	139.60	9.1981
<b>Ours (on IP2P)</b>	<b>0.9146</b>	<b>0.9736</b>	<b>0.9337</b>	<b>0.2042</b>	<b>77.07</b>	<b>10.4931</b>
MagicBrush Zhang et al. (2024)	0.8939	0.9646	0.9057	0.2448	88.96	<b>9.8105</b>
<b>Ours (on MB)</b>	<b>0.9040</b>	<b>0.9664</b>	<b>0.9207</b>	<b>0.2263</b>	<b>85.61</b>	9.6078
ZONE Li et al. (2024)	0.9569	0.9906	0.9319	0.0650	44.63	10.7947
<b>Ours (on ZONE)</b>	<b>0.9729</b>	<b>0.9955</b>	<b>0.9332</b>	<b>0.0409</b>	<b>27.53</b>	<b>11.3038</b>
CDS Nam et al. (2024)	0.9605	0.9941	0.9239	0.0938	29.50	10.9578
DDPM Inverse H-Spiegelglas et al. (2024)	0.8433	0.9611	0.8584	0.2567	105.45	10.7182

#### 4 EVALUATION

**Baselines and evaluation metrics.** We evaluate our method against several strong text-conditioned image editing baselines, including InstructPix2Pix Brooks et al. (2023), MagicBrush Zhang et al. (2024), ZONE Li et al. (2024), CDS Nam et al. (2024), and DDPM Inversion H-Spiegelglas et al. (2024). For fair comparison, we report results for our framework integrated into each baseline backbone, referred to as *Ours (on IP2P)*, *Ours (on MB)*, and *Ours (on ZONE)*. Evaluation is conducted on the HumanEdit benchmark using a set of both language-image and perceptual quality metrics. We adopt six standard evaluation metrics: CLIP-I and CLIP-T measure CLIP-based similarity between the edited image and the input image or the target text, respectively, as in Radford et al. (2021); DINO score Caron et al. (2021) assesses semantic consistency using DINO-ViT features; LPIPS Zhang et al. (2018) captures perceptual similarity; FID Heusel et al. (2018) evaluates distributional realism against the ground-truth images; and IS Salimans et al. (2016) (Inception Score) quantifies the diversity and quality of generated outputs. As shown in the

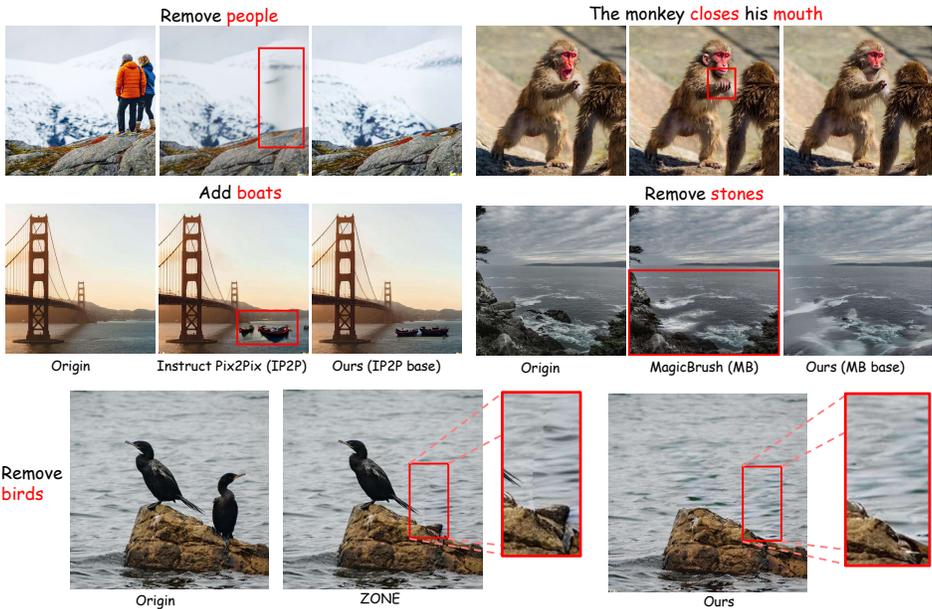


Figure 5: Qualitative comparison with baseline methods. Our framework produces edits that are both semantically accurate and visually coherent across various instructions. Our results consistently exhibit better blending and alignment with both regional and global semantics.

Table 1, our framework consistently enhances the performance of existing diffusion-based editing models across all evaluation metrics. Applied to InstructPix2Pix, we observe a +21.5% increase in CLIP-I (from 0.7531 to 0.9146) and a 59.2% reduction in LPIPS (from 0.5008 to 0.2042), alongside a substantial FID drop from 139.60 to 77.07. Similar improvements are seen with MagicBrush, where CLIP-T improves from 0.9057 to 0.9207, and LPIPS decreases from 0.2448 to 0.2263. When integrated into ZONE, our method achieves state-of-the-art results, with CLIP-I of 0.9729, DINO of 0.9955, and FID reduced by 38.3% (from 44.63 to 27.53). These consistent gains across semantic, perceptual, and generative metrics highlight the generalizability and effectiveness of our framework in improving both local edit fidelity and global scene coherence.

As illustrated in Fig. 5, our framework improves the visual quality and semantic consistency of edited images across various scenarios. For the prompt “remove people”, InstructPix2Pix (IP2P) manages to erase the figures but leaves behind artifacts and unnatural textures in the edited area, whereas our method produces a more seamless and visually coherent background. Similarly, in the “add boats” case, IP2P fails to integrate the boats naturally into the scene. With MagicBrush (MB), when prompted to “The monkey closes his mouth”, the baseline not only alters the mouth but also distorts unrelated regions such as the hand, resulting in an unnatural appearance. Moreover, in the “remove stones” example, MB cannot nearly remove these objects. Although ZONE exhibits strong localization due to its segmentation-based approach, the edited regions often appear visually detached from the rest of the image. This is evident in cases like “remove birds”, where the edit is structurally correct but lacks smooth blending with the background. In comparison, our framework maintains both spatial accuracy and perceptual coherence, yielding edits that are semantically faithful and visually harmonious.

## 5 ABLATION STUDIES

We have finetuned the models IP2P Brooks et al. (2023) and MagicBrush (MB) Zhang et al. (2024) without our proposed framework, using the identical experimental setup described in Sec. 3.2.3, to demonstrate that the observed performance gains stem from our training framework rather than from other experimental factors. The results, shown in Table 2, clearly indicate consistent improvements across all metrics when our framework is applied. Specifically, for IP2P, our method improves CLIP-I from 0.9054 to 0.9146 and reduces LPIPS from 0.2287 to 0.2042, demonstrating better semantic preservation and perceptual fidelity. FID also drops from 83.13 to 77.07, with IS increasing from 9.6015 to 10.4931. Similarly, for MB, CLIP-T improves from 0.9046 to 0.9207, and FID decreases significantly from 104.82 to 85.61. These results confirm that our framework generalizes well across architectures and yields better alignment between the generated image and both regional and global textual guidance. Qualitatively, as illustrated in Fig. 6, we observe that models trained without our

Table 2: Effectiveness of the proposed training framework through comparing models finetuned with and without our framework.

Model	IP2P (without)	IP2P (with)	MB (without)	MB (with)
CLIP-I↑	0.9054	<b>0.9146</b>	0.8793	<b>0.9040</b>
DINO↑	0.9680	<b>0.9736</b>	0.9517	<b>0.9664</b>
CLIP-T↑	0.9249	<b>0.9337</b>	0.9046	<b>0.9207</b>
LPIPS↓	0.2287	<b>0.2042</b>	0.2696	<b>0.2263</b>
FID↓	83.13	<b>77.07</b>	104.82	<b>85.61</b>
IS↑	9.6015	<b>10.4931</b>	8.9593	<b>9.6078</b>

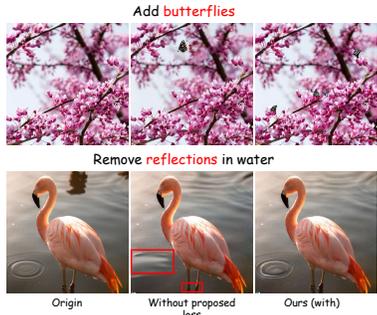


Figure 6: Qualitative ablation study on loss components: the model with vs without proposed loss.

proposed framework often perform imprecise or overly broad edits. For instance, with the prompt “Remove reflection in water”, the model trained without our method not only removes the reflection but also unintentionally alters unrelated regions such as the water surface and the bird’s feet. In

contrast, our approach yields more focused and semantically faithful edits, preserving surrounding details while accurately fulfilling the instruction.

Table 3: Evaluate the impact of removing each component in our training framework. Region-level semantic alignment and gated fusion contribute significantly to the model’s performance, as removing either leads to notable degradation in quality.

Model	IP2P (without full desc loss)	IP2P (without region desc loss)	IP2P (without gated fusion)	IP2P (with proposed loss)
CLIP-I↑	0.8851	0.8806	0.8688	<b>0.9146</b>
DINO↑	0.9565	0.9569	0.9518	<b>0.9736</b>
CLIP-T↑	0.9089	0.9136	0.8996	<b>0.9337</b>
LPIPS↓	0.2583	0.2600	0.2942	<b>0.2042</b>
FID↓	102.63	99.59	102.79	<b>77.07</b>
IS↑	9.5514	9.1468	9.2767	<b>10.4931</b>

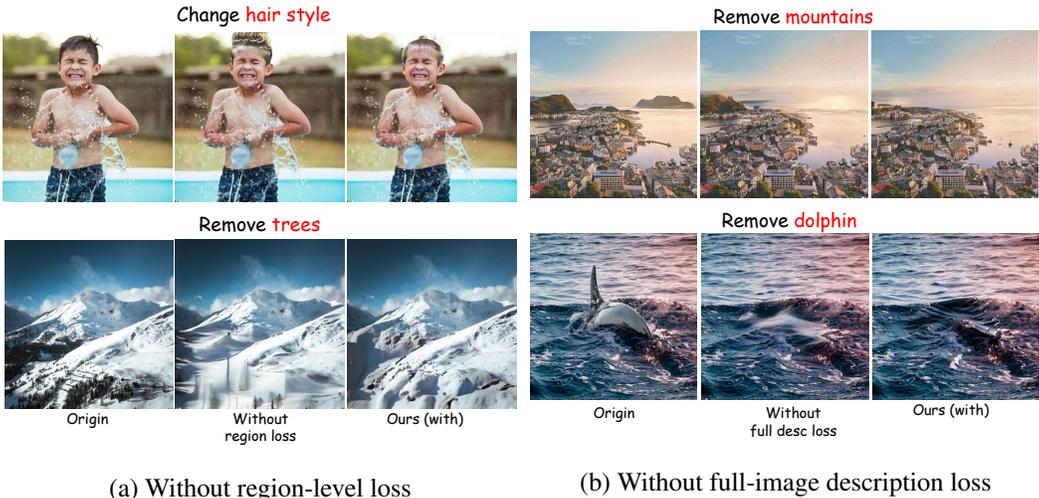


Figure 7: Visual comparison of models trained without region-level (left) and full-image (right) description losses.

To better understand the contribution of each component in our framework, we conduct an ablation study by selectively removing key elements: the full description loss, the region description loss, and the gated fusion module. As shown in Table 3, removing the region description loss results in greater performance degradation (e.g., CLIP-I drops from 0.9146 to 0.8806 and FID increases from 77.07 to 99.59) than removing the full description loss (CLIP-I drops to 0.8851, FID increases to 102.63), indicating that localized semantic alignment plays a more critical role in image editing. Moreover, disabling the gated fusion while keeping the region description loss leads to the worst performance across most metrics (e.g., LPIPS rises to 0.2942 and CLIP-I drops to 0.8688), emphasizing the importance of integrating contextual information during region embedding. These findings support our hypothesis that both region-aware guidance and context fusion are essential for achieving fine-grained, coherent edits.

Fig. 7 illustrates the qualitative improvements achieved by incorporating our proposed loss components. For instance, in the example of “remove dolphin”, although the baseline model (trained without our losses) succeeds in removing the dolphin, it leaves behind an unnaturally flat water region. In contrast, our method produces a more visually coherent result by synthesizing natural water textures—such as waves—within the edited region, blending it seamlessly into the context.

## 6 CONCLUSION

In this work, we propose a novel region-aware training framework for text-conditioned image editing that integrates both local and global semantic alignment. Inspired by how humans contextualize edits within an entire scene, our method enables each editable region to understand its role in the broader visual composition through a gated region-context fusion module and dual-level vision-language supervision. We optimize region-level embeddings using contrastive loss against fine-grained textual descriptions and align the entire image with detailed scene-level descriptions generated by a vision-language model. Extensive experiments demonstrate that our framework significantly improves edit accuracy, semantic coherence, and visual fidelity across multiple pretrained diffusion-based editors. These results confirm the effectiveness and generalizability of our approach in achieving fine-grained, contextually consistent image editing.

## REFERENCES

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing, 2025. URL <https://arxiv.org/abs/2412.04280>.
- Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models, 2024. URL <https://arxiv.org/abs/2311.16711>.
- Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023. doi: 10.1109/CVPR52729.2023.01764.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.
- Sherry X Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Kuo-Chin Lien, Misha Sra, and Pradeep Sen. Tino-edit: Timestep and noise optimization for robust diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6337–6346, 2024.
- Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models, 2024. URL <https://arxiv.org/abs/2309.17102>.
- Vidit Goel, Elia Peruzzo, Yifan Jiang, DeJia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair diffusion: A comprehensive multimodal object-level image editor. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8609–8618, 2024. doi: 10.1109/CVPR52733.2024.00822.
- Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. <https://github.com/huggingface/accelerate>, 2022.
- Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6986–6996, 2024. doi: 10.1109/CVPR52733.2024.00667.
- Inbar H-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*, pp. 12469–12478, 2024.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL <https://arxiv.org/abs/2208.01626>.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL <https://arxiv.org/abs/1706.08500>.

- 540 Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal  
541 Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and*  
542 *Pattern Recognition 2023*, 2023.
- 543 Hyunsoo Lee, Minsoo Kang, and Bohyung Han. Diffusion-based conditional image editing through optimized  
544 inference with guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*,  
545 2025.
- 546 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for  
547 unified vision-language understanding and generation. In *International conference on machine learning*, pp.  
548 12888–12900. PMLR, 2022.
- 549 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training  
550 with frozen image encoders and large language models. In *International conference on machine learning*,  
551 pp. 19730–19742. PMLR, 2023.
- 552 Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu,  
553 Jianzhuang Liu, and Baochang Zhang. Zone: Zero-shot instruction-guided local editing. In *2024 IEEE/CVF*  
554 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6254–6263, 2024. doi: 10.1109/  
555 CVPR52733.2024.00598.
- 556 Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-driven image editing via  
557 learnable regions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.  
558 7059–7068, 2024. doi: 10.1109/CVPR52733.2024.00674.
- 559 Chang Liu, Xiangtai Li, and Henghui Ding. Referring image editing: Object-level image editing via referring  
560 expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*  
561 *(CVPR)*, pp. 13128–13138, June 2024.
- 562 Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhu-  
563 oshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint*  
564 *arXiv:2403.05525*, 2024.
- 565 Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided  
566 latent diffusion image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*  
567 *Recognition (CVPR)*, pp. 9192–9201, June 2024.
- 568 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-  
569 try, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural  
570 language supervision. In *International conference on machine learning*, pp. 8748–8763. PmlR, 2021.
- 571 Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved  
572 techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- 573 Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled  
574 diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.  
575 6327–6336, 2024a. doi: 10.1109/CVPR52733.2024.00605.
- 576 Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma,  
577 Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi  
578 Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao,  
579 Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced  
580 multimodal understanding, 2024b. URL <https://arxiv.org/abs/2412.10302>.
- 581 Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for  
582 instruction-guided image editing, 2024. URL <https://arxiv.org/abs/2306.10012>.
- 583 Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness  
584 of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern*  
585 *Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.
- 586  
587  
588  
589  
590  
591  
592  
593

## APPENDIX

Table 4: Prompt for the LLM to make the image description in a template

594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647	<p><b>Prompt</b></p> <ul style="list-style-type: none"> <li>• <b>Input:</b> &lt;image&gt;</li> <li>• <b>Instruction:</b> You are a meticulous visual analyst. Carefully examine the given image and describe it in a single, flowing paragraph (maximum 520 tokens). Focus on every visually observable detail—such as color, texture, material, size, shape, and spatial relationships. Do not use bullet points or lists.</li> <li>• <b>Constraints:</b> Avoid assumptions or inferences about unseen factors (e.g., time of day, season, emotions, story). Describe only what is directly visible in the image.</li> <li>• <b>Your paragraph must naturally include:</b> <ul style="list-style-type: none"> <li>– A clear overview of the setting (e.g., indoor/outdoor, environment type, lighting conditions, background elements, overall mood)</li> <li>– Detailed description of each major object: <ul style="list-style-type: none"> <li>* Appearance, color, material (wood, metal, fabric, etc.)</li> <li>* Texture (smooth, rough, shiny, soft, etc.)</li> <li>* Size (relative to others)</li> <li>* Spatial position (e.g., foreground, center-left)</li> </ul> </li> <li>– If humans or animals are present, describe each one in full detail: <ul style="list-style-type: none"> <li>* Hair, face, visible skin or fur, and accessories</li> <li>* Clothing: color, texture, material, style, condition</li> <li>* Pose: orientation and position of each body part (head, arms, legs, torso, hands, feet)</li> <li>* Stance or motion—only if clearly visible and grounded in the image</li> </ul> </li> <li>– Distinct, thorough description for multiple people or animals</li> <li>– Supporting/background elements: furniture, walls, ground, vegetation, distant objects</li> <li>– Clear spatial relationships (e.g., in front of, behind, next to, overlapping, under)</li> <li>– Explicit description of the visual features of each object or region targeted in the editing instruction: "{edit_prompt}". For example, if the instruction is "The girl bent and raised her two hands", then describe: <ul style="list-style-type: none"> <li>* Her posture (e.g., leaning forward, bent knees)</li> <li>* The position and gesture of her hands (e.g., raised above shoulders, palms open)</li> </ul> </li> </ul> </li> <li>• <b>Style Requirement:</b> Use vivid, sensory-rich language. Every detail must be grounded in what can actually be seen. Avoid summarizing—immerse the reader in a scene constructed entirely from the image’s visible content.</li> </ul>
--	---

648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701

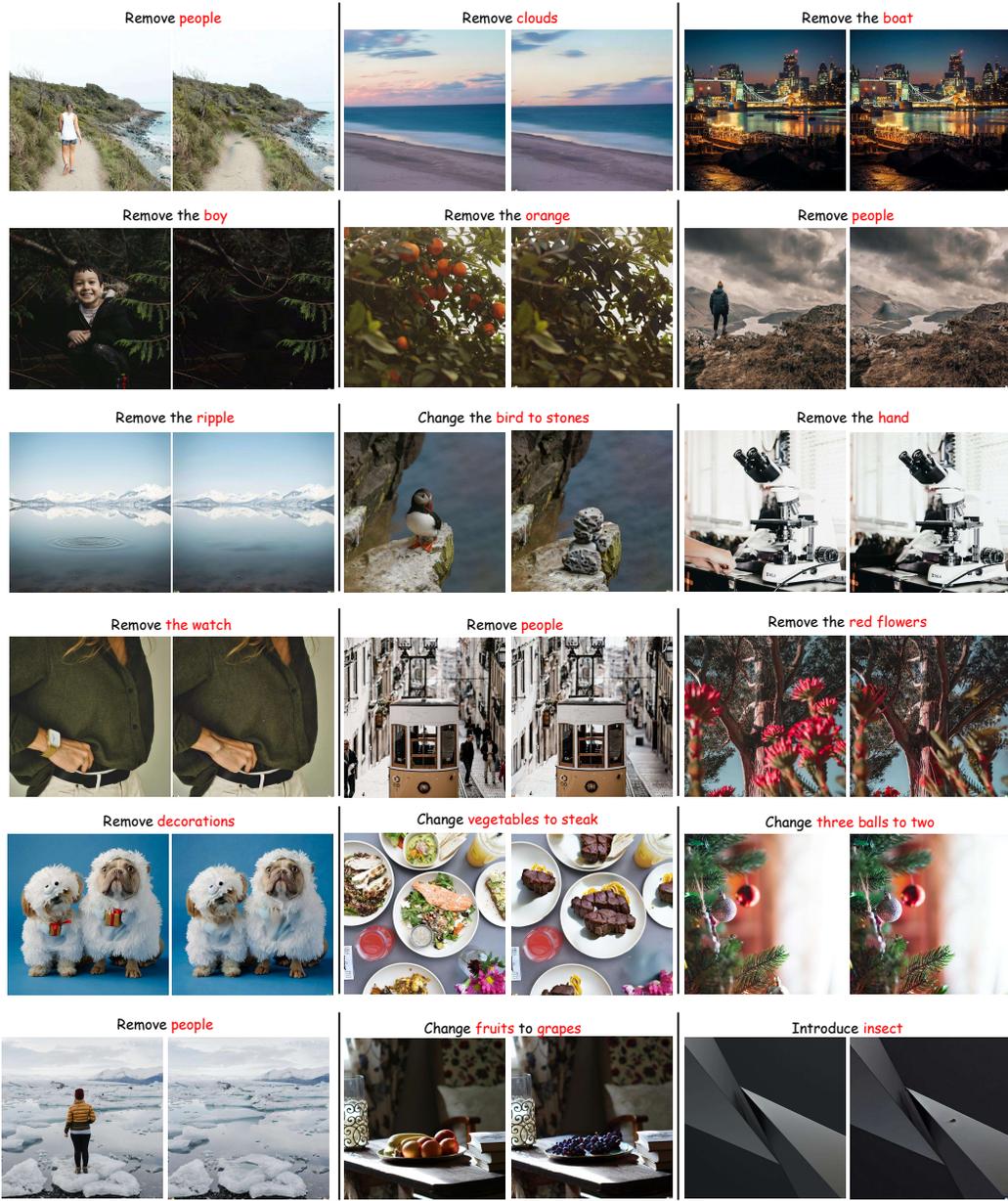


Figure 8: Some qualitative results of our proposal

Table 5: Some examples of the image description generated by Deepseek-VL-Tiny with the prompt in Table 4

702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

Image	Editing Instruction	Image Description
	Remove the woman.	<p>The image captures a serene nighttime landscape, where the starry sky dominates the upper portion of the frame. The celestial display is dense with stars, creating a textured canvas of light and dark. Below, the terrain is a lush green field, dotted with wildflowers and grasses. In the middle ground, a cluster of coniferous trees stands tall, their silhouettes contrasting against the bright moon. The moon itself is a focal point, casting a soft glow that illuminates the scene with a gentle light. To the right, a rocky outcrop rises, adding a rugged element to the otherwise tranquil setting. The overall mood is peaceful and contemplative, inviting the viewer to pause and appreciate the beauty of the night sky.</p>
	Replace wildflowers with red roses.	<p>The image captures a close-up view of a hand holding a single red rose. The rose is vibrant, with deep red petals that appear soft and velvety, suggesting it might be made of silk or satin. The stem is slender and green, indicating a natural material. The hand is positioned in the foreground, with fingers gently curled around the stem, showcasing the delicate texture of the skin. The background is blurred, creating a bokeh effect that highlights the rose and hand, giving the image a dreamy, ethereal quality. The lighting is soft and diffused, casting gentle shadows and enhancing the rich color of the rose. The overall mood is intimate and serene, evoking a sense of romance and tranquility.</p>
	Change a white dress into a black dress.	<p>The image depicts a young woman standing amidst a vibrant backdrop of pink bougainvillea flowers. She is wearing a black dress adorned with white floral patterns, which contrasts beautifully with the bright colors of the flowers. Her hair is styled in a sleek, straight manner, and she has a subtle smile on her face. The lighting is soft and natural, suggesting that the photo was taken outdoors during the daytime. The overall mood of the image is serene and elegant, with the woman appearing relaxed and at ease in her surroundings.</p>
	Add a bird	<p>The image depicts a bronze statue of a man in formal attire, including a suit jacket, vest, and bow tie. The statue is situated outdoors, with bare branches of a tree partially obscuring the view. A bird is perched on the shoulder of the statue, adding a touch of life to the scene. The background features a modern building with large windows, suggesting an urban environment. The lighting appears soft, possibly indicating an overcast day or the time of day being late afternoon. The statue's detailed craftsmanship is evident in the texture of the clothing and the realistic portrayal of the man's features. The bird adds a dynamic element to the otherwise static composition.</p>