# InstructRestore: Region-Customized Image Restoration with Human Instructions

**Shuaizheng Liu**[1,2], **Jianqi Ma**[1], **Lingchen Sun**[1,2], **Xiangtao Kong**[1,2], **Lei Zhang**[1,2,†]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute
shuaizhengliu21@gmail.com, {jianqi.ma, ling-chen.sun, xiangtao.kong}@connect.polyu.hk
cslzhang@comp.polyu.edu.hk

## Abstract

Despite the significant progress in diffusion prior-based image restoration for real-world scenarios, most existing methods apply uniform processing to the entire image, lacking the capability to perform region-customized image restoration according to user preferences. In this work, we propose a new framework, namely **InstructRestore**, to perform region-adjustable image restoration following human instructions. To achieve this, we first develop a data generation engine to produce training triplets, each consisting of a high-quality image, the target region description, and the corresponding region mask. With this engine and careful data screening, we construct a comprehensive dataset comprising 536,945 triplets to support the training and evaluation of this task. We then examine how to integrate the low-quality image features under the ControlNet architecture to adjust the degree of image details enhancement. Consequently, we develop a ControlNet-like model to identify the target region and allocate different integration scales to the target and surrounding regions, enabling region-customized image restoration that aligns with user instructions. Experimental results demonstrate that our proposed InstructRestore approach enables effective human-instructed image restoration, including restoration with controllable bokeh blur effects and region-specific restoration with continuous intensity control. Our work advances the investigation of interactive image restoration and enhancement techniques. Data, code, and models are publicly available at `https://github.com/shuaizhengliu/InstructRestore.git`.

## 1 Introduction

Image restoration (IR) is a fundamental problem in computer vision to recover high-quality images from degraded inputs. Early works have achieved significant progress on individual IR tasks based on specific simulated degradation assumptions, including denoising [59, 60], deblurring [28, 36], and super-resolution [9, 24]. While demonstrating strong performance within their target domains, these approaches exhibit inherent limitations when generalizing to real-world scenarios characterized by unknown and composite degradations. This has motivated the emerging paradigm of real-world image restoration, which aims to handle complex degradation in practical imaging scenarios, particularly for challenging cases like real-world super-resolution [5, 43]. To address this challenge, recent advances have developed sophisticated degradation models to better approximate real-world conditions [57, 39]. Building upon these advanced degradation models, and with the advent of pretrained text-to-image (T2I) generation models such as Stable Diffusion (SD) [33], which can more effectively model the complex distribution of natural images, researchers have started to explore the use of powerful SD priors to produce realistic IR outcomes [38, 25, 50, 46, 52, 45, 35, 41, 2, 8, 51].

† Corresponding author. This work is supported by the PolyU-OPPO Joint Innovative Research Center.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

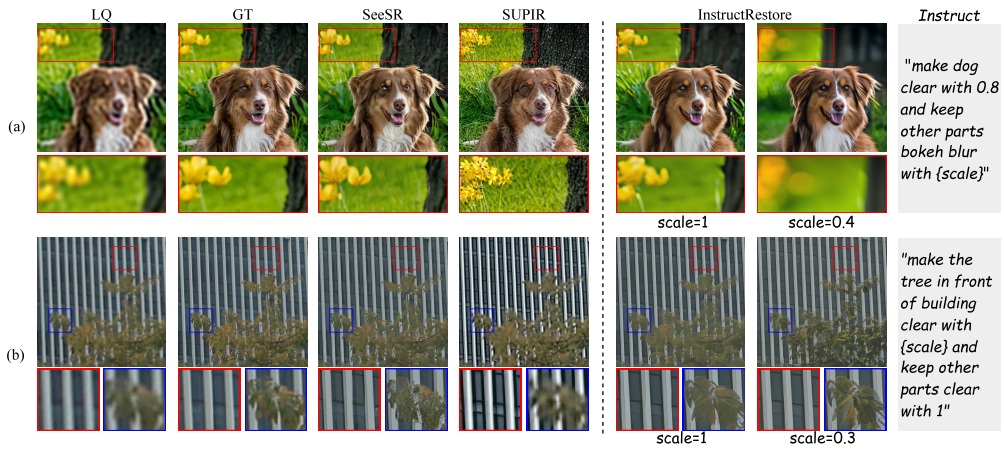


Figure 1: Our proposed **InstructionRestore** framework enables region-customized restoration following human instruction. As shown in (a), current methods [46, 52] tend to restore the bokeh blurry region incorrectly, while our approach allows for adjustable control over the degree of blur based on user instructions. In (b), existing methods fail to achieve region-specific enhancement intensities, while our approach can simultaneously suppress the over-enhancement in areas of building and improve the visual quality in areas of leaves.

By generating details semantically consistent with the underlying content of the image, SD-based methods achieve significantly better perceptual quality than previous approaches [23, 26, 6, 9, 34, 29, 40, 39, 22]. However, image restoration is an ill-posed problem that admits multiple plausible solutions. Existing methods are limited to a single restoration outcome applied uniformly across the image, lacking the ability to accommodate varied user preferences across different image regions. For example, in the photography of targeted objects (*e.g.*, portrait), the background is intentionally blurred for aesthetic focus. During restoration, users typically want to preserve or even adjust bokeh effects, yet existing generative prior-based IR methods may produce unnecessary textures that disrupt the intended bokeh effects on the background regions, as shown in Fig. 1(a). In addition, user preferences for content fidelity and perceptual quality vary across image semantic regions. For irregular texture regions (*e.g.*, trees), it's challenging to accurately recover pixel-wise details due to signal aliasing in the degradation process [22]. In these cases, strictly enforcing fidelity often leads to over-smoothed results. Therefore, users generally prioritize perceptual quality for irregular texture regions, favoring more aggressive detail generation. Conversely, for structural regions (*e.g.*, architecture) or flat areas (*e.g.*, skies), large pixel-wise differences are more perceptually sensitive and easily detected as artifacts [47]. Therefore, user preference may shift towards content fidelity to preserve accuracy, as illustrated in Fig. 1(b). Unfortunately, existing methods cannot achieve such customized restoration of different regions.

To address the limitation mentioned above, we propose **InstructRestore**, a novel framework that enables users to realize region-specific restoration through natural language instruction for real-world scenarios, including bokeh adjustment and region-aware tuning of content fidelity and perceptual quality. Our InstructRestore approach can precisely adjust restoration effects in target semantic regions while keeping other areas unaffected, showing the ability of instruction following. To begin with this novel task, we need a dataset for training and evaluation, which should offer descriptions of target regions to construct human instruction, along with corresponding region masks. To the best of our knowledge, there is not a publicly available dataset that provides such triplets of high-quality images, referring descriptions, and the corresponding region masks. The most relevant datasets to our task can be the referential segmentation datasets such as RefCOCO [53]. However, its image quality and resolution are insufficient to support IR tasks. To bridge this gap, we develop a data generation engine. Utilizing Semantic-Sam [18] and Osprey [55] models, we obtain masks and initial descriptions from a set of selected high-quality images. We then use large language models (LLMs), more specifically Qwen [48], to iteratively parse and refine these descriptions, formatting them to meet the instructional requirements of IR tasks. Finally, we build a dataset of 536,945 triplets, covering diverse scenes such as plants, buildings, animals, *etc*.

Building upon this dataset, we train the InstructRestore model for region-customized IR with user instructions. To ensure that the model can accurately identify the human-specified region and properly

enhance the designated area, we propose integrating the conditional features of low-quality input images into a ControlNet-like architecture. Instructions are used as text prompts to the control-branch of ControlNet [61]. Trained on our curated dataset, the control-branch could simultaneously generate region masks and conditional features. By applying distinct integration scales to the conditional features of user-customized regions and their surroundings, our InstructRestore model achieves locally controlled restoration that aligns with user intentions.

Our key contributions are summarized as follows. (1) First, we introduce the task of region-customized image restoration with human instruction, which represents an important class of practical IR tasks. (2) Second, we develop a data generation engine and construct a large-scale dataset with $536,945$ triplets to support this task. (3) Finally, we present InstructRestore, the first model that understands user instructions for region-customized restoration for real-world complex degradation. Our experiments demonstrate the capability and effectiveness of our InstructRestore model, showcasing its great potential for interactive and user-instructed image restoration.

## 2 Related Work

**Diffusion-based Restoration in Real World**. Recent diffusion models have significantly advanced the task of IR in real world, addressing mixed degradations such as noise, blur, JPEG compression, and resolution reduction. StableSR [38] and DiffBIR [25] treat the low-quality (LQ) input as condition to guide reverse diffusion process. PASD [50] and SeeSR [46] introduce the semantic prompts like short captions or tags to enrich the result with finer semantic details. SUPIR [52] scales up datasets along with long descriptions to boost perceptual quality with SDXL [30] pre-trained model. DreamClear [2] and FluxIR [8] introduce Diffusion Transformer (DiT)-based models designed for enhanced performance in image restoration. To tackle the inefficiency of iterative sampling, one-step diffusion methods [45, 35, 56] have emerged, ensuring quality with faster inference. Despite their advancements, existing methods perform restoration uniformly, failing to accommodate user preferences for region-specific refinements.

**Instruction-guided Editing and Restoration**. Natural language instructions enable intuitive human-AI collaboration by translating high-level intent into pixel-level operations. Instruction-guided image editing methods like InstructPix2Pix [4] and MagicBrush [58] have demonstrated remarkable capabilities in spatially aware manipulations. Subsequent works like MGIE [10] and SmartEdit [13] further advance instruction comprehension through multimodal LLMs. Others [19, 11] focus on region-specific control, ensuring editing explicitly defined areas by user instructions. However, these breakthroughs remain confined to semantic-level manipulation rather than physically grounded restoration. To address this, recent efforts have incorporated user instructions into restoration frameworks. InstructIR [7] and PromptFix [54] leverage task-specific instructions to enable a single model to handle multiple restoration tasks, including denoising, deblur, rain removal, *etc*. SPIRE [31] incorporates semantic descriptions to handle in-the-wild restoration scenarios. However, these methods primarily use instructions for task differentiation or global parameter tuning, lacking the ability to perform region-specific refinements. Our work introduces the first instruction-guided restoration framework for real-world scenarios that enables region-specific refinements through natural language commands, addressing the critical limitation of global-only operations in prior arts.

## 3 Dataset Construction

InstructRestore aims to adjust restoration effects on user-specified regions following human instructions. To achieve this goal, the model needs to understand the semantic information of the target regions for performing localized restoration. A critical requirement for training such a model is the availability of a large-scale dataset, which simultaneously offers high-quality images, descriptions of target regions, and corresponding region masks. In this paper, we develop a data generation engine to build such a comprehensive dataset, named **Tri-IR**, to facilitate the research of InstructRestore tasks. The data generation process is detailed in Fig. 2.

### 3.1 Dataset Construction Pipeline

**High-quality ground-truth image collection.** High-resolution and high-quality ground-truth (GT) images are critical for training IR models. Therefore, we collect high-quality images from LSDIR [20],

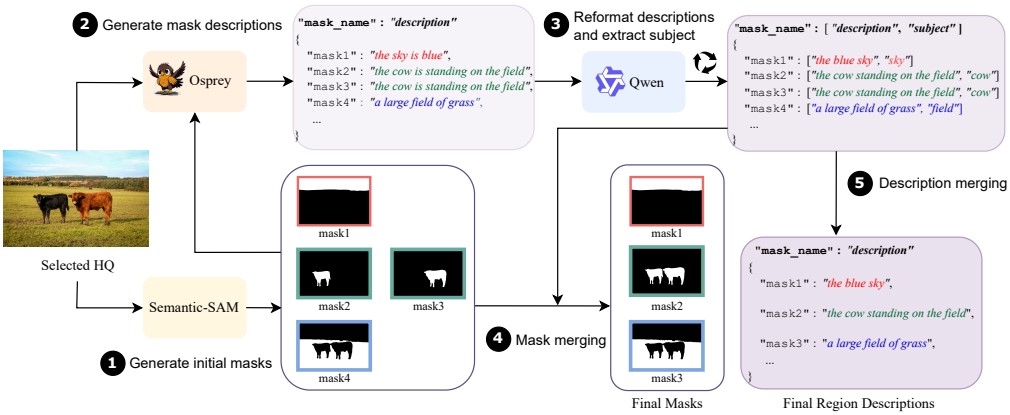


Figure 2: Illustration of the annotation pipeline. For selected high-quality images, Semantic-SAM [18] generates initial masks, followed by Osprey [55] for region-level descriptions. Qwen [48] reformats descriptions into noun phrases and extracts semantic subjects. Identical semantics are merged to produce final masks and region captions.

EntitySeg [32] train set and EBB! [14] bokeh train set with shorter side larger than 512 pixels and MUSIQ [17] score larger than 60.

**Annotation pipeline.** To obtain high-quality images, we design an automatic annotation pipeline to extract the semantic region masks and their corresponding descriptions with a combination of state-of-the-art models. In the mask extraction phase, we first utilize a state-of-the-art segmentation model, *e.g.*, Semantic-SAM [18], to generate coarse-grained semantic segmentation masks for the semantic region of the images. For images from EntitySeg [32], we directly reuse their pre-annotated masks. Once the masks are obtained, we pair each image with its mask and feed them into a multi-modal large language model, Osprey [55], to generate region-level descriptions. These descriptions serve as part of instructions to specify the regions to be processed or restored. At this stage, though we obtain preliminary masks and descriptions, they are still far from perfect as our training data due to two key issues below: (1) Semantic-SAM [18] occasionally produces multiple mask pieces for one semantic meaning, leading region ambiguity and harmful for the region customization learning; (2) the descriptions are not always in noun phrase format due to the response arbitrariness, making them unsuitable for embedding into instructions.

To address these issues, we first utilize Qwen-7B [3], a large language model (LLM), to perform the following tasks through prompt tuning: (1) parsing the subject from the descriptions and (2) reformatting them into noun phrases. Due to the randomness in LLM's outputs, we iteratively perform the refinement process. Specifically, we identify error cases and re-execute the above process by a larger LLM, Qwen-72B [3]. This cycle is repeated 3 times to ensure high-quality outputs. More details can be found in the **appendices**. Finally, based on the parsed subject, we merge the masks and their corresponding descriptions for regions with identical semantics.

### 3.2 Dataset Statistics

As shown in Fig. 2, our Tri-IR dataset, provides triplets of high-quality GT images, region masks, and descriptive captions. To underscore the relevance and utility of our dataset, we compare it with the most relevant referential segmentation datasets including RefClef [16], RefCOCO [53], RefCOCO+ [53] and RefCOCOg [27] in Table 1, which also provide masks and captions for semantic regions. The comparison focuses on the number of annotations, the range of image resolutions, and MUSIQ-based quality scores.

Table 1: Statistics of our dataset and related datasets.

| Datasets | Annotation Amount | Min Resolution | Max Resolution | MUSIQ |
|---|---|---|---|---|
| RefClef [16] | 99,523 | 320×480 | 360×480 | 67.06 |
| RefCOCO [53] | 196,771 | 157×160 | 640×637 | 69.73 |
| RefCOCO+ [53] | 196,737 | 157×160 | 640×637 | 69.73 |
| RefCOCOg [27] | 208,960 | 157×160 | 640×637 | 69.73 |
| Ours | 536,945 | 540×540 | 4464×2244 | 71.87 |

As can be seen from Table 1, existing datasets, while widely used for segmentation, exhibit critical limitations for IR tasks. Their images are capped at resolution less than 650 pixels and their MUSIQ scores fall significantly below ours. In contrast, our dataset not only provides 536,945 annotated regions (surpassing other datasets in scale) but also delivers higher-resolution images with superior

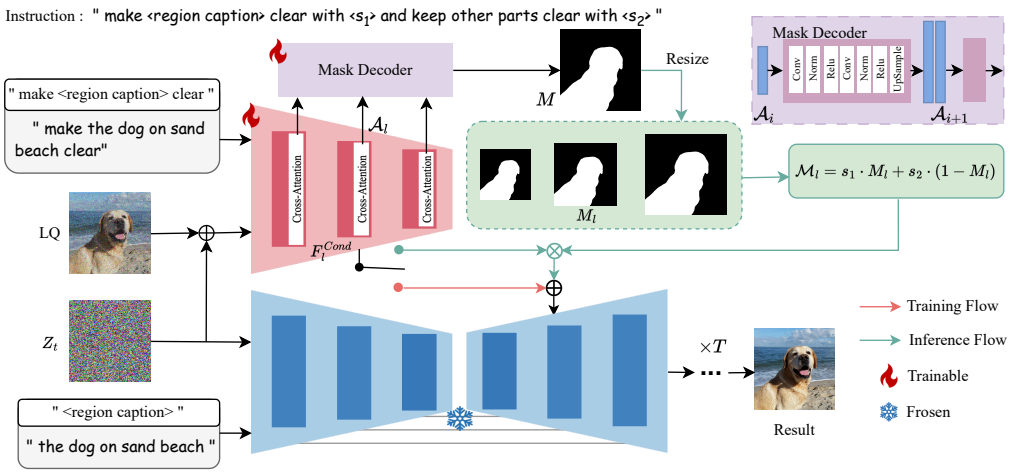


Figure 3: Framework of InsturctRestore. The framework uses red and green arrows to denote training and inference processes respectively. During testing, user instructions are parsed to generate target-region semantic masks, with differentiated coefficient modulation applied to conditional features inside/outside mask regions, enabling instruction-guided region-specific restoration effects.

perceptual quality, meeting the need for IR tasks. Our dataset enables both precise semantic control and photorealistic restoration. To further illustrate the semantic diversity and applicability of our dataset, we plot a word cloud reflecting the relative frequency of different semantic content in the **appendices**, which demonstrates that our dataset covers a wide range of semantic regions that are commonly targeted in restoration tasks, such as plants, buildings, animals, *etc*.

## 4 InstructRestore Model Design

Our network is designed to achieve region-specific restoration effects based on user instructions, where each instruction contains both spatial region specifications and restoration strength. The core challenge lies in how to accurately localize specified regions and implement continuous controllable local restoration effects with given strength, while keeping the restoration of remaining regions unchanged. A straightforward approach might be to construct training data pairs corresponding to different local restoration strengths. However, this approach is too labor-intensive. We develop a more elegant solution to address the above challenges without requiring different strength pairs. Specifically, existing SD-based IR models often employ a ControlNet architecture with the low-quality (LQ) image as a conditional signal. In this architecture, the pre-trained SD backbone generates text-guided features while the ControlNet branch extracts LR-derived features. The fusion between these two pathways determines the final restoration output. We observe that scaling the ControlNet features by a coefficient $\alpha$ during inference provides flexible control of the data fidelity and semantic enhancement. Intuitively, a smaller coefficient allows the SD backbone to dominate, resulting in richer generated details and enhanced perceptual quality, while a larger coefficient makes the output closer to the degraded image, increasing fidelity but reducing generated details. Building on this insight, we develop **InstructRestore**, which employs a ControlNet-like architecture during training using only standard restoration data. The network learns to perform restoration while predicting region masks from user instructions. During inference, InstructRestore generates region masks and scales ControlNet feature by different coefficients inside and outside the masks based on user instructions to achieve region-specific restoration.

### 4.1 Training Framework

**Architecture design**. As shown in Fig. 3, our InstructRestore model consists of a pre-trained SD backbone, the ControlNet adaptor, and a lightweight mask decoder. The SD model is frozen during the entire training stage. The region captions $c_R$ extracted from the user instructions $c_I$ act as text prompts for the SD model, providing semantic guidance to generate semantic details. ControlNet duplicates the encoder and middle blocks of the SD UNet as trainable copies. It receives features

extracted from the LQ image and user instructions $c_I$ as input, then extracts hierarchical conditional features from the input and injects them into the UNet decoder blocks at multiple scales.

To accurately localize the target regions in user instructions, we design a mask decoder to predict a spatial mask $\hat{M}$. Since ControlNet is initialized from the pre-trained SD UNet, it has been revealed [12] that the cross-attention features between textual and visual embeddings exhibit strong responses to text-described semantic regions. We then extract cross-attention features $\{\mathcal{A}_l\}_{l=1}^{L}$ between textual embeddings and visual features at each scale of the ControlNet as input to the Mask decoder, which is designed with a pyramidal structure to effectively process multi-scale features. The features of each scale $\mathcal{A}_l$ are first passed through two blocks, each consisting of a convolutional layer (Conv), group normalization (GN), and a ReLU activation. The processed features are then upsampled and concatenated with the cross-attention features $\mathcal{A}_{l+1}$ at the next scale. The combined features are processed by another Conv-GN-ReLU block and passed to the subsequent scale.

**Training process**. Our constructed dataset consist of triplets $[I_{\text{HQ}}, M, c_M]$, where $I_{\text{HQ}}$ denotes the high-quality GT image, $M$ is the binary mask specifying the target region, and $c_M$ is the textual caption of the masked area. To generate training samples, we first apply the Real-ESRGAN degradation pipeline to $I_{\text{HQ}}$ to obtain the LQ input $I_{\text{LQ}}$. Subsequently, we construct specific instructions $c_I$ and region caption $c_R$ based on $c_M$, tailored to different restoration purposes. For region-specific restoration, $c_I$ follows the template *"make* $\{c_M\}$ *clear"*, while $c_R$ is the same as $c_M$. For bokeh-aware restoration, the template becomes *"make* $\{c_M\}$ *clear and keep other parts bokeh blur"*, while $c_R$ follows the template *"*$\{c_M\}$ *in front of bokeh background"*.

During training, $I_{\text{HQ}}$ is first encoded into the latent space by the pre-trained VAE encoder, yielding $z_0$. The diffusion process progressively corrupts $z_0$ with Gaussian noise over randomly sampled timesteps $t$, resulting in noisy latent states $z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\alpha_t$ follow a cosine noise schedule. We utilize $z_t$ and region caption $c_R$ as inputs to pre-trained SD backbone. The ControlNet takes $I_{\text{LQ}}$, $z_t$ and $c_I$ as input to produce conditional features $\{F_l^{\text{cond}}\}_{l=1}^{L}$, which are added to the frozen SD UNet decoder with scaling factor $\alpha = 1$ during training. The training flow is highlighted in red in Fig. 3.

The mask decoder takes the cross-attention features $\{\mathcal{A}_l\}_{l=1}^{L}$ from ControlNet as input to generate target region masks $\hat{M}$, supervised by the GT masks $M$ with Cross-Entropy loss. The InstructRestore network, denoted by $\epsilon_\theta$, is conditioned on noisy latent $z_t$, LQ images $I_{\text{LQ}}$, instruction $c_I$, and region caption $c_R$. The training objective $\mathcal{L}$ combines the diffusion loss and mask supervision $\mathcal{L}_{\text{mask}}$:

$$\mathcal{L} = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, I_{\text{LQ}}, c_I, c_R)\|_2^2\right] + \lambda \mathcal{L}_{\text{mask}}(\hat{M}, M), \tag{1}$$

where $\mathcal{L}_{\text{mask}}(\hat{M}, M) = \text{CrossEntropy}(\hat{M}, M)$ and $\lambda$ balances the two terms.

### 4.2 Region-customized Inference

After training, our framework enables users to specify target regions and restoration intensities through structured instructions during inference, as shown in the green flow in Fig. 3. The user instructions follow task-specific templates. For general restoration, we set the template as *"make {region caption} clear with {*$s_1$*}, and make other parts clear with {*$s_2$*}"*; for bokeh-aware restoration, the template is *"make {region caption} clear with {*$s_1$*}, and keep other parts bokeh blur with {*$s_2$*}"*. The {region caption} specifies the textual caption of region of interest (*e.g.*, "the dog on the sand beach"), and $s_1, s_2 \in \mathbb{R}^+$ define the enhancement scales towards fidelity for the target and other regions, respectively. Here, larger values of $s_1$ and $s_2$ result in higher fidelity (closer to the degraded input), while smaller values allow more semantic enhancement. The instruction parsing process extracts three key components: the region caption for SD backbone text conditioning, the main instruction body for ControlNet text encoding, and the fidelity scales $s_1, s_2$ for mask modulation. For region-customized restoration, the region caption is directly used as SD text condition, while for bokeh-aware restoration, it is modified to *"{region caption} in front of bokeh background"* to stimulate the pre-trained SD backbone to generate bokeh blur effect features. Similarly, the main instruction body differs between tasks: *"make {region caption} clear"* for general restoration and *"make {region caption} clear and keep other parts bokeh blur"* for bokeh-aware restoration.

The trained ControlNet branch processes the degraded image and the parsed instruction to generate a mask $M \in [0, 1]$ indicating the target region. It is dynamically resized to match the spatial dimensions of each U-Net upsampling decoder layer, producing masks at multiple scales $\{M_l\}_{l=1}^{L}$. At each layer

$l$, a modulation map is computed as:

$$\mathcal{M}_l = s_1 \cdot M_l + s_2 \cdot (1 - M_l), \tag{2}$$

where $s_1$ and $s_2$ control the fidelity scales for the target and background regions, respectively. This modulation map determines how much ControlNet features contribute to the final output: higher values preserve more original content, while lower values allow more semantic enhancement. The modulated ControlNet features $F_l^{\text{cond}}$ are fused with the base SD features $F_l^{\text{sd}}$ via element-wise multiplication:

$$F_l^{\text{out}} = F_l^{\text{sd}} + \mathcal{M}_l \odot F_l^{\text{cond}}. \tag{3}$$

By applying this modulation progressively across all decoder layers, our framework ensures precise alignment with user intent, *i.e.*, enhancing target regions with intensity $s_1$ while maintaining natural fidelity in non-target areas with intensity $s_2$. Our architecture seamlessly transitions between differently enhanced regions, producing photorealistic restoration results that follow user instructions.

# 5 Experiments

## 5.1 Experiment Settings

**Training details.** Our method is built on SD2.1 [33]. Training data is generated by the data generation engine described in Section 3. The LQ images are obtained by the Real-ESRGAN [39] degradation pipeline. The LQ images and instructions serve as inputs to the model, while the GT images and region masks provide supervision. Our model is first trained on the general degradation dataset for 120K iterations, guided by the instruction template *"make the { region caption } clear "*. The training continues by combining the bokeh dataset with the general degradation dataset for 14k iterations. During this stage, the sampling probability is set to $25\%$ for the general degradation dataset and $75\%$ for the bokeh dataset, which is paired with the instruction template *"make the { region caption } clear and keep other parts bokeh blur."*. The training is conducted on two A100 GPUs with a batch size of 64 and an initial learning rate of $5e^{-5}$. AdamW is adopted as the optimizer for network training.

**Comparison methods.** As the first instruction-based region-customized IR approach, InstructRestore mainly benchmarks against: (1) GAN-based Real-ESRGAN [39]; (2) Diffusion-based methods like StableSR [38], DiffBIR [25], PASD [50], SeeSR [46], SUPIR [52], and OSEDiff [44].

## 5.2 Results on Localized Enhancement

We first show InstructRestore's results with user instructions. Then we demonstrate its precise restoration of specified regions. Finally, we compare it with existing methods.

**Test dataset**. We curate 100 real-world images from multiple sources, including RealSR [5], DRealSR [43], and the RAIM challenge [21], to construct our Instruct100Set. Specifically, we select and crop images with clear semantic region to ensure meaningful evaluation. The foreground masks are generated using the pipeline described in Section 3, ensuring accurate and consistent ROI extraction. The user instructions used in the experiment are in the format of *"make { target region caption } clear with { fidelity level 1 } and keep other parts clear with { fidelity level 2 }."*

**Evaluation metrics.** To comprehensively assess the performance of our method, we adopt both reference-based and no-reference metrics, evaluating both target regions and the entire image. The reference-based metrics include PSNR, SSIM [42] (on the Y channel in YCbCr space) and LPIPS [62]. The no-reference metrics include MANIQA [49], MUSIQ [17] and CLIPIQA [37]. For region-specific evaluation, we compute PSNR and SSIM exclusively within human-specified regions by using the provided GT mask. For other metrics, we zero out pixels outside the target region based on GT mask for computation. This ensures the evaluation focusing on the target regions while being compatible with standard implementations of these metrics.

**Localized enhancement with user-instruction**. We first showcase our method's ability to perform localized enhancement with user-instructions. By specifying the target region and enhancement strength, our method allows users to explicitly control the balance between data fidelity and generative details. As illustrated in Fig. 4, by applying different fidelity scale instructions to the flower region, we successfully adjust the level of details in the flower region while keeping other regions (*e.g.*, leaves and soil) largely unchanged. To our best knowledge, our method is the first one to allow user-instructed local enhancement. To quantitatively validate the instruction-following capability

Table 2: Quantitative evaluation on the instruction following capability of InstructRestore. Experiments are conducted on the Instruct100Set with instruction of "make { region caption of target area} clear with { fidelity scale } and keep other parts clear with 1."

| Fidelity Scale | Target Area | | | | | | Remaining Area | |
|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | CLIPIQA↑ | MUSIQ↑ | MANIQA↑ | PSNR↑ | SSIM↑ |
| 0.5 | 29.71 | 0.7522 | 0.1610 | 0.6801 | 67.86 | 0.6108 | 31.27 | 0.8949 |
| 0.7 | 30.37 | 0.8188 | 0.1439 | 0.6931 | 68.23 | 0.6161 | 31.55 | 0.9047 |
| 0.9 | 30.64 | 0.8494 | 0.1331 | 0.6832 | 67.91 | 0.6091 | 31.61 | 0.9087 |
| 1.1 | 30.73 | 0.8649 | 0.1253 | 0.6659 | 66.92 | 0.5934 | 31.56 | 0.9108 |

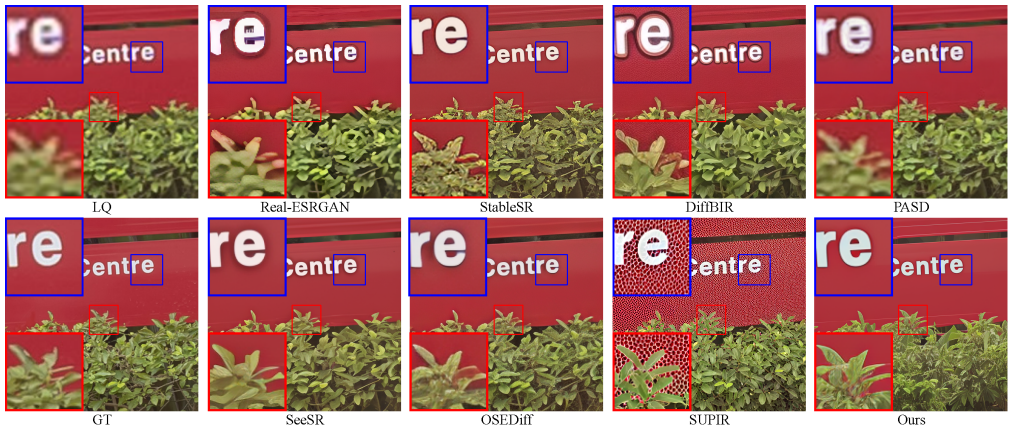


Figure 5: Visual comparison of different methods. We set the instruction as *"make the bush in front of sign clear with 0.5 and keep other parts clear with 0.9"* to keep the fidelity of sign and prioritize detail enhancement of bushes.

of our method, we conduct experiments by varying the enhancement fidelity scales exclusively within the target region while keeping the fidelity scale of the surrounding areas unchanged. For the surrounding regions, we calculate reference-based metrics to assess their stability. The quantitative results are shown in Table 2. We see that when the fidelity scale is small, the non-reference metrics for the target region are significantly higher, indicating that the method tends to generate more details. As the fidelity scale increases, the reference-based metrics (*e.g.*, PSNR and SSIM) improve, while the non-reference metrics gradually decrease. This demonstrates that the method effectively follows the instructions, transitioning from detail-oriented generation to a more input-faithful reconstruction. Furthermore, the PSNR of the target region varies by 1.02 db, while the surrounding regions vary only by 0.29 db. This stark contrast confirms that the enhancement process is localized to the target region, leaving the surrounding areas largely unaffected. Due to space limitations, ablation studies on the mask decoder and feature modulation mechanism are provided in the **appendices**. We also conduct instruction variation experiments in the appendix, testing with instructions that do not follow the standard templates. Although not trained on such variations, our model demonstrates reasonable robustness in generating masks. Please refer to the **appendices** for details.

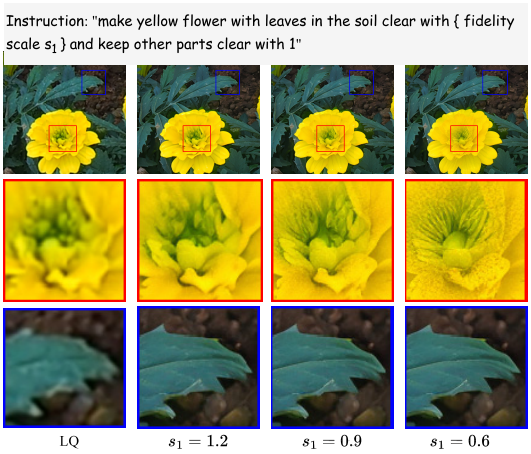


Figure 4: Localized enhancement following instruction on real-world test data. The details in flowers are enhanced gradually while the other regions keeping almost unchanged.

**Comparison with other methods**. We then compare InstructRestore with the competing methods. For images with heavier degradations, we prioritize stronger generative prior to synthesize more details; for regions with high-frequency and irregular textures (*e.g.*, flowers, brushes), we favor generative enhancement to achieve realistic appearances; while for regions with regular structures (*e.g.*, sign and buildings), a conservative enhancement level is selected to avoid unnatural artifacts. As shown in Figure 5, our method can handle well distinct regions, namely the sign and the bushes,

Table 3: Quantitative comparison between our InstructRestore method and other methods on Instruct100Set. The best and second best results of each metric are highlighted in red and blue.

| Method | Target Area | | | | | Full Image | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | CLIPIQA↑ | MUSIQ↑ | MANIQA↑ | PSNR↑ | SSIM↑ | LPIPS↓ | CLIPIQA↑ | MUSIQ↑ | MANIQA↑ |
| RealESRGAN | 31.69 | 0.9065 | 0.7124 | 58.63 | 0.4991 | 27.69 | 0.7871 | 0.3185 | 0.7280 | 60.39 | 0.5030 |
| StableSR | 30.36 | 0.8522 | 0.6707 | 65.75 | 0.5915 | 25.39 | 0.7072 | 0.3001 | 0.7072 | 69.19 | 0.6691 |
| DiffBIR | 30.95 | 0.8804 | 0.6820 | 66.80 | 0.5971 | 26.64 | 0.6897 | 0.3434 | 0.7456 | 69.96 | 0.6609 |
| PASD | 31.80 | 0.9176 | 0.5724 | 61.02 | 0.5323 | 28.37 | 0.7893 | 0.2590 | 0.5768 | 62.92 | 0.5866 |
| SeeSR | 30.90 | 0.8788 | 0.6758 | 67.73 | 0.5974 | 26.75 | 0.7324 | 0.2879 | 0.7246 | 71.49 | 0.6691 |
| SUPIR | 30.74 | 0.8682 | 0.6868 | 62.98 | 0.5655 | 26.29 | 0.6997 | 0.3235 | 0.6840 | 64.40 | 0.6085 |
| OSEDiff | 30.21 | 0.8657 | 0.6417 | 66.75 | 0.5851 | 26.07 | 0.7340 | 0.2870 | 0.7342 | 71.88 | 0.6635 |
| Ours | 30.55 | 0.8368 | 0.6887 | 68.17 | 0.6137 | 25.65 | 0.6999 | 0.3245 | 0.7278 | 71.95 | 0.6809 |

within the same scene. In comparison, methods such as DiffBIR and SUPIR tend to over-enhance the sign, introducing unnecessary artifacts and distortions, while other methods fail to adequately reconstruct the bush, resulting in a smeared and over-smoothed appearance.

Our method allows adjusting the fidelity scale to meet the specific requirements of each region. For example, for the sign, which requires high fidelity, we set the fidelity scale to $0.9$ for faithful restoration. For the bush, we prioritize detail enhancement with a fidelity scale of $0.5$ to generate richer textures. To provide an example of quantitative evaluation, we simply set the fidelity scale for the foreground at $0.8$, while that for other regions to $1$. The evaluation results are reported in Table 3. Since this setting prioritizes generative enhancement in target regions to achieve richer details, it shows better no-reference metrics but relatively lower scores in reference metrics that favor strict fidelity preservation. It is important to note that our InstructRestore enables users to adaptively adjust restoration results based on their preferences. The metrics here only serve as an example to demonstrate that our approach can produce visually pleasing results following user instructions.

### 5.3 Results on Images with Bokeh Effects

In this section, we perform experiments to demonstrate that our method can perform image restoration while preserving bokeh effects and controlling the bokeh intensity.

**Test dataset**. We construct a test dataset by selecting images from two sources: the EBB! dataset [14] and images with bokeh background carefully curated from Pixabay [1]. We select 70 images from them with distinct semantic foregrounds and bokeh background. Masks are generated for the foreground regions to precisely define the ROI, based on which the images are center-cropped to ensure a consistent resolution of $512 \times 512$. Subsequently, we apply Real-ESRGAN [39] degradations to generate LQ and GT image pairs for evaluation. To support instructed interaction, we leverage the masks and GT images to generate foreground descriptions using the pipeline illustrated in Section 3. The user instructions are formatted as *"make {foreground description} clear with {enhancement fidelity level} and keep other parts bokeh blur with the {bokeh level}."*

**Evaluation Metrics.** We compute reference-based metrics for full image and background regions. In addition, we employ D-DFFNet [15], a model for detecting blurred background, to generate the background mask and compute the Intersection-over-Union (IoU) with GT of background mask as a measure of bokeh preservation performance.

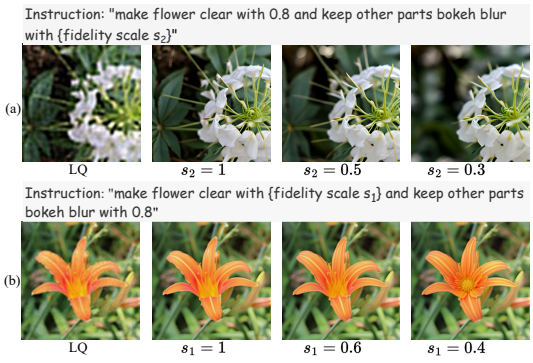


Figure 6: Control of bokeh effect and foreground enhancement. (a) Restoration with controlled bokeh effect while restoring foreground. (b) Restoration with varying foreground enhancement levels while preserving background bokeh.

**Control of bokeh effect**. Our method allows users to specify the desired intensity of bokeh effects and foreground enhancement level via instructions. As illustrated in Fig. 6 (a), our method successfully adjusts the background blur based on user instructions, simulating varying depth-of-field effects while maintaining the sharpness and details in the foreground. More importantly, the adjusted blur is not merely a uniform increase in blur intensity. It faithfully replicates the circular light spots of realistic bokeh, mimicking the optical effects produced by high-quality digital single-lens reflex (DSLR) cameras. As mentioned in Section 4.2, we stimulate the


Figure 7: Visual Comparison of bokeh preservation results between the compared methods.

generation of authentic bokeh features by incorporating "bokeh blur background" text into the SD backbone. So smaller scale values result in less integration of LR input features, leading to increased blur effects, enabling simulation of different depth-of-field effects. In addition to specifying the intensity of bokeh effects, users can further specify the enhancement strength for the foreground, achieving flexible control on the level of details in focal regions. As illustrated in Fig. 6 (b), the fine details in the flower stamens become more pronounced as the instruction changes. Note that such a feature is not supported by existing restoration methods. We also provide quantitative results validating our controllable blur adjustment in the **appendices**.

**Comparison with other methods**. Since the foreground semantics in EBB! mainly include objects like cars and road signs requiring high fidelity, we set the foreground enhancement strength to 1.0. For simplicity and fairness, the bokeh fidelity scale is also set to a default value of 1, representing the weakest depth-of-field effect. The quantitative comparison results are presented in Tab. 4.

Table 4: Quantitative comparison on Bokeh testset

| Method | Background | | | Full Image | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | Bokeh IoU↑ | PSNR↑ | SSIM↑ | LPIPS↓ |
| RealESRGAN | 30.86 | 0.8305 | 0.7203 | 23.69 | 0.7060 | 0.3700 |
| StableSR | 30.24 | 0.8049 | 0.7405 | 22.55 | 0.6305 | 0.3965 |
| DiffBIR | 30.46 | 0.8017 | 0.6289 | 22.20 | 0.5943 | 0.4415 |
| PASD | 31.87 | 0.8453 | 0.8234 | 24.27 | 0.7280 | 0.3523 |
| SeeSR | 30.42 | 0.8149 | 0.7580 | 22.95 | 0.6652 | 0.3677 |
| SUPIR | 29.92 | 0.7847 | 0.7739 | 21.21 | 0.5745 | 0.4375 |
| OSEDiff | 29.89 | 0.8175 | 0.7990 | 22.58 | 0.6707 | 0.3609 |
| Ours | 31.46 | 0.8462 | 0.8482 | 24.69 | 0.7437 | 0.3394 |

Our method demonstrates significantly better performance in fidelity-oriented metrics compared to competing methods, reflecting its ability to accurately approximate the GT's bokeh characteristics. In contrast, existing methods fail to preserve bokeh effects, leading to deviations from the GT. Visual comparison is shown in Fig. 7. More comparisons are provided in the **appendices**. Competing methods tend to restore the background with sharp details, disrupting the bokeh effect, whereas our method preserves natural background blur while enhancing foreground details, ensuring both fidelity and artistic quality.

## 6 Conclusion

We presented InstructRestore, the first framework for region-customized image restoration guided by human instructions. To support this task, we designed a scalable data annotation engine and constructed a dedicated dataset comprising 536,945 triplets, each containing a high-quality image, the region mask and region caption. Building on this dataset, we developed an InstructRestore model that parsed human instructions to achieve region-specific restoration. Our framework allowed users to apply distinct enhancement intensities to different regions and adjust background bokeh effects. By enabling fine-grained control via user instructions, our work advanced research in interactive image restoration and enhancement techniques.

**Limitations.** While InstructRestore offers a baseline for region-customized restoration guided by human instructions, it has several limitations. Currently, it lacks support for instance-level object specification, which requires instance-level masks and captions. Moreover, users are recommended to follow a predefined instruction format, though an off-the-shelf LLM can convert free-form inputs. Furthermore, although our method achieves competitive results, it focuses more on localized customization, while it deserves further exploration of global quality optimization. Reducing the number of inference steps is also worth exploring. Addressing these limitations would boost the applicability and performance of user-instructed image restoration in real world scenarios.

## References


[1] Pixabay: 5.3 million+ stunning free images to use anywhere. https://pixabay.com/.

[2] Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. *Adv. Neural Inform. Process. Syst.*, 2024.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18392–18402, 2023.

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Int. Conf. Comput. Vis.*, pages 3086–3095, 2019.

[6] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Eur. Conf. Comput. Vis.*, pages 17–33, 2022.

[7] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *Eur. Conf. Comput. Vis.*, pages 1–21, 2024.

[8] Junyuan Deng, Xinyi Wu, Yongxing Yang, Congchao Zhu, Song Wang, and Zhenyao Wu. Acquire and then adapt: Squeezing out text-to-image model for image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23195–23206, 2025.

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, 2015.

[10] 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. *arXiv preprint arXiv:2309.17102*, 2023.

[11] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6986–6996, 2024.

[12] 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.

[13] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8362–8371, 2024.

[14] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 418–419, 2020.

[15] Yuxin Jin, Ming Qian, Jincheng Xiong, Nan Xue, and Gui-Song Xia. Depth and dof cues make a better defocus blur detector. In *Int. Conf. Multimedia and Expo*, pages 882–887, 2023.

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Conf. Empir. Methods Nat. Lang. Process.*, pages 787–798, 2014.

[17] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Int. Conf. Comput. Vis.*, pages 5148–5157, 2021.

[18] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *Eur. Conf. Comput. Vis.*, pages 467–484, 2024.

[19] Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu, Jianzhuang Liu, et al. Zone: Zero-shot instruction-guided local editing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6254–6263, 2024.

[20] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1775–1787, 2023.

[21] Jie Liang, Radu Timofte, Qiaosi Yi, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Lei Zhang, Yibin Huang, et al. Ntire 2024 restore any image model (raim) in the wild challenge. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6632–6640, 2024.

[22] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5657–5666, 2022.

[23] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Int. Conf. Comput. Vis.*, pages 1833–1844, 2021.

[24] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[25] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *Eur. Conf. Comput. Vis.*, pages 430–448, 2024.

[26] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 457–466, 2022.

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11–20, 2016.

[28] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3883–3891, 2017.

[29] JoonKyu Park, Sanghyun Son, and Kyoung Mu Lee. Content-aware local gan for photo-realistic super-resolution. In *Int. Conf. Comput. Vis.*, pages 10585–10594, 2023.

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[31] Chenyang Qi, Zhengzhong Tu, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Qifeng Chen, and Hossein Talebi. Spire: Semantic prompt-driven image restoration. In *Eur. Conf. Comput. Vis.*, pages 446–464, 2024.

[32] Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. In *Int. Conf. Comput. Vis.*, October 2023.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022.

[34] Lingchen Sun, Jie Liang, Shuaizheng Liu, Hongwei Yong, and Lei Zhang. Perception-distortion balanced super-resolution: A multi-objective optimization perspective. *IEEE Trans. Image Process.*, 2024.

[35] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. *arXiv preprint arXiv:2412.03017*, 2024.

[36] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8174–8182, 2018.

[37] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, pages 2555–2563, 2023.

[38] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *Int. J. Comput. Vis.*, pages 1–21, 2024.

[39] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Int. Conf. Comput. Vis.*, pages 1905–1914, 2021.

[40] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Eur. Conf. Comput. Vis. Worksh.*, pages 0–0, 2018.

[41] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25796–25805, 2024.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[43] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Eur. Conf. Comput. Vis.*, pages 101–117, 2020.

[44] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024.

[45] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Adv. Neural Inform. Process. Syst.*, 37:92529–92553, 2025.

[46] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25456–25467, 2024.

[47] Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, and Chao Dong. Desra: detect and delete the artifacts of gan-based real-world super-resolution models. *arXiv preprint arXiv:2307.02457*, 2023.

[48] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[49] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1191–1200, 2022.

[50] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Eur. Conf. Comput. Vis.*, pages 74–91, 2024.

[51] Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Yuhui Wu, and Lei Zhang. Fine-structure preserved real-world image super-resolution via transfer vae training. *arXiv preprint arXiv:2507.20291*, 2025.

[52] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25669–25680, 2024.

[53] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Eur. Conf. Comput. Vis.*, pages 69–85, 2016.

[54] Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785*, 2024.

[55] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 28202–28211, 2024.

[56] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23153–23163, 2025.

[57] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Int. Conf. Comput. Vis.*, pages 4791–4800, 2021.

[58] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Adv. Neural Inform. Process. Syst.*, 36, 2024.

[59] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.*, 26(7):3142–3155, 2017.

[60] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Trans. Image Process.*, 27(9):4608–4622, 2018.

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pages 3836–3847, October 2023.

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to content from line 73 to line 79

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to content from line 366

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: the paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Have described the detail of data and method

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data is too big to be zipped as single file. Open access for data and code will be offered after review. We offer details in main paper and appendices.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have described the detail

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the type of compute resources in main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research in the paper fully conforms to the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we analyze the potential social impact in appendix

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we clearly indicate the baseline methods and data used in the paper. Their licenses permit use with academic scope.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce the detail of data and model, and will release all if accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Please refer to section 3 in main paper and corresponding content in appendix.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.