RAP-SR: RestorAtion Prior Enhancement in Diffusion Models for Realistic Image Super-Resolution

Jiangang Wang[1,2], Qingnan Fan[2†], Jinwei Chen[2], Hong Gu[2], Feng Huang[3], Wenqi Ren[1†]

[1]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[2]vivo Mobile Communication Co. Ltd
[3]School of Mechanical Engineering and Automation, Fuzhou University
wangjg33@mail2.sysu.edu.cn, fqnchina@gmail.com, renwq3@mail.sysu.edu.cn
Project Page: https://github.com/W-JG/RAP-SR

Abstract

Benefiting from their powerful generative capabilities, pretrained diffusion models have garnered significant attention for real-world image super-resolution (Real-SR). Existing diffusion-based SR approaches typically utilize semantic information from degraded images and restoration prompts to activate prior for producing realistic high-resolution images. However, general-purpose pretrained diffusion models, not designed for restoration tasks, often have suboptimal prior, and manually defined prompts may fail to fully exploit the generated potential. To address these limitations, we introduce RAP-SR, a novel restoration prior enhancement approach in pretrained diffusion models for Real-SR. First, we develop the High-Fidelity Aesthetic Image Dataset (HFAID), curated through a Quality-Driven Aesthetic Image Selection Pipeline (QDAISP). Our dataset not only surpasses existing ones in fidelity but also excels in aesthetic quality. Second, we propose the Restoration Priors Enhancement Framework, which includes Restoration Priors Refinement (RPR) and Restoration-Oriented Prompt Optimization (ROPO) modules. RPR refines the restoration prior using the HFAID, while ROPO optimizes the unique restoration identifier, improving the quality of the resulting images. RAP-SR effectively bridges the gap between general-purpose models and the demands of Real-SR by enhancing restoration prior. Leveraging the plug-and-play nature of RAP-SR, our approach can be seamlessly integrated into existing diffusion-based SR methods, boosting their performance. Extensive experiments demonstrate its broad applicability and state-of-the-art results. *Codes and datasets will be available upon acceptance.*

1 Introduction

Image super-resolution (SR) is a fundamental task in computer vision, aiming to reconstruct high-resolution (HR) images from low-resolution (LR) inputs, with broad applications in mobile photography (Chen et al. 2019), autonomous driving (Li et al. 2023b), and robotics (Wang et al. 2021a). SR remains a highly ill-posed problem due to the complexity and variability of degradation models in real-world scenarios. Early SR solutions focus on improving fidelity (Dong et al. 2014; Kim, Lee, and Lee 2016; Haris, Shakhnarovich, and Ukita 2018) by employing pixel-level losses such as ℓ_1

---

† Corresponding author.
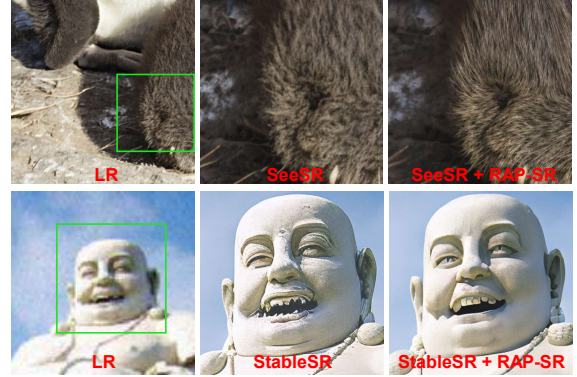This work was completed during an internship at vivo.

Figure 1: Visual Comparison: RAP-SR enhances the restoration prior of pretrained diffusion models. Our proposed RAP-SR method can be seamlessly integrated into diffusion-based SR methods, generating more realistic details and textures without the need for fine-tuning the original model.

and MSE, often resulting in over-smoothed details (Liang, Zeng, and Zhang 2022). Advanced architectures (Liang et al. 2021) have improved performance, but issues like artifacts and poor visual quality remain when applied to real-world scenarios. To address this, Real-SR methods aim to reproduce realistic details by optimizing both fidelity and perceptual quality, often employing generative adversarial networks (GANs) (Goodfellow et al. 2014). However, GAN-based approaches suffer from problems such as model collapse and difficult training (Li et al. 2022).

Recently, diffusion models (Ho, Jain, and Abbeel 2020) have gained prominence in image generation, leading to the development of large-scale pretrained text-to-image (T2I) diffusion models, such as Stable Diffusion (SD). For a wide range of natural images, diffusion prior is more effective than GAN-based prior. Additionally, these models have shown significant potential in various downstream low-level vision tasks, including image editing (Meng et al. 2022; Avrahami, Lischinski, and Fried 2022), image restoration (Lugmayr et al. 2022; Chung, Sim, and Ye 2022), and image-to-image translation (Song et al. 2021; Saharia et al. 2022). Methods such as StableSR (Wang et al. 2024), PASD (Yang et al. 2024), DiffBIR (Lin et al. 2024), SUPIR (Yu et al. 2024), and SeeSR (Wu et al. 2024) lever-

age pretrained T2I models to tackle the Real-SR problem by capturing semantic structures from LR images and using handcrafted restoration prompts to activate restoration prior for generating realistic HR images.

However, employing pretrained diffusion models for Real-SR tasks presents two primary challenges: the inadequacy of restoration prior and inaccuracies in prompt activation. General pretrained diffusion models are not inherently designed for restoration tasks. While these models have strong prior knowledge and can generate images of varying quality, their limited restoration prior hinders their ability to produce high-quality, rich-detail images (Dai et al. 2023; Chen et al. 2024). Moreover, previous diffusion-based SR methods often rely on manually crafted restoration prompts to activate restoration prior. Natural language often fails to accurately describe image quality under multiple degradations, leading to incorrectly activated restoration prior.

To address these limitations, we propose **RAP-SR**, a novel restoration prior enhancement approach in pretrained diffusion models for Real-SR. Firstly, we develop the *High-Fidelity Aesthetic Image Dataset (HFAID)* using a Quality-Driven Aesthetic Image Selection Pipeline (QDAISP). Although large-scale datasets are available, the images within these datasets often suffer from poor and inconsistent quality due to varied purposes. We propose a meticulous four-stage image selection process (QDAISP) facilitated by a large-scale multi-modality model to filter images by evaluating both the image quality and aesthetic attributes. As a result, HFAID consists of 5,000 high-fidelity and aesthetic images from a pool of 1 million, surpassing the quality of all existing datasets tailored for image restoration. This dataset serves as the foundation for enhancing the prior of pretrained models, enabling the transition from low-quality image generation to high-quality output production. Secondly, we establish a *Restoration Prior Enhancement framework*, including restoration prior refinement (RPR) and restoration-oriented prompt optimization (ROPO). RPR refines the restoration prior by fine-tuning the model using HFAID. ROPO optimizes specific identifiers during the prior refinement phase. The method combines unique identifiers with the image's semantic caption, strengthening the association between prompt and image quality, and enabling accurate activation of the restoration prior. As a result, our framework effectively strengthens the restoration prior, bridging the gap between general-purpose models and Real-SR tasks.

RAP-SR's plug-and-play design allows it to be seamlessly integrated with existing diffusion-based SR methods, such as StableSR (Wang et al. 2024), DiffBIR (Lin et al. 2024), and SeeSR (Wu et al. 2024), improving both visual quality and objective metrics. Overall, our contribution is summarized as follows:

- We collected the High-Fidelity Aesthetic Image Dataset (HFAID), which surpasses existing datasets not only in

fidelity but also in aesthetic quality, effectively enhancing the priors of diffusion models.
- We proposed a Prior Restoration Enhancement framework, which includes the Restoration Prior Refinement (RPR) and Restoration-Oriented Prompt Optimization (ROPO) modules, designed to improve and accurately activate the model's restoration priors.
- Our method can be seamlessly integrated to improve existing diffusion-based SR methods. Extensive experiments demonstrate its broad applicability and excellent performance.

2 Related Work

2.1 Real-world Image Super Resolution

Deep learning has emerged as the predominant approach for SR tasks, with the pioneering work of SRCNN (Dong et al. 2014) utilizing deep neural networks for SR. Subsequent methods that incorporate residual connections and attention mechanisms (Liang et al. 2021; Kim, Lee, and Lee 2016; Haris, Shakhnarovich, and Ukita 2018) often aim to minimize fidelity loss through pixel-level supervised loss functions. However, this approach typically results in overly smoothed details (Liang, Zeng, and Zhang 2022). Recent research has shifted towards addressing Real-SR challenges, which involve complex and unknown degradation processes. Some researchers propose collecting real-world LR and HR paired data to train networks (Cai et al. 2019; Wei et al. 2020), although this approach can be costly. Alternatively, other methods focus on synthesizing realistic data pairs for training. Notably, BSRGAN (Zhang et al. 2021) and RealESRGAN (Wang et al. 2021b) offer efficient degradation pipelines for Real-SR. Given the generative capabilities of GAN-based methods (Goodfellow et al. 2014), they have become dominant in Real-SR tasks. Adversarial training improves the perceptual quality of images. However, GANs face limitations such as training instability and model collapse, which often result in unnatural artifacts (Li et al. 2022). Consequently, recent research has begun exploring advanced generative models, such as diffusion models, which are capable of generating high-quality, detailed images.

2.2 Diffusion Model

Diffusion models utilize Markov chains to transform latent variables into complex data distributions, as exemplified by DDPM (Ho, Jain, and Abbeel 2020) and its accelerated variant, DDIM (Song, Meng, and Ermon 2021). The Latent Diffusion Model (LDM)(Rombach et al. 2022) achieves impressive results with reduced computational costs. These advancements enable large-scale pretrained text-to-image (T2I) models like Stable Diffusion (SD) and ImgGen. ControlNet (Zhang, Rao, and Agrawala 2023) allows for external control over the generation process, while EMU (Dai et al. 2023) enhances aesthetic quality through fine-tuning. InstructPix2Pix (Brooks, Holynski, and Efros 2023) refines T2I models using editing instructions. Diffusion models excel in various image generation tasks, including restora-

---

e.g., SeeSR positive prompt: "clean, high-resolution, 8k"; negative prompt: "dotted, noise, blur, lowres, smooth". SUPIR positive prompt: "cinematic, high contrast, highly detailed, 32k, ultra HD, extreme meticulous detailing, *etc*"; SUPIR negative prompt: "painting, oil painting, illustration, drawing, art, sketch, *etc*".
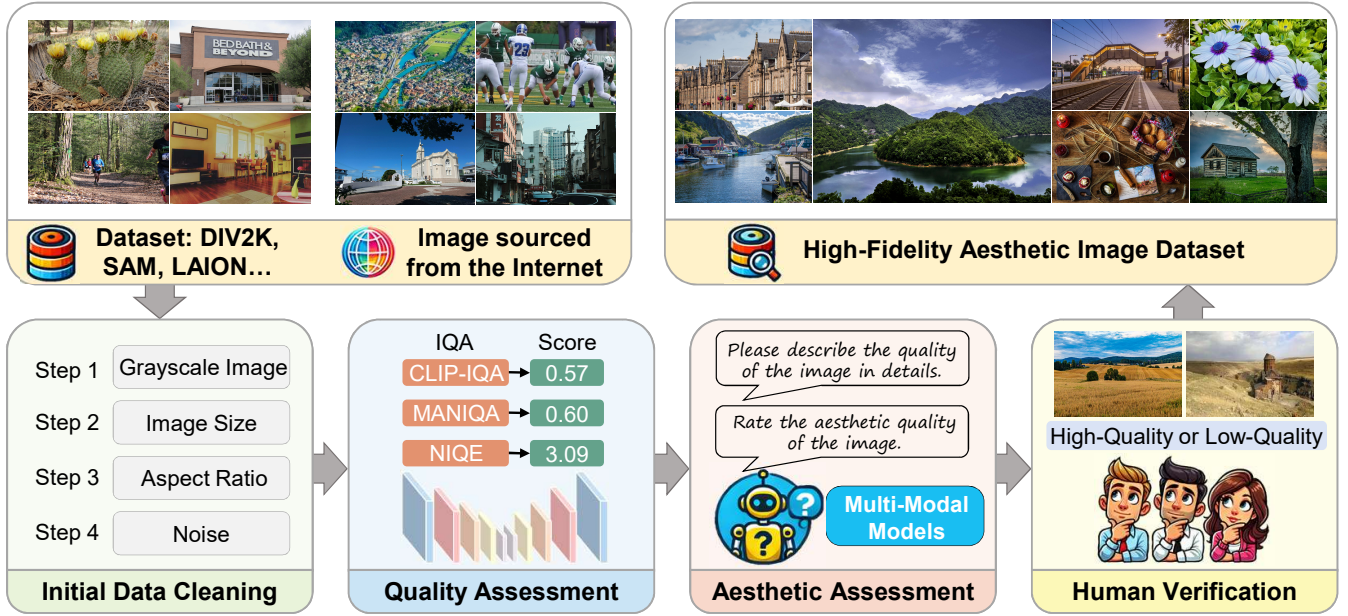
Figure 2: Quality-Driven Aesthetic Image Selection Pipeline. This process is divided into four stages. Unlike previous methods that focus solely on image quality, our approach incorporates the multi-modality model to evaluate both image quality and aesthetic performance. Ultimately, we meticulously select 5,000 ultra-high-quality images from the initial pool of one million images to create the High-Fidelity Aesthetic Image Dataset.

tion (Lugmayr et al. 2022; Chung, Sim, and Ye 2022), editing, and colorization (Song et al. 2021; Saharia et al. 2022).

2.3 Diffusion-Based Super-Resolution

Diffusion-based SR methods can be categorized into three main types. The first type adjusts the inverse process of pretrained diffusion models using gradient descent (Wang, Yu, and Zhang 2023; Kawar et al. 2022; Fei et al. 2023). These methods do not require retraining but assume a predefined image degradation model, limiting their applicability in Real-SR scenarios. The second type involves training the diffusion model on paired data (Rombach et al. 2022; Shang et al. 2024; Yue, Wang, and Loy 2023), but the restoration quality is heavily dependent on the quantity of training data, constraining their potential to achieve exceptional results. The third type leverages the robust generative prior of large-scale pretrained diffusion models by introducing adapters for control (Wu et al. 2024; Yu et al. 2024; Yang et al. 2024). By utilizing LR images as control information, pretrained diffusion models can produce high-quality results, making this approach the mainstream for diffusion model-based SR methods. StableSR (Wang et al. 2024) balances fidelity and perceptual quality by incorporating a time-aware encoder and feature warping. DiffBIR (Lin et al. 2024) employs SwinIR (Liang et al. 2021) for initial degradation removal, enhancing details with a diffusion model. PASD (Yang et al. 2024) utilizes semantic models like ResNet (He et al. 2016) to extract information from LR images, thereby bolstering the generative capability of T2I models. SeeSR (Wu et al. 2024) improves T2I model generation using tags and additional conditions. CoSeR (Sun et al. 2024) augments T2I model generation by providing reference images from LR

inputs. SUPIR (Yu et al. 2024) achieves superior outcomes through the use of a larger diffusion model, coupled with a robust language model. These methods guide T2I diffusion models in generating high-quality HR images by extracting additional semantic information from degraded LR images. However, they often overlook the restoration prior inherent in pretrained diffusion models, which are crucial for image reconstruction tasks.

3 Methodology

3.1 High-Fidelity Aesthetic Image Dataset

Observation and Motivation Previous research (Dai et al. 2023; Chen et al. 2024) indicates that training large-scale T2I diffusion models involves multiple phases. The initial phase focuses on aligning text and images, where the diffusion model establishes a mapping between the two by leveraging billions of text-image pairs. The subsequent phase, known as quality-tuning, aims to enhance image quality. Once text-image alignment is achieved, the pretrained model can be fine-tuned using a small dataset tailored to specific task domains. In the Real-SR task, it is particularly important to enhance the model's restoration prior through quality-tuning.

During the quality-tuning phase, the quality of the dataset is crucial to the effectiveness of model training (Dai et al. 2023). An ideal dataset should contain high-quality, detail-rich images with informative captions. However, existing datasets like LAION-5B (Schuhmann et al. 2022) suffer from poor image quality, incomplete captions, and misalignment between images and text. Widely adopted datasets such as SAM (Kirillov et al. 2023), COCO (Lin et al. 2014), and
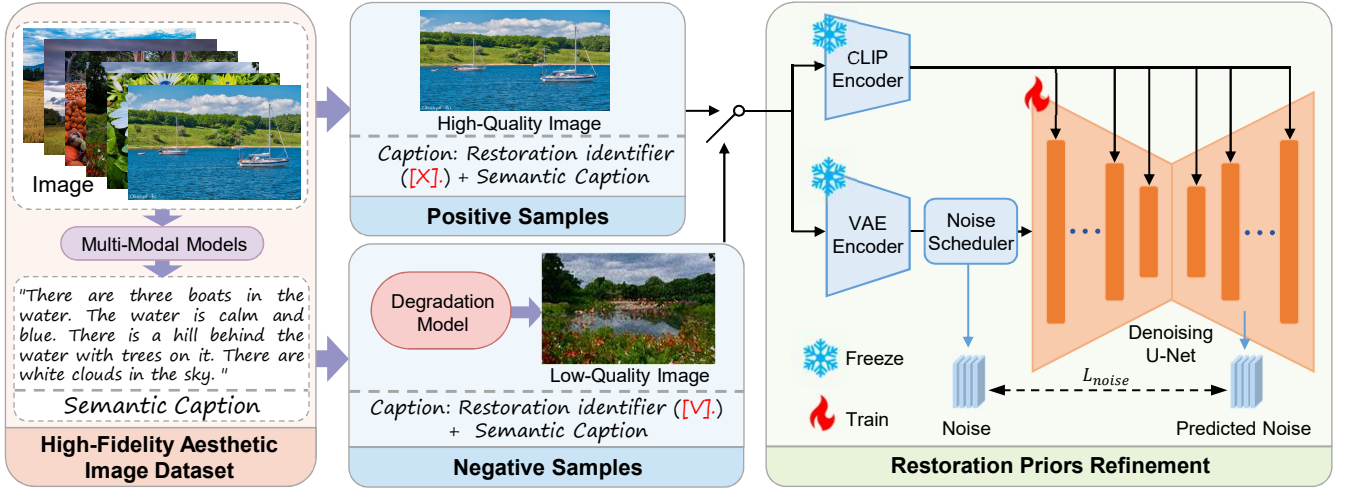
Figure 3: Restoration Priors Enhancement Framework: This framework includes Restoration Priors Refinement (RPR) and Restoration-Oriented Prompt Optimization (ROPO). ROPO optimizes the restoration prompt by constructing both positive and negative samples. For negative samples, a degradation model generates low-quality images and then combines the unique restoration identifier with the image's semantic caption to create training data. Through the subsequent RPR process, the model enhances its restoration prior, learning to associate image quality with the restoration identifier.

ImageNet (Deng et al. 2009) also have low quality and inadequate labeling. Although datasets such as DIV2K (Agustsson and Timofte 2017), Flickr2K (Timofte et al. 2017), and LSDIR (Li et al. 2023c) provide relatively high quality, they still do not meet ultra-high-quality standards due to quality inconsistencies in their datasets.

Quality-Driven Aesthetic Image Selection Pipeline To address these issues, we curate a high-fidelity aesthetic image dataset (HFAID) by selecting 5,000 ultra-high-quality images from an initial pool of 1 million images. To effectively filter the image data, we design a quality-driven aesthetic image selection pipeline that considers both image quality and aesthetic performance. As shown in Figure 2, the selection process is divided into four stages: Initial Data Cleaning, Quality Assessment, Aesthetic Assessment, and Human Verification.

Initially, we collect approximately 1 million images from existing datasets and publicly available online sources. Initial data cleaning is performed, including checks for grayscale images, verification of image size and aspect ratio, and the use of Laplacian variance (Li et al. 2023c) to detect image noise. In the second stage, to accurately assess image quality, we employ state-of-the-art no-reference image quality assessment metrics. The currently available no-reference image quality metrics (e.g., CLIP-IQA, MANIQA, MUSIQ, NIQE, BRISQUE, etc.) each focus on different aspects. Our goal is to select extremely high-quality data that meets human aesthetic standards. Therefore, the choice of metrics is crucial. We first select 200 images with the best and worst performance under each metric from the LSDIR dataset and conduct a 10-person user evaluation, ultimately selecting MANIQA (Yang et al. 2022), CLIP-IQA (Wang, Chan, and Loy 2023), and NIQE (Zhang, Zhang, and Bovik 2015) as the core evaluation metrics.

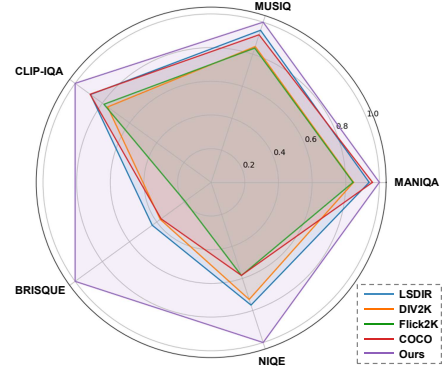Previous image restoration datasets primarily focus on

Figure 4: Comparison of No-Reference Metrics Across Different Datasets. Our proposed dataset significantly outperforms existing datasets across all evaluation metrics.

image quality and detail richness but lack exploration of aesthetic evaluation. The aesthetic quality of images is equally crucial for image generation tasks. Studies have shown that multi-modality models have surpassed traditional models in the field of image understanding (Li et al. 2023a). In the third phase, we use existing multi-modality models for aesthetic evaluation, utilizing the mPLUG-Owl2 (Ye et al. 2024) model to query images and obtain precise aesthetic evaluation metrics. In the final phase, we conduct human verification to accurately assess the quality of each image, ensuring that each image is evaluated by at least two people. We provide a detailed explanation of the selection process in the supplementary material.

Comparison With Other Datasets As illustrated in figure 2, our dataset excels in both image quality and detail richness. We evaluate the quality of our dataset using five no-reference image quality assessment metrics: MANIQA, MUSIQ, CLIP-IQA, BRISQUE, and NIQE. Fig-

ure 4 presents the results, showing that our dataset significantly outperforms others across all metrics. Additional examples are provided in the supplementary material.

3.2 Restoration Priors Enhancement Framework

Restoration Priors Refinement   High-quality image captions are also crucial for training diffusion models. To generate text labels with high information density, we utilize the advanced vision-language model Florence-2 (Xiao et al. 2024), a large-scale end-to-end multi-modality model. Leveraging its image understanding capabilities, we produce high-quality text labels. The quality of our labels surpasses that of existing text-image datasets (Schuhmann et al. 2022), with further details provided in the supplementary materials.

Since the pre-trained diffusion model has already completed text-image alignment, we can achieve quality-tuning with high-quality image-text data in a short period, thereby enhancing the model's ability to restoration prior knowledge. Using a smaller batch size of 40, the model converges within 3,000 steps, significantly reducing training time. Additionally, we find that both the quantity and quality of data significantly influence the tuning effect, which will be discussed in detail in section 4.2.

Restoration-Oriented Prompt Optimization   Previous diffusion-based SR models often rely on manually designed restoration prompts to activate restoration prior, frequently resulting in defocused images and artifacts that degrade the overall quality of image restoration. DreamBooth (Ruiz et al. 2023) introduces a novel method for fine-tuning pre-trained diffusion models by associating a unique identifier with a specific object using a small number of images, enabling the generation of realistic images that accurately represent the object. While DreamBooth focuses on binding specific object concepts to pretrained diffusion models, we extend its application to restoration prompt optimization for image quality. We develop a restoration-oriented prompt optimization method to precisely activate the model's restoration prior. The method is shown in figure 3.

Given the challenges of fully expressing image quality through natural language, especially under conditions of multiple degradations, we create new identifiers to represent various levels of image quality. During the restoration prior enhancement phase in diffusion models, we redefine the categories of high-quality and degraded low-quality image data to better align with the needs of models.

For the high-quality category, we utilize high-quality image data and combine positive restoration identifiers with the semantic caption of the images to form positive samples. These positive samples serve as a benchmark for the model to understand what constitutes high-quality imagery. In contrast, for the low-quality category, we first generate degraded low-quality image data using a sophisticated image degradation model. We then combine negative restoration identifiers with the semantic captions of the original images to create negative samples. By merging these unique identifiers with their semantic captions, we effectively link image quality with image semantics, making it easier for the model to distinguish between different quality levels.

Algorithm 1: Restoration-oriented Prompt Optimization Algorithm

Require: Ultra-high-quality Dataset of image-text pairs $S = \{(x_i, c_i)\}_{i=1}^N$, diffusion model f_θ, degradation model d_μ, number of timesteps T, noise schedule $\{\beta_t\}_{t=1}^T$, positive identifier c_p, negative identifier c_n, positive ratio r, learning rate η

1: Initialize model parameters θ from a pre-trained model
2: **for** each (x_i, c_i) in S **do**
3:    Sample a random value $u \sim \text{Uniform}(0, 1)$
4:    **if** $u < r$ **then**
5:       Append positive identifier:
         $c_i \leftarrow \text{concatenate}(c_p, c_i)$
6:    **else**
7:       Append negative identifier:
         $c_i \leftarrow \text{concatenate}(c_n, c_i)$
8:       Degrade the image: $x_i \leftarrow d_\mu(x_i)$
9:    **end if**
10:    Sample timestep $t \sim \text{Uniform}(1, T)$
11:    Sample noise $\epsilon \sim \mathcal{N}(0, I)$
12:    Compute noisy image: $x_t \leftarrow \sqrt{\bar{\alpha}_t} \cdot x_i + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon$,
       where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$
13:    Predict noise: $\hat{\epsilon}_\theta \leftarrow f_\theta(x_t, t, c_i)$
14:    Compute loss: $L \leftarrow \|\epsilon - \hat{\epsilon}_\theta\|^2$
15:    Update model parameters: $\theta \leftarrow \theta - \eta \nabla_\theta L$
16: **end for**

During the training process, we perform a random sampling of positive and negative samples according to a predetermined ratio r. This is crucial for fine-tuning the pre-trained diffusion model, enabling it to learn how to associate high-quality and low-quality images with their corresponding identifiers. As a result, the model becomes adept at accurately activating the restoration prior, which significantly enhances its ability to generate high-quality images from degraded inputs. This method ensures a more reliable and consistent image restoration process across various scenarios. The pseudo-code of our restoration-oriented prompt optimization algorithm is summarized as Algorithm 1.

During inference, we adopt a classifier-free guidance strategy, which enables the diffusion model to generate higher-quality images using negative prompts without additional training. At each inference step, we calculate the positive prompt c_{pos} and negative prompt c_{neg} and mix these predictions to obtain the final output:

$$\hat{\epsilon} = \epsilon_\theta(z_{lr}^t, t, c_{pos}, x_{lr}), \tag{1}$$

$$\hat{\epsilon}_{\text{neg}} = \epsilon_\theta(z_{lr}^t, t, c_{neg}, x_{lr}), \tag{2}$$

$$\tilde{\epsilon} = \hat{\epsilon} + \lambda_s(\hat{\epsilon} - \hat{\epsilon}_{\text{neg}}). \tag{3}$$

where λ_s is the guidance scale and z_{lr}^t represents the noise potential of the low-resolution image. In practice, We employ unique identifiers defined during our training as positive prompts c_{pos} and negative prompts c_{neg} to generate higher-quality images.

Datasets	Metric	StableSR	Stable + RAP-SR	DiffBIR	DiffBIR + RAP-SR	SeeSR	SeeSR + RAP-SR
DIV2K	PSNR ↑	23.28	**23.85**	23.66	**23.83**	**23.70**	23.59
	SSIM ↑	0.5733	**0.5808**	0.5651	**0.5694**	**0.6052**	0.5897
	LPIPS ↓	**0.3118**	0.3542	**0.3516**	0.3536	**0.3168**	0.3501
	MANIQA ↑	**0.6193**	0.6017	0.6211	**0.6255**	0.6246	**0.6271**
	MUSIQ ↑	65.85	**66.32**	65.77	**66.90**	68.66	**68.86**
	CLIPIQA ↑	0.6771	**0.7517**	0.6693	**0.6928**	0.6936	**0.7254**
	BRISQUE ↓	15.62	**11.84**	14.66	**8.69**	20.41	**16.08**
RealSR	PSNR ↑	24.60	**25.03**	24.81	**25.06**	**25.00**	24.85
	SSIM ↑	0.7045	**0.7081**	0.6595	**0.6663**	**0.7187**	0.7003
	LPIPS ↓	**0.3096**	0.3457	0.3608	**0.3564**	**0.3096**	0.3388
	MANIQA ↑	0.6252	**0.6259**	0.6238	**0.6349**	0.6456	**0.6512**
	MUSIQ ↑	66.00	**67.39**	64.93	**67.46**	69.93	**70.0644**
	CLIPIQA ↑	0.6315	**0.6956**	0.6448	**0.6911**	0.6553	**0.7217**
	BRISQUE ↓	19.51	**15.39**	18.98	**13.36**	29.06	**19.01**
DrealSR	PSNR ↑	27.99	**28.3019**	26.82	**27.15**	27.98	**28.01**
	SSIM ↑	0.7504	**0.7566**	0.6633	**0.6858**	**0.7719**	0.7600
	LPIPS ↓	**0.3275**	0.3644	0.4497	**0.4365**	**0.3243**	0.3542
	MANIQA ↑	0.5620	**0.5667**	0.5924	**0.6173**	0.5941	**0.6161**
	MUSIQ ↑	59.03	**60.70**	60.85	**63.7761**	64.95	**65.08**
	CLIPIQA ↑	0.6385	**0.6752**	0.6369	**0.6729**	0.6757	**0.7075**
	BRISQUE ↓	21.39	**17.36**	22.60	**15.60**	32.12	**23.21**

Table 1: Quantitative comparison. The best results of each group are highlighted in **bold**. ↑ and ↓ mean that the larger or smaller score is better, respectively.

4   Experiments

Enhancing Restoration Prior For Pretrained Diffusion Model.   We conduct the proposed restoration prior enhancement experiments using Stable Diffusion 2.1 and our HFAID dataset. As described in Section 3.2, we use the Florence-2 (Xiao et al. 2024) model to generate highly informative text labels. During the enhancement process, the images are resized to 512 pixels on the longer side and center-cropped. Low-quality images are synthesized through the RealESRAGN (Wang et al. 2021b) pipeline. In the optimization process, we employ the AdamW (Loshchilov and Hutter 2019) optimizer with a learning rate of 5e-5 and train the model using two NVIDIA L40 GPUs with a batch size of 40 for 3,000 iterations. The ratio r of high-quality (HQ) to low-quality (LQ) images is set at 0.8, where '[X]' denotes positive identifier and '[V]' denotes negative identifier. To maintain the consistency of CFG (Ho and Salimans 2022), there is a 5% chance of leaving the text labels empty.

Evaluation Setting.   To evaluate the proposed restoration prior enhancement method, we test three methods: SeeSR, DiffBIR, and StableSR. The results are obtained for SeeSR + RAP-SR, DiffBIR + RAP-SR, and StableSR + RAP-SR. All these methods use Stable Diffusion as the pre-trained model, with their original models replaced by our RAP model. **It is important to note that we do not perform any fine-tuning on these replaced models.** During the evaluation, we conduct comprehensive tests on synthetic data from the DIV2K val dataset and real-world datasets RealSR (Cai et al. 2019) and DrealSR (Wei et al. 2020). In the tests, the resolution of high-resolution images is set to 512 × 512, while low-resolution images are cropped to 128 × 128.

_____
https://huggingface.co/stabilityai/stable-diffusion-2-1

Evaluation Metrics.   To better align with human perception, we use seven evaluation metrics: PSNR, SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018), MANIQA (Yang et al. 2022), MUSIQ (Ke et al. 2021), BRISQUE (Shao and Mou 2021) and CLIPIQA (Hessel et al. 2021). PSNR and SSIM measure pixel-level differences, while LPIPS assesses perceptual distances. MANIQA, MUSIQ, BRISQUE, and CLIPIQA are no-reference image quality metrics. Previous studies have shown that reference-based metrics have a weaker correlation with human perception of image quality in real-world scenarios (Yu et al. 2024; Wang et al. 2024). A discussion of the test metrics is detailed in the supplementary material.

4.1   Comparison with state-of-the-arts

Quantitative Comparison   Table 1 provides quantitative comparisons on three synthetic and real-world datasets. We have the following observations. Firstly, the method we proposed has achieved great improvements in almost all no-reference metrics such as MANIQA, MUSIQ, CLIPIQA, and BRISQUE on all three data sets. This shows that our method significantly improves the image generation capabilities of the original method and can generate richer details. Secondly, for reference metrics such as PSNR, SSIM, and LPIPS, our method only improves the original performance on some data sets. This is primarily because the DM-based method generates more realistic details, which impact these metrics. Overall, our PAR-SR achieves better no-reference metric scores while maintaining competitive full-reference metric scores.

Qualitative Comparison   Figure 5 shows visual examples from synthetic and real-world datasets. The visual results are consistent with the quantitative findings: our model sig-
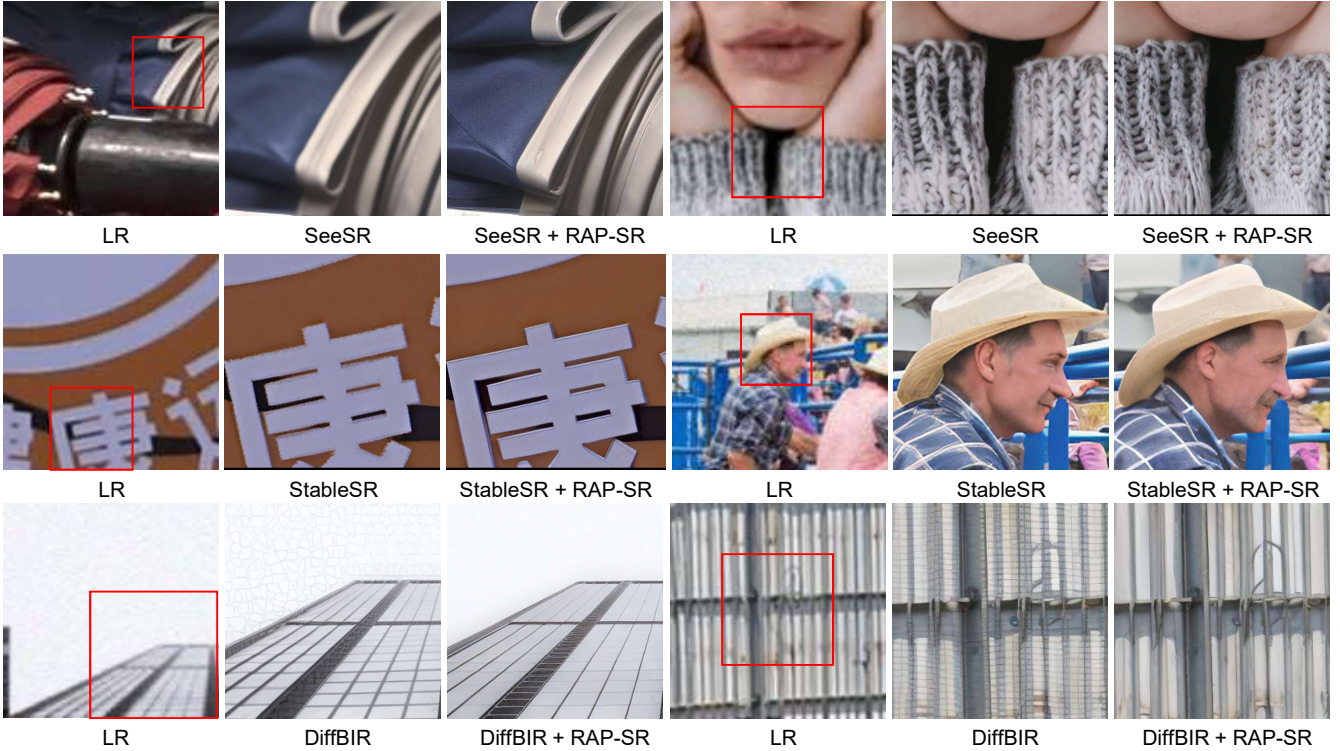
Figure 5: Qualitative comparisons on real-world test datasets. RAP-SR obtains the best visual performance.

	Configurations	MANIQA ↑	CLIPIQA ↑
Dataset size	1000	0.5681	0.639
	3000	0.5727	0.6423
	8000	0.5926	0.6866
Prompt	w/o prompt	0.5806	0.6308
	w/o negative prompt	0.6063	0.6864
	w/o positive prompt	0.6107	0.6629
Ours	Default	**0.6161**	**0.7075**

Table 2: Ablation study. Test on the DrealSR dataset.

nificantly improves perceptual quality, produces more realistic textures, and enhances the overall realism of images (e.g., sweaters and landscapes). Furthermore, our method significantly reduces issues such as blurring and artifacts (e.g., in windows and skies). In summary, RAP-SR enables diffusion-based super-resolution methods to more accurately reconstruct image details in real-world scenes. More visual results are provided in the supplementary material.

4.2 Ablation Study

Due to the superior generative capabilities of SeeSR (Wu et al. 2024), all ablation experiments for our model are conducted using the SeeSR model.

Effect of Different Dataset Sizes We conducted random split tests on the proposed HFAID dataset with varying sizes, as shown in Table 2. When the dataset size is small, there is a significant decline in no-reference metrics. When the dataset

size is expanded to 8,000 images, the model begins to converge, and both reference and no-reference metrics show a decline compared to the default 5,000-image dataset, resulting in further performance degradation.

Effect of Different Prompts We test our fine-tuned T2I model using various restoration prompts to validate their effects. The prompts included: positive prompt only, negative prompt only, and no prompt. The results are shown in Table 2. Firstly, we observe a significant drop in no-reference perceptual quality metrics when no prompts are used, underscoring the crucial role of restoration prompts. Additionally, negative prompts prove more beneficial for image generation compared to positive prompts. When we use both positive and negative prompts, all metrics achieve optimal results.

5 Conclusion

This paper introduces RAP-SR, a novel approach that enhances restoration priors in pretrained diffusion models for real-world image super-resolution (Real-SR) tasks. We develop the High-Fidelity Aesthetic Image Dataset (HFAID) through a Quality-Driven Aesthetic Image Selection Pipeline (QDAISP), surpassing existing datasets in fidelity and aesthetic quality. The Restoration Priors Enhancement Framework, including Restoration Priors Refinement (RPR) and Restoration-Oriented Prompt Optimization (ROPO), refines priors and optimizes restoration identifiers. RAP-SR seamlessly integrates into diffusion-based SR methods, significantly boosting performance. Extensive

experiments demonstrate its broad applicability and state-of-the-art results.

References

Agustsson, E.; and Timofte, R. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2017*, 1122–1131.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 18187–18197.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, 18392–18402.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward Real-World Single Image Super-Resolution: A New Benchmark and a New Model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2019*, 3086–3095.

Chen, C.; Xiong, Z.; Tian, X.; Zha, Z.; and Wu, F. 2019. Camera Lens Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 1652–1660.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wang, Z.; Kwok, J. T.; Luo, P.; Lu, H.; and Li, Z. 2024. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Chung, H.; Sim, B.; and Ye, J. C. 2022. Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 12403–12412.

Dai, X.; Hou, J.; Ma, C.-Y.; Tsai, S.; Wang, J.; Wang, R.; Zhang, P.; Vandenhende, S.; Wang, X.; Dubey, A.; et al. 2023. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2009*, 248–255.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a Deep Convolutional Network for Image Super-Resolution. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Proceedings of European Conference on Computer Vision, Part IV*, volume 8692, 184–199.

Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; and Dai, B. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, 9935–9946.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.;

Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems, NeurIPS 2014*, 2672–2680.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep Back-Projection Networks for Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 1664–1673.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 770–778.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in neural information processing systems, NeurIPS 2020*, volume 33, 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising Diffusion Restoration Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems, NeurIPS 2022*.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 5128–5137.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Deeply-Recursive Convolutional Network for Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 1637–1645.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, 3992–4003.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems, NeurIPS 2023*.

Li, W.; Zhou, K.; Qi, L.; Lu, L.; and Lu, J. 2022. Best-Buddy GANs for Highly Detailed Image Super-resolution. In *AAAI Conference on Artificial Intelligence, AAAI 2022,*, 1412–1420.

Li, Y.; Hunt, S.; Park, J.; O'Toole, M.; and Kitani, K. 2023b. Azimuth Super-Resolution for FMCW Radar in Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, 17504–17513.

Li, Y.; Zhang, K.; Liang, J.; Cao, J.; Liu, C.; Gong, R.; Zhang, Y.; Tang, H.; Liu, Y.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Gool, L. V. 2023c. LSDIR: A Large Scale Dataset for Image Restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2023*, 1775–1787.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Gool, L. V.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021*, 1833–1844.

Liang, J.; Zeng, H.; and Zhang, L. 2022. Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 5647–5656.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Proceedings of European Conference on Computer Vision, Part V*, volume 8693, 740–755.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Ouyang, W.; Qiao, Y.; and Dong, C. 2024. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. arXiv:2308.15070.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Gool, L. V. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 11451–11461.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *The Tenth International Conference on Learning Representations, ICLR 2022*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 10674–10685.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, 22500–22510.

Saharia, C.; Chan, W.; Chang, H.; Lee, C. A.; Ho, J.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Palette: Image-to-Image Diffusion Models. In Nandigjav, M.; Mitra, N. J.; and Hertzmann, A., eds., *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference*, 15:1–15:10.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022.

LAION-5B: An open large-scale dataset for training next generation image-text models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems, NeurIPS 2022*.

Shang, S.; Shan, Z.; Liu, G.; Wang, L.; Wang, X.; Zhang, Z.; and Zhang, J. 2024. ResDiff: Combining CNN and Diffusion Model for Image Super-resolution. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, 8975–8983.

Shao, W.; and Mou, X. 2021. No-Reference Image Quality Assessment Based on Edge Pattern Feature in the Spatial Domain. *IEEE Access*, 9: 133170–133184.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR 2021*.

Sun, H.; Li, W.; Liu, J.; Chen, H.; Pei, R.; Zou, X.; Yan, Y.; and Yang, Y. 2024. CoSeR: Bridging Image and Language for Cognitive Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 25868–25878.

Timofte, R.; Agustsson, E.; Gool, L. V.; Yang, M.; Zhang, L.; Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K. M.; Wang, X.; Tian, Y.; Yu, K.; Zhang, Y.; Wu, S.; Dong, C.; Lin, L.; Qiao, Y.; Loy, C. C.; Bae, W.; Yoo, J.; Han, Y.; Ye, J. C.; Choi, J.; Kim, M.; Fan, Y.; Yu, J.; Han, W.; Liu, D.; Yu, H.; Wang, Z.; Shi, H.; Wang, X.; Huang, T. S.; Chen, Y.; Zhang, K.; Zuo, W.; Tang, Z.; Luo, L.; Li, S.; Fu, M.; Cao, L.; Heng, W.; Bui, G.; Le, T.; Duan, Y.; Tao, D.; Wang, R.; Lin, X.; Pang, J.; Xu, J.; Zhao, Y.; Xu, X.; Pan, J.; Sun, D.; Zhang, Y.; Song, X.; Dai, Y.; Qin, X.; Huynh, X.; Guo, T.; Mousavi, H. S.; Vu, T. H.; Monga, V.; Cruz, C.; Egiazarian, K. O.; Katkovnik, V.; Mehta, R.; Jain, A. K.; Agarwalla, A.; Praveen, C. V. S.; Zhou, R.; Wen, H.; Zhu, C.; Xia, Z.; Wang, Z.; and Guo, Q. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2017*, 1110–1121.

Wang, J.; Chan, K. C. K.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, 2555–2563.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C. K.; and Loy, C. C. 2024. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *Int. J. Comput. Vis.*, 132(12): 5929–5949.

Wang, R.; Zhang, D.; Li, Q.; Zhou, X.; and Lo, B. 2021a. Real-time Surgical Environment Enhancement for Robot-Assisted Minimally Invasive Surgery Based on Super-Resolution. In *IEEE International Conference on Robotics and Automation, ICRA 2021*, 3434–3440.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021b. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021*, 1905–1914.

Wang, Y.; Yu, J.; and Zhang, J. 2023. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. In *The Eleventh International Conference on Learning Representations, ICLR 2023*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component Divide-and-Conquer for Real-World Image Super-Resolution. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Proceedings of European Conference on Computer Vision, Part VIII*, volume 12353 of *Lecture Notes in Computer Science*, 101–117.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024. SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 25456–25467.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2024. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 4818–4829.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2022*, 1190–1199.

Yang, T.; Wu, R.; Ren, P.; Xie, X.; and Zhang, L. 2024. Pixel-Aware Stable Diffusion for Realistic Image Super-Resolution and Personalized Stylization. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Proceedings of European Conference on Computer Vision, Part XI*, volume 15069 of *Lecture Notes in Computer Science*, 74–91.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 13040–13051.

Yu, F.; Gu, J.; Li, Z.; Hu, J.; Kong, X.; Wang, X.; He, J.; Qiao, Y.; and Dong, C. 2024. Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 25669–25680.

Yue, Z.; Wang, J.; and Loy, C. C. 2023. ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems, NeurIPS 2023*.

Zhang, K.; Liang, J.; Gool, L. V.; and Timofte, R. 2021. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. In *IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 4771–4780.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, 3813–3824.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Trans. Image Process.*, 24(8): 2579–2591.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 586–595.

Appendix

A Details of the HFAID

In this section, we provide a detailed comparison of our proposed High-fidelity Aesthetic Image Dataset (HFAID) with existing datasets, highlighting differences in visual quality, quantitative metrics, and caption accuracy.

A.1 Quantitative Comparisons

To accurately assess the effectiveness of our proposed HFAID Dataset, we conduct a comparison of no-reference metrics across different datasets, as shown in Table 3. In our selection pipeline, we use CLIPIQA (Wang, Chan, and Loy 2023), MANIQA (Yang et al. 2022), and NIQE (Zhang, Zhang, and Bovik 2015) as the primary evaluation metrics. To avoid data bias, we also employ MUSIQ (Ke et al. 2021) and BRISQUE (Shao and Mou 2021) as additional testing metrics. The results indicate that our dataset achieves the best outcomes across all metrics. Compared to the previously high-quality restoration dataset LSDIR (Li et al. 2023c), our dataset shows significant improvements across multiple no-reference metrics: The NIQE metric improves by 23%, the BRISQUE improves by 56% and the CLIPIQA improves by 11%. This demonstrates the advantages of our proposed dataset.

A.2 Qualitative Comparisons

In Figure 10, we present the visual results of our dataset. To ensure a fair comparison, we randomly sample eight images from the dataset for comparison. Compared to existing datasets, the HAFID dataset excels in both image quality and aesthetic performance. Specifically, the HAFID dataset demonstrates higher image quality and better alignment with human aesthetic preferences than the commonly used LAION-5B (Schuhmann et al. 2022) dataset in text-to-image tasks. The SAM (Kirillov et al. 2023) dataset applies blurring to all faces to protect privacy, but this significantly impacts the dataset's usefulness in model training.

	Image	
Raw Caption	Best Apps For Pancake Day Banner	Old fashioned table clock Royalty Free Stock Photo
Our Caption	There is a stack of pancakes on a white plate. There are raspberries and blueberries on top of the pancakes. There is two white bowls next to the plate with berries in them.	A clock is sitting on a wooden table. The clock has a gold chain around it. The face of the clock is white. The numbers on the clock are black. There is a blue wall behind the table.

Figure 6: Comparison of Different Image Captions. In existing text-image datasets, such as LAION, the caption quality is generally low and prone to errors. In contrast, we utilize advanced vision-language models to generate more refined captions, thereby providing richer data for model training.

	MANIQA↑	MUSIQ↑	CLIPIQA↑	BRISQUE↓	NIQE↓
DIV2K	0.6041	64.17	0.5729	15.14	3.092
Filck2K	0.6017	63.43	0.5933	29.54	3.886
COCO	0.6844	69.73	0.6696	15.40	3.891
LSIDR	0.6702	71.86	0.6675	13.08	2.950
Ours	**0.7136**	**75.69**	**0.7524**	**5.671**	**2.260**

Table 3: Quantitative comparisons across different datasets. The results indicate that our dataset achieves the best performance across all metrics.

A.3 Comparison of Image Captions

In the training process of Diffusion models, the quality of text labels is just as important as image quality. To better describe the images, we use advanced vision-language models to generate high-density captions. As shown in Figure 6, the existing LAION dataset contains numerous errors in its captions, which are often brief and fail to fully describe the image's content. In contrast, the captions we generate show significant improvements in both descriptive quality and information density, enabling our model to perform more effectively in quality control during generation.

B Details of the QDAISP

In the second phase of the Quality-Driven Aesthetic Image Selection Pipeline (QDAISP), our goal is to accurately assess image quality in alignment with human aesthetic preferences. The core of this process lies in identifying image quality assessment metrics that best match human evaluation standards. To achieve this, we conduct a detailed user study.

First, we test the LSDIR (Li et al. 2023c) dataset using commonly used no-reference metrics, such as CLIPIQA, MANIQA, MUSIQ, NIQE, and BRISQUE. We then select 200 images from both the best and worst-performing results for each metric for analysis, as shown in Figure 11. Next, we organize a group of 10 researchers to rate these metrics to identify those that most accurately reflect human aesthetic preferences. After voting, CLIPIQA, NIQE, and MANIQA are selected as the key metrics for evaluating image quality. The voting results are presented in Figure 7. It is important to note that we do not claim BRISQUE and MUSIQ are use-

Figure 7: User Study Results. The voting results of this study are based on feedback from 10 volunteers. The image quality assessment metrics that best align with human quality preferences are identified by evaluating the performance of the images and the metrics.

Metric	LAION	Flick2K	SAM	DIV2K	LSDIR	HFAID
MANIQA	0.6434	0.6247	0.6275	0.6475	0.6333	**0.6512**
CLIPIQA	0.6258	0.6683	0.6841	0.6806	0.6923	**0.7217**

Table 4: Effectiveness of HFAID. Our proposed HFAID significantly improves the results.

less in image quality assessment; rather, we emphasize that metrics like CLIPIQA more closely align with human aesthetic preferences.

As shown in Figure 11, images with better metric performance in CLIPIQA and MANIQA exhibit richer details, brighter visuals, prominent subjects, and a sense of aesthetic appeal. In contrast, images with worse metric performance display noticeable color casts and poorer image quality. NIQE effectively identifies synthetic images, particularly in worse metric performance. Furthermore, we find that BRISQUE and MUSIQ correlate less with human aesthetic preferences.

C Ablation Study

This section provides additional ablation experiments on High-fidelity Aesthetic Image Dataset (HFAID) and Restoration-Oriented Prompt Optimization (ROPO).

C.1 Effectiveness of HFAID

We train RAP-SR on various datasets to demonstrate the superiority of our proposed HFAID. The datasets used include DIV2K (Agustsson and Timofte 2017), LAION-5B, SAM, Flick2K (Timofte et al. 2017), LSDIR, and our HFAID dataset. Each dataset is trained for the same number of epochs to ensure fairness. We use SeeSR as the base model, train RAP-SR with different datasets, and test it on the RealSR dataset. As shown in Table 4, the HFAID dataset significantly improves the model's performance across multiple metrics compared to other datasets. This underscores the significant advantage of our dataset in enhancing result quality.

Sematic Prompt | Sematic Prompt + SeeSR Negative Prompt | Sematic Prompt + Restoration Identifier

Figure 8: Comparison of the Effectiveness of Our ROPO Method on T2I Tasks. When we apply different restoration prompts (such as SeeSR's negative prompts: "dotted, noise, blur, lowres, smooth") and our method (Restoration Negative Identifier), ROPO generates more realistic degradation effects than SeeSR prompts. Consequently, this allows the diffusion model to produce more realistic results when utilizing CFG.

C.2 Effectiveness of Restoration-Oriented Prompt Optimization

To validate the effectiveness of our proposed Restoration-Oriented Prompt Optimization (ROPO) strategy, we conduct tests on text-to-image tasks using default semantic prompts, SeeSR's negative prompts, and our Restoration Identifier. As shown in Figure 8, SeeSR's negative prompts produce simple degradation effects like grayscale images, while our Restoration Identifier generates more realistic degradations. These effects are often hard to describe precisely with language, but our optimized approach enables the Diffusion model to correlate prompts with realistic degradations strongly.

During the inference phase, by using the Classifiers-Free Guidance (CFG) (Ho and Salimans 2022) strategy, which combines positive and negative prompts, the model generates more realistic results. Positive prompts guide the model to produce images that match the target description, while negative prompts help steer the output away from undesired features. By optimizing the Restoration Identifier, the model

(a) Result-1   (b) Result-2

Metric	Result-1	Result-2
PSNR	26.69	24.93
SSIM	0.7434	0.6903
LPIPS	0.4342	0.3434
MANIQA	0.2584	0.6594
MUSIQ	45.87	72.35
CLIPIQA	0.6298	0.8452
BRISQUE	58.52	14.12

(c) Quantitative comparison

Figure 9: Comparison of No-Reference and Full-Reference Metrics. In two different results (Result-1 and Result-2), although Result-1 performs better on full-reference metrics such as PSNR and SSIM, it fails to deliver highly realistic outcomes compared to Result-2. Therefore, in Real-SR tasks, no-reference metrics are more valuable than full-reference metrics.

effectively avoids real degradation features, leading to generate more realistic images.

D Misalignment Between Human Perception and Image Quality Assessment Metrics

In Figure 9, we present an illustrative example that highlights the differences between image quality assessment metrics and human perception. We compare two different results and evaluate them using multiple metrics. Quantitative evaluation shows that Result-2 scores lower than Result-1 on full-reference metrics, such as PSNR and SSIM, but higher on no-reference metrics, such as CLIPIQA and MANIQA. However, visual assessment reveals that Result-2 produces a more realistic effect than Result-1, effectively reducing smoothness and blurring. This suggests that no-reference image quality assessment metrics, such as MANIQA, MUSIQ, and CLIPIQA, align more closely with human visual perception trends. This further underscores the importance of no-reference image quality assessment metrics in real-world super-resolution tasks.

E Additional Visual Results from RAP-SR

In Figure 12, we present additional visual results to demonstrate the superior applicability and performance of RAP-SR.

HFAID (Ours)

LAION-5B

SAM

Figure 10: Qualitative Comparisons on different datasets. Our HFAID dataset is compared with the LAION-5B and SAM datasets. The results demonstrate that the HFAID dataset excels in image quality and aesthetic performance. In contrast, the LAION-5B dataset shows lower image quality, while the SAM dataset has even poorer quality, with facial features of individuals intentionally obscured for privacy reasons.

	Better Metric Performance	Worse Metric Performance
CLIPIPQA		
MANIQA		
NIQE		
MUSIQ		
BRISQUE		

Figure 11: Image quality assessment and its corresponding images used in the user study. The study results indicate that CLIPIQA, MANIQA, and NIQE more accurately reflect human aesthetic preferences and excel at distinguishing between high-quality and low-quality images. In contrast, metrics like MUSIQ and BRISQUE demonstrate poor separability between different image qualities and diverge from human aesthetic preferences.
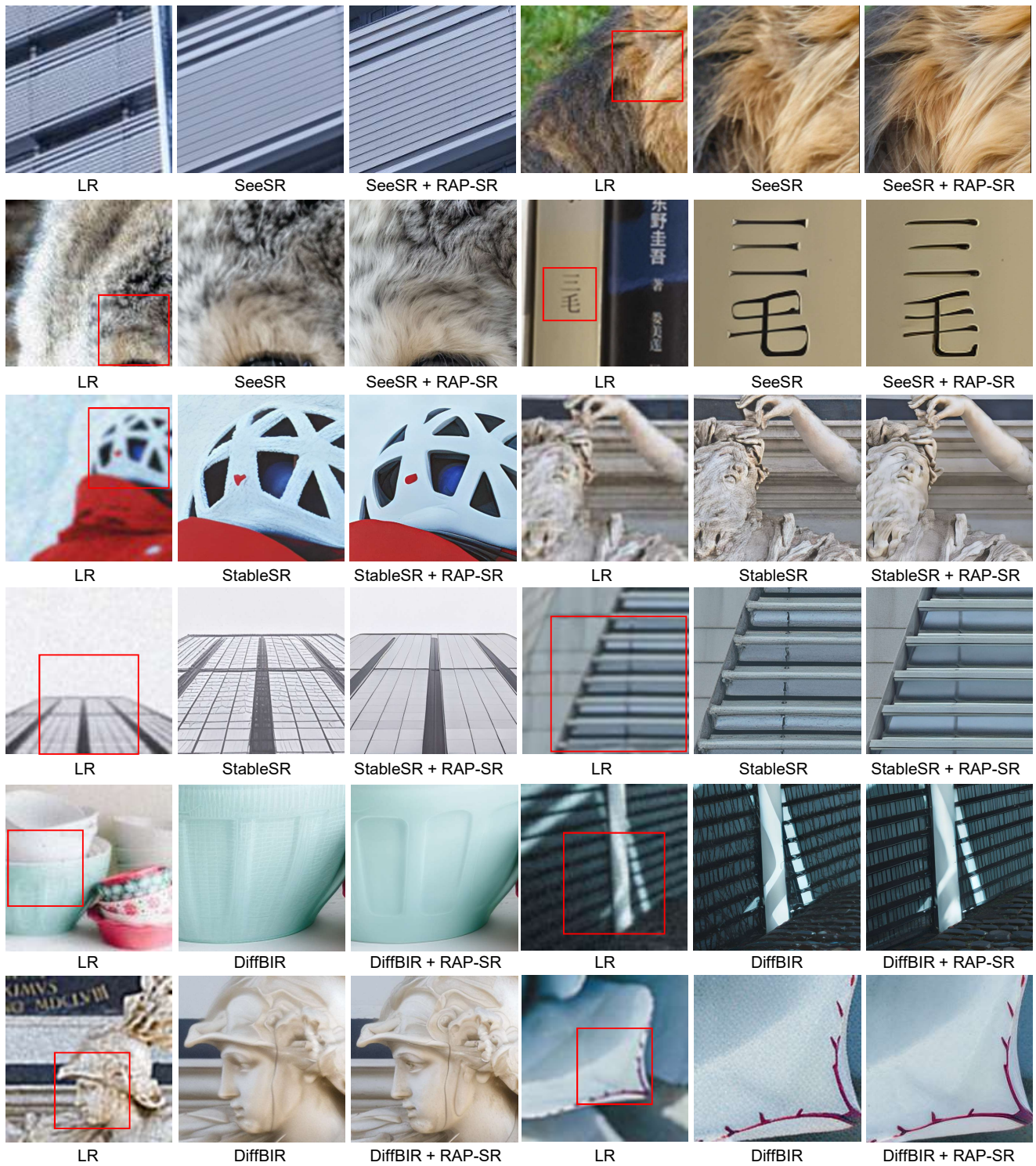
Figure 12: Qualitative comparisons on different test datasets. RAP-SR obtains the best visual performance. Please zoom in for a detailed view.