

SPIRE: Semantic Prompt-Driven Image Restoration

Chenyang Qi^{1,2*}, Zhengzhong Tu¹, Keren Ye¹, Mauricio Delbracio¹,
Peyman Milanfar¹, Qifeng Chen², and Hossein Talebi¹

¹ Google Research

² HKUST

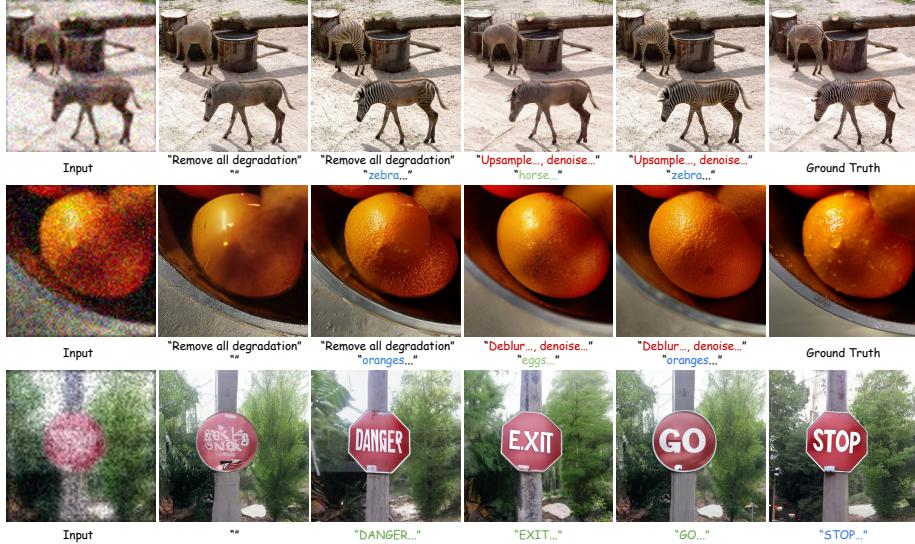


Fig. 1: We present **SPIRE: Semantic Prompt-Driven Image Restoration**, a text-based foundation model for all-in-one, instructed image restoration. SPIRE allows users to flexibly leverage either semantic-level content prompt, or quantitative degradation-aware restoration prompt, or both, to obtain their desired enhancement results based on personal preferences. In other words, SPIRE can be easily prompted to conduct blind restoration, semantic restoration, or task-specific granular treatment. Our framework also enables a new paradigm of **instruction-based image restoration**, providing a reliable evaluation benchmark to facilitate vision-language models for low-level computational photography applications.

Abstract. Text-driven diffusion models have become increasingly popular for various image editing tasks, including inpainting, stylization, and object replacement. However, it still remains an open research problem to adopt this language-vision paradigm for more fine-level image processing tasks, such as denoising, super-resolution, deblurring, and compression artifact removal. In this paper, we develop SPIRE, a Semantic and restoration Prompt-driven Image Restoration framework that leverages

*This work was done during an internship at Google Research.

natural language as a user-friendly interface to control the image restoration process. We consider the capacity of prompt information in two dimensions. First, we use content-related prompts to enhance the semantic alignment, effectively alleviating identity ambiguity in the restoration outcomes. Second, our approach is the first framework that supports fine-level instruction through language-based quantitative specification of the restoration strength, without the need for explicit task-specific design. In addition, we introduce a novel fusion mechanism that augments the existing ControlNet architecture by learning to rescale the generative prior, thereby achieving better restoration fidelity. Our extensive experiments demonstrate the superior restoration performance of SPIRE compared to the state of the arts, alongside offering the flexibility of text-based control over the restoration effects.

基於語言提供的「定量規範」來調整復原強度

1 Introduction

Image restoration or enhancement aims to recover high-quality, pixel-level details from a degraded image, while preserving as much the original semantic information as possible. Although neural network models [11, 12, 30, 63, 67, 71, 73, 74] have marked significant progress, it still remains challenging to design an effective task-conditioning mechanism instead of building multiple individual models for each task (such as denoise, deblur, compression artifact removal) in practical scenarios. The advancement of text-driven diffusion models [48, 50, 53] has unveiled the potential of natural language as a universal input condition for a broad range of image processing challenges, which improves interactivity with users and reduces the cost of task-specific fine-tuning. However, the existing applications of natural language in stylization [38, 44], and inpainting [3, 4, 10, 21, 44, 64] predominantly focus on high-level semantic editing, whereas the uniqueness and challenges for low-level image processing have been less explored.

Natural language text prompts in image restoration can play two crucial roles: alleviating semantic ambiguities and resolving degradation type ambiguities. Firstly, the same degraded image can correspond to different visual objects, leading to ambiguities in content interpretation (*e.g.*, discerning whether the blurred animal in Fig. 1 is a horse or a zebra). Typically, unconditional image-to-image restoration leads to a random or average estimation of the identities, resulting in neither fish nor fowl. Secondly, certain photographic effects that are deliberately introduced for aesthetic purposes, such as bokeh with a soft out-of-focus background, can be misconstrued as blur distortions by many existing deblurring models, leading to unwanted artifacts or severe hallucinations in the outputs. Although blind restoration methods can produce clean images by either leveraging frozen generative priors [67, 70, 77], or using end-to-end regression [71], they do not consider the aforementioned ambiguities.

In this paper, we introduce **SPIRE**—a Semantic Prompt-Driven Image REstoration framework that provides a user-friendly interface to fully control both the restored image semantics and the enhancement granularity using natural language instructions. Traditional blind restoration without prompts (Fig-

ure 1) has limited flexibility in solving semantic and degradation ambiguity, thus tends to generate average blurry images. Leveraging the semantic and quantitative restoration prompts, we show that the visual quality can be significantly improved. Moreover, our framework can also be used as blind restoration with null prompts. In addition to higher perceptual quality, our framework provides the flexibility for users to generate more than one plausible result. Specifically, in imaging scenarios with large levels of degradation there are multiple plausible solutions. The proposed interactive framework has the capability to personalize the output by leveraging input prompts according to user preference. In concurrent studies, the focus has typically been isolated to one of three areas: either blind restoration [34, 67], only semantic prompting [78], or only discrete restoration types [9, 36]. To the best of our knowledge, this is the first unified model to support the following three distinct features simultaneously:

1. **Blind Restoration**: When instructed with a general restoration prompt “remove all degradations” and empty semantic prompt “”, SPIRE operates as a blind restoration model (“Ours w/o text” in Tab. 1).
2. **Semantic Restoration**: When provided with a text description of the (desired) visual content, SPIRE concentrates on restoring the specified identity of the uncertain or ambiguous objects in the degraded image.
3. **Quantitative Task-Specific Restoration**: Receiving specific restoration type hints (e.g., “deblur...”, “deblur... denoise...”), SPIRE transforms into a task-specific model (e.g., deblur, denoise, or both). Moreover, it understands the numeric nuances of different degradation parameters in language prompts so that the users can control the restoration strength (e.g., “deblur with sigma 3.0” in Fig. 7).

To train SPIRE, we first build a synthetic data generation pipeline on a large-scale text-image dataset [8] upon the second-order degradation process proposed in Real-ESRGAN [71]. Additionally, we embed the degradation parameters into the restoration prompts to encode finer-grained degradation information (e.g., “deblur with sigma 3.5”). Our approach provides richer degradation-specific information compared to contemporary works [26, 36] which only employ the degradation types (e.g., “gaussian blur”). We then finetune a ControlNet adaptor [84]—which learns a parameter-efficient, parallel branch on top of the latent diffusion models (LDMs) [50]—on a mixture of diverse restoration tasks (see Fig. 1). The semantic text prompts are processed through the LDM backbone, aligning with LDM’s text-to-image pretraining. The degradation text prompts and image-conditioning are implemented in the ControlNet branch, as these aspects are not inherently addressed by the vanilla LDM. We further improve the ControlNet by introducing a new modulation connection that adaptively fuses the degradation condition with the generative prior with only a few extra trainable parameters yet showing impressive performance gain.

Our extensive experiments demonstrate remarkable restoration quality achieved by our proposed SPIRE method. Moreover, SPIRE offers additional freedom for both content and restoration prompt engineering. For example, the semantic

prompt “*a very large giraffe eating leaves*” helps to resolve the semantic ambiguity in Fig. 2; the degradation prompt “deblur with sigma 3.0” reduces the gaussian blur while maintaining the intentional motion blur in Fig. 3. We also find out that our model learns a continuous latent space of the restoration strength in Fig. 7 even though we do not explicitly fine-tune the CLIP embedding. Our contribution can be summarized as follows:

- We introduce the **first unified text-driven image restoration model that supports both semantic prompts and restoration instructions**. Our experiments demonstrate that incorporating semantic prompts and restoration instructions significantly enhances the restoration quality.
- Our proposed paradigm **empowers users to fully control the semantic outcome of the restored image using different semantic prompts** during test time.
- Our proposed approach provides a mechanism for users to **adjust the category and strength of the restoration effect** based on their subjective preferences.
- We demonstrate that text can serve as universal guidance control for low-level image restoration, eliminating the need for task-specific model design.

2 Related Work

First, we review the literature on image restoration. Our proposed framework aims to address some of the limitations of the existing restoration methods. Next, we compare multiple text-guided diffusion editing methods. Our method is designed to deal with two different text ambiguities concurrently, which is an unprecedented challenge for existing text-driven diffusion methods. Finally, we discuss the parameter-efficient fine-tuning of diffusion models, which further motivates our proposed modulation connection.

Image Restoration is the task of recovering a high-resolution, clean image from a degraded input. Pioneering works in the fields of super-resolution [14, 32], motion and defocus deblurring [1, 61, 79], denoising [82], and JPEG and artifact removal [16, 45] primarily utilize deep neural network architectures to address specific tasks. Later, Transformer-based [30, 63, 73] and adversarial-based [71] formulations were explored, demonstrating state-of-the-art performance with unified model architecture. Recently there has been a focus on exploiting iterative restorations, such as the ones from diffusion models, to generate images with much higher perceptual quality [11, 49, 52, 54, 74, 75, 78].

There have been some recent works attempting to control the enhancement strength of a single model, such as the scaling factor condition in arbitrary super-resolution [68, 69] and noise-level map for image denoising [83]. However, these methods require hand-crafting dedicated architectures for each task, which limits their scalability. In contrast to the task-specific models, other approaches seek to train a single blind model by building a pipeline composed of several classical degradation process, such as BSRGAN [80] and Real-ESRGAN [71].

Text-guided Diffusion Editing. Denoising diffusion models are trained to reverse the noising process [23, 57, 58]. Early methods focused on unconditioned generation [13], but recent trends pivot more to conditioned generation, such as

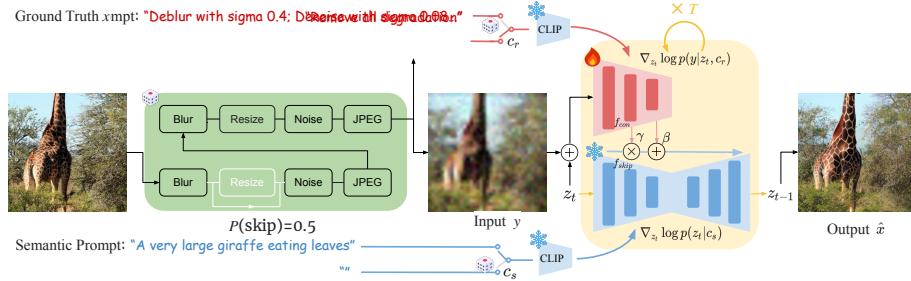


Fig. 2: Framework of SPIRE. In the training phase, we begin by synthesizing a degraded version y , of a clean image x . Our degradation synthesis pipeline also creates a **restoration prompt** c_r , which contains **numeric parameters that reflects the intensity of the degradation introduced**. Then, we inject the synthetic restoration prompt into a **ControlNet adaptor**, which uses our proposed **modulation fusion blocks** (γ, β) to **connect with the frozen backbone** driven by the semantic prompt c_s . During **test time**, the **users can employ** the SPIRE framework as either a blind restoration model with restoration prompt “*Remove all degradation*” and empty semantic prompt \emptyset , or manually adjust the restoration c_r and semantic prompts c_s to obtain what they want.

image-to-image translation [52, 55] and text-to-image generation [50, 53]. Latent Diffusion [50] is a groundbreaking approach that introduces a versatile framework for improving both training and sampling efficiency, while flexible enough for general conditioning inputs. Subsequent works have built upon their pretrained text-to-image checkpoints, and designed customized architectures for different tasks like text-guided image editing. For instance, SDEdit [37] generates content for a new prompt by adding noise to the input image. Attention-based methods [21, 44, 64, 81] show that images can be edited via reweighting and replacing the cross-attention map of different prompts. Aside from the previously mentioned approaches driven by target prompts, instruction-based editing [7, 17] entail modifying a source image based on specific instructions. These textual instructions are typically synthesized using large language models [41, 47]. As CLIP serves as a cornerstone for bridging vision and language, several studies have extensively investigated the representation space of CLIP embeddings [46]. Instead of directly applying CLIP to fundamental discriminative tasks [19, 85], these works either tailor CLIP to specific applications, such as image quality assessment [28, 31], serve as semantic guidance in generative adversarial network [5] or aligning image embedding with degradation types in the feature space [26, 36]. To enhance CLIP’s ability to understand numerical information, some works [42, 43] finetune it using contrastive learning on synthetic numerical data, and then incorporate the fine-tuned text embedding into diffusion training.

Parameter-Efficient Diffusion Model Finetuning. To leverage the powerful generative prior in pretrained diffusion models, parameter-efficient components such as text embedding [15], low-rank approximations of model weights [20, 25], and cross attention layers [29] can be finetuned to personalize the pretrained model. Adaptor-based finetuning paradigms [39, 84] propose to keep the original

UNet weights frozen, and add new image or text conditions. This adaptor generates residual features that are subsequently added to the frozen UNet backbone.

3 Method

We introduce a universal approach to combine the above mentioned task-specific, strength-aware, and blind restoration methods within a unified framework (illustrated in Sec. 3.2). We further propose to decouple the learning of content and restoration prompts to better preserve the pre-trained prior while injecting new conditions, as unfolded in Sec. 3.3. Sec. 3.4 details our design on how to accurately control the restoration type and strength, as well as our proposed modulation fusion layer that adaptively fuses the restoration features back to the frozen backbone. We start with some preliminaries in the following section.

3.1 Preliminaries

Latent Diffusion Models (LDMs) [50] are probabilistic generative models that learn the underlying data distribution by iteratively removing Gaussian noise in the latent space, which is typically learned using a VAE autoencoder. Formally, the VAE encoder \mathcal{E} compresses an input image \mathbf{x} into a compact latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$, which can be later decoded back to the pixel space using the coupled VAE decoder \mathcal{D} , often learned under an image reconstruction objective: $\mathcal{D}(\mathcal{E}(\mathbf{x})) \approx \mathbf{x}$. During training stage, the output of a UNet [51] $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{y})$ conditioned on \mathbf{y} (such as text, images, etc.) is parameterized [13, 56] to remove Gaussian noise ϵ in the latent space \mathbf{z}_t as:

$$\min_{\theta} \mathbb{E}_{(\mathbf{z}_0, \mathbf{y}) \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(0, I), t} \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{y})\|_2^2, \quad (1)$$

where \mathbf{z}_t is a noisy sample of \mathbf{z}_0 at sampled timestep t . The condition \mathbf{y} is randomly dropped out as \emptyset to make the model unconditional.

At test time, deterministic DDIM sampling [57] is utilized to convert a random noise \mathbf{z}_T to a clean latent \mathbf{z}_0 which is decoded as final result $\mathcal{D}(\mathbf{z}_0)$. In each timestep t , classifier-free guidance [13, 24] can be applied to trade-off sample quality and condition alignment:

$$\bar{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{y}) = \epsilon_\theta(\mathbf{z}_t, t, \emptyset) + w(\epsilon_\theta(\mathbf{z}_t, t, \mathbf{y}) - \epsilon_\theta(\mathbf{z}_t, t, \emptyset)),$$

where w is a scalar to adjust the guidance strength of \mathbf{y} . Note that the estimated noise $\bar{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{y})$ is used to update \mathbf{z}_t to \mathbf{z}_{t-1} , which is an approximation of the distribution gradient score [60] as $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t | \mathbf{y}) \propto \bar{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{y})$.

3.2 Text-driven Image Restoration

Based on the LDM framework, we propose a new restoration paradigm—text-driven image restoration. Our method target to restore images (\mathbf{x} or \mathbf{z}_0) based

on conditions $\{\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r\}$. Specifically: \mathbf{y} denotes the degraded image condition, \mathbf{c}_s is the semantic prompt describing the clean image \mathbf{x} (e.g., “a panda is sitting by the bamboo” or “a panda”), and \mathbf{c}_r is the restoration prompt that describes the details of the degradation in terms of both the operation and parameter (e.g., “deblur with sigma 3.0”). We use $\mathbf{y} = \text{Deg}(\mathbf{x}, \mathbf{c}_r)$ to denote the degradation process which turns the clean image \mathbf{x} into its degraded counterpart \mathbf{y} .

The above text-driven image restoration model $p(\mathbf{z}_t | \{\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r\})$ can be trained using paired data. We use the text-image dataset [8] so that each clean image \mathbf{x} is paired with a semantic prompt \mathbf{c}_s (the paired alt-text). Then $\mathbf{y} = \text{Deg}(\mathbf{x}, \mathbf{c}_r)$ is simulated using the synthetic degradation [71] pipeline (Sec. 4.1), yielding the final paired training data $(\mathbf{x} \text{ or } \mathbf{z}_0, \{\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r\}) \sim p_{\text{data}}$.

Comparison to Blind Restoration. Recent existing blind restoration models [34, 55, 67] also leverage diffusion priors to generate high-quality images. However, these methods are prone to hallucinate unwanted, unnatural, or over-sharpened textures given that the semantic- and degradation-ambiguities persist everywhere. Fig. 3 provides an example, where the input image is degraded simultaneously with Gaussian blur and motion blur. The fully blind restoration method removes all of the degradation, leaving no way to steer the model to remove only the Gaussian blur while preserving the motion effect for aesthetic purposes. Therefore, having the flexibility to control the restoration behavior and its strength at test time is a crucial requirement. Our text-driven formulation of the problem introduces additional controlling ability using both the semantic content and the restoration instruction. In our SPIRE design, $\mathbf{c}_r = \text{"Remove all degradation"}$ is also a plausible restoration prompt which makes our SPIRE to be compatible with the blind restoration. Further, our text-driven restoration takes advantages of the pretrained content-aware LDM using the additional semantic prompt \mathbf{c}_s , hence able to better handle semantic ambiguities on the noisy, blurry regions whose visual objects are less recognizable.

3.3 Decoupling Semantic and Restoration Prompts

To effectively learn the latent distribution $p(\mathbf{z}_t | \mathbf{y}, \mathbf{c}_s, \mathbf{c}_r)$, we further decouple the conditions $\{\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r\}$ into two groups—one for the text-to-image prior ($\mathbf{c}_s \rightarrow \mathbf{z}_t$) already imbued in the pretrained LDM model, and the other for both the image-to-image ($\mathbf{y} \rightarrow \mathbf{z}_t$) and restoration-to-image ($\mathbf{c}_r \rightarrow \mathbf{z}_t$) that needs to be learnt from the synthetic data. This decoupling strategy prevents catastrophic forgetting in the pretrained diffusion model and enables independent training of the restoration-aware model, whose score function [60] is expressed as:

$$\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t | \mathbf{y}, \mathbf{c}_s, \mathbf{c}_r) \approx \underbrace{\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t | \mathbf{c}_s)}_{\text{Semantic-aware (frozen)}} + \underbrace{\nabla_{\mathbf{z}_t} \log p(\mathbf{y} | \mathbf{z}_t, \mathbf{c}_r)}_{\text{Restoration-aware (learnable)}}. \quad (2)$$

In the above equation, the first part $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t | \mathbf{c}_s)$ aligns with the text-to-image prior inherent in the pretrained LDM. The second term $\nabla_{\mathbf{z}_t} \log p(\mathbf{y} | \mathbf{z}_t, \mathbf{c}_r)$ approximates the consistency constraint with degraded image \mathbf{y} , meaning the latent image \mathbf{z}_t should be inferred by the degraded image \mathbf{y} and the degradation

details \mathbf{c}_r . This is somewhat similar to the reverse process of $\mathbf{y} = \text{Deg}(\mathbf{x}_t, \mathbf{c}_r)$ in the pixel space. More derivation is provided in our supplement. Providing degradation information through a restoration text prompt \mathbf{c}_r can largely reduce the uncertainty of estimation $p(\mathbf{y}|\mathbf{z}_t, \mathbf{c}_r)$, thereby alleviating the ambiguity of the restoration task and leading to better results that are less prone to hallucinations.

3.4 Learning to Control the Restoration.

Inspired by ControlNet [84], we employ an adaptor to model the restoration-aware branch in Eq. (2). First, the pretrained latent diffusion UNet $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}_s)$ is **frozen** during training to preserve **text-to-image prior** $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{c}_s)$. Secondly, an **encoder initialized from pretrained UNet** is finetuned to learn the restoration condition $\nabla_{\mathbf{z}_t} \log p(\mathbf{y}|\mathbf{z}_t, \mathbf{c}_r)$, which takes degradation image \mathbf{y} , current noisy latents \mathbf{z}_t and the **restoration prompt** \mathbf{c}_r as input. The output of encoder is **fusion features** \mathbf{f}_{con} for frozen UNet. In this way, the choice of ControlNet implementation follows our decoupling analysis in Eq. (2).

Conditioning on the Degraded Image \mathbf{y} . We apply downsample layers on \mathbf{y} to make the feature map matching to the shape of the \mathbf{z}_t , denoted as $\mathcal{E}'(\mathbf{y})$. \mathcal{E}' is more compact [84] than VAE encoder \mathcal{E} , and its trainable parameters allow model to perceive and adapt to the degraded images from our synthetic data pipeline. Then, we concatenate the downsampled feature map $\mathcal{E}'(\mathbf{y})$ with the latent \mathbf{z}_t , and feed them to the ControlNet encoder.

Conditioning on the Restoration Prompt \mathbf{c}_r . Following [50], we first use the vision-language CLIP [46] to infer the text sequence embedding $\mathbf{e}_r \in \mathbb{R}^{M \times d}$ for \mathbf{c}_r , where M is the number of tokens, and $d = 768$ is the dimension of the image-text features. Specifically, given a restoration instruction “*deblur with sigma 3.0*”, CLIP first tokenizes it into a sequence of tokens (e.g., “3.0” is tokenized into [“3”, “.”, “0”]) and embeds them into M distinct token embeddings. Then, stacked causal **transformer layers** [65] are applied on the embedding sequence to contextualize the token embeddings, resulting in \mathbf{e}_r . Hence, tokens such as “*deblur*” may arm with the numeric information such as “3”, “.”, and “0”, making it possible for the model to distinguish the semantic difference of strength definition in different restoration tasks. Cross-attention process ensures that the information from \mathbf{e}_r is propagated to the ControlNet feature \mathbf{f}_{con} . While [42, 43] finetuned CLIP using contrastive learning, while we found frozen CLIP still works. The reason is that the learnable cross-attention layers can somehow ensemble the observations from frozen CLIP and squeeze useful information from it.

Modulation Fusion Layer. The decoupling of **learning two branches** (Sec. 3.3), despite offering the **benefit of effective separation** of prior and conditioning,



Fig. 3: Degradation ambiguities in real-world problems. By adjusting the restoration prompt, our method can preserve the motion effect that is coupled with Gaussian blur, while fully blind restoration models do not provide this flexibility.

Method	Prompts	Parameterized Degradation with synthesized c_r								Real-ESRGAN Degradation without c_r					
		Sem	Res	FID \downarrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	CLIP-I \uparrow	CLIP-T \uparrow	FID \downarrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	CLIP-I \uparrow	CLIP-T \uparrow
SwinIR [30]		X	X	43.22	0.423	24.40	0.717	0.856	0.285	48.37	0.449	23.45	0.699	0.842	0.284
StableSR [67]		X	X	20.55	0.313	21.03	0.613	0.886	0.298	25.75	0.364	20.42	0.581	0.864	0.298
DiffBIR [34]		X	X	17.26	0.302	22.16	0.604	0.912	0.297	19.17	0.330	21.48	0.587	0.898	0.298
ControlNet-SR [84]		X	X	13.65	0.222	23.75	0.669	0.938	0.300	16.99	0.269	22.95	0.628	0.924	0.299
Ours w/o text		X	X	12.70	0.221	23.84	0.671	0.939	0.299	16.25	0.262	23.15	0.636	0.929	0.300
DiffBIR [34] + SDEdit [37]		✓	X	19.36	0.362	19.39	0.527	0.891	0.305	17.51	0.375	19.15	0.521	0.887	0.308
DiffBIR [34] + CLIP [46]		✓	X	18.46	0.365	20.50	0.526	0.896	0.308	20.31	0.374	20.45	0.539	0.885	0.307
ControlNet-SR + CLIP [46]		✓	X	13.00	0.241	23.18	0.648	0.937	0.307	15.16	0.286	22.45	0.610	0.926	0.308
Ours		✓	✓	11.34	0.219	23.61	0.665	0.943	0.306	14.42	0.262	23.14	0.633	0.935	0.308

Table 1: Quantitative results on the MS-COCO dataset (with c_s) using our parameterized degradation (left) and Real-ESRGAN degradation (right). We also denote the prompt choice at test time. ‘Sem’ stands for semantic prompt; ‘Res’ stands for restoration prompt. The first group of baselines are tested without prompt. The second group are combined with semantic prompt in zero-shot way.

Degradation Process	$p(\text{choose})$	Parameterized				Degradation Process	$p(\text{choose})$	Real-ESRGAN	
		Restoration	Prompt	Restoration	Prompt			Restoration	Prompt
Gaussian Blur	0.5	Deblur with $\{\sigma \in [0.2, 3.0]\}$ or Deblur	Blur				1.0		
Downsample	0.5	Upsample to $\{\text{resizing factor} \in [1.0, 7.0]\}$ or Upsample	Resize				1.0		
Gaussian Noise	0.5	Denoise with $\{\sigma \in [0.0, 0.12]\}$ or Denoise	Noise				1.0		Remove all degradation
JPEG	0.5	Dejepg with quality $\{\text{quality factor} \in [30, 92]\}$ or Dejepg	JPEG				1.0		

Table 2: Our training degradation is randomly sampled in these two pipeline with 50% each. (1) Images generated by our **parameterized pipeline** are paired with either a restoration type (*e.g.*, “*Deblur*”) or a restoration parameter prompt (*e.g.*, “*Deblur with sigma 0.3;*”). (2) In other 50% iterations, degraded images y synthesized by Real-ESRGAN are paired with the same restoration prompt c_r = “*Remove all degradation*”

may cause **distribution shift** in the learned representations between the frozen backbone feature f_{skip} and ControlNet feature f_{control} . To allow for adaptive alignment of the above features, here we propose a **new modulation fusion layer** to **fuse the multi-scale features from ControlNet to the skip connection features** of the frozen backbone: $\hat{f}_{\text{skip}} = (1 + \gamma)f_{\text{skip}} + \beta$; $\gamma, \beta = \mathcal{M}(f_{\text{con}})$, where γ, β are the **scaling and bias parameters** from \mathcal{M} a lightweight 1×1 zero-initialized convolution layer.

4 Experiments

4.1 Text-based Training Data and Benchmarks

Our method opens up an entirely new paradigm of instruction-based image restoration. However, existing image restoration datasets such as DIV2K [2] and Flickr2K [40] do not provide high-quality semantic prompts, and Real-ESRGAN degradation is for blind restoration. To address this, we construct a new setting including training data generation and test benchmarks for evaluation.

Our parameterized degradation pipeline is based on the Real-ESRGAN [71] which contains a large number of degradation types. Since the original full Real-ESRGAN degradation is difficult to be parameterized and represented as user-friendly natural language (*e.g.*, not every one understands “*anisotropic plateau blur with beta 2.0*”), we **choose the 4 most general degradations** which are practical for parameterization to support degradation parameter-aware restoration, as shown in Tab. 2. Our parameterized pipeline skips each degradation stage

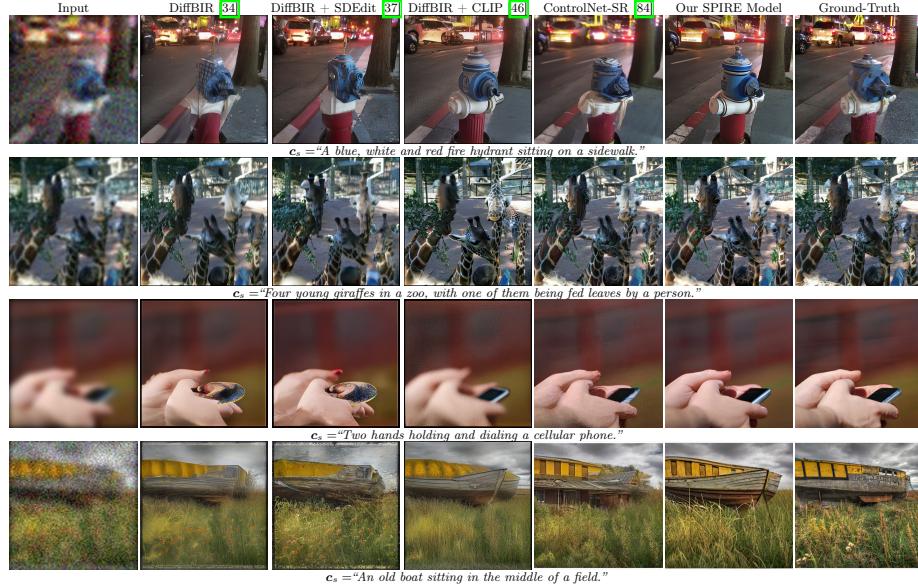


Fig. 4: Visual Comparison with other baselines. Our method of integrating both the semantic prompt c_s and the restoration prompt c_r outperforms image-to-image restoration (DiffBIR, Retrained ControlNet-SR) and naive zero-shot combination with semantic prompt. It achieves more sharp and clean results while maintaining consistency with the degraded image.

with a 50% probability to increase the diversity of restoration prompts and support single task restoration. The descriptions of the selected degradations are appended following the order of the image degradations to synthesize c_r . In real application, users can control the restoration type and its strength by modifying restoration prompt. As shown in third column of Fig. 3, driven by restoration prompt "Deblur with sigma 3.0", our framework removes the Gaussian blur, while preserving motion blur. The representation of the restoration strength is also general enough to decouple different tasks in natural language as shown in Fig. 7. More visual results are presented in our supplement.

Training data construction. In the training stage, we sample near 100 million text-image pairs from internal data source and use the alternative text label with highest relevance as the semantic prompt c_s in the framework. We drop out the semantic prompt to empty string \emptyset by a probability of 10% to support semantic-agnostic restoration. We mix the Real-ESRGAN degradation pipeline and our parameterized pipeline by 50% each in training iterations.

Evaluation Setting. We randomly choose 3000 pairs of image x and semantic prompt c_s from the MS-COCO [33] validation set. Then, we build two test sets on parameterized and the Real-ESRGAN processes separately. At the left half of Tab. 1, images are sent to the parameterized degradation to form the synthesize degradation image y with restoration prompt c_r . In this setting, we expect "Ours" model to fully utilize the c_s describing the semantic and the c_r

Method	FID↓	LPIPS↓	PSNR↑	SSIM↑	CLIP-I↑
Real-ESRGAN [72]	32.37	0.312	22.52	0.646	0.683
DiffBIR [34] (zero-shot)	30.71	0.354	22.01	0.526	0.921
StableSR	24.44	0.311	21.62	0.533	0.928
Ours w/o text (zero-shot)	28.80	0.352	21.68	0.549	0.927
Ours w/o text (finetuned)	22.45	0.321	21.38	0.532	0.932

Table 3: Numerical results on the DIV2K testset without any prompt.

describing the degradations. On the right half, the same images are fed to the Real-ESRGAN degradation—so the degradation is aligned with open-sourced checkpoints [30, 34, 67]. Because not every model can utilize both the semantic (c_s) and restoration (c_r) information, Tab. 1 provides a “Prompts” column denoting what information the model relied on. In summary, our constructed benchmarks based on MS-COCO can serve multiple purposes—from fully-blind, content-prompted, to all-prompted restoration.

Evaluation Metrics. We use the generative metric FID [22] and perceptual metric LPIPS [86] for quantitative evaluation of image quality. PSNR, and SSIM are reported for reference. We also evaluate similarity scores in the clip embedding space with ground truth image (CLIP-I) and caption (CLIP-T).

4.2 Comparing with baselines

We compare our method with three categories of baselines:

- Open-sourced image restoration checkpoints, including the regression model **SwinIR** [30] and the latent diffusion models **StableSR** [67] and **DiffBIR** [34].
- Combining the state-of-the-art method [34] with the zero-shot post-editing [37] (DiffBIR+SDEdit) or zero-shot injection through CLIP [46] (DiffBIR + CLIP).
- ControlNet-SR (retrained): We adapt the image-to-image ControlNet model [84] for super-resolution. For consistency, we maintain the same training iterations as our own model. We also evaluate its zero-shot capabilities with CLIP guidance (ControlNet-SR + CLIP).

Quantitative comparison with baselines is presented in Tab. 1. On the left, we evaluate our full model with the parameterized degradation test set. Since open-sourced baselines are pretrained on Real-ESRGAN degradation, not perfectly aligned with our parameterized degradation, we also evaluate our full model on MS-COCO with Real-ESRGAN degradation by setting our restoration prompt to “*Remove all degradation*”. Thanks to our prompts guidance and architecture improvements, our full model achieves best FID, LPIPS, CLIP-Image score, which means better image quality and semantic restoration. “Ours w/o text” is the same checkpoint conditioned on only degradation image and has high pixel-level similarity (the second best PSNR and SSIM) with GT. Although SwinIR has highest PSNR, its visual results are blurry. Although combining semantic prompt in zero-shot way can bring marginal improvement in FID and CLIP similarity with caption, it deteriorates the image restoration capacity and results in worse LPIPS and CLIP-Image. In contrast, semantic prompt guidance improves the CLIP image similarity of our full model. Tab. 3 shows the evaluation results of our model on DIV2K test set provided by StableSR. In zero-shot

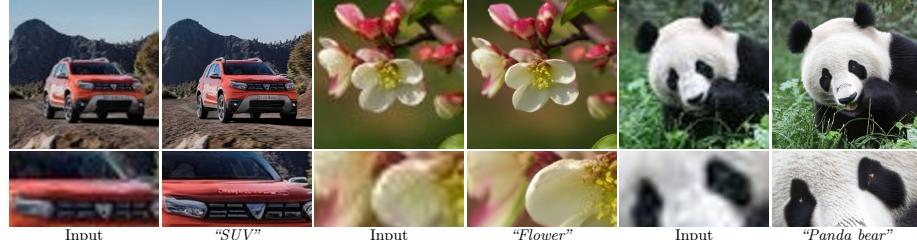


Fig. 5: Real-world image restorations. Our framework enhances real-world low-quality images following semantic prompts provided by language models or user input.

Metric	SwinIR [30]	Diffbir [34]	Ablation of restoration prompt c_r				Ablation of semantic prompt c_s		
	X	X	X	short	unmatched	full			
Sem c_s	X	X							
Res c_r	X	X	"Deblur 0.6"	"... 1.0"	"... 2.0"	"... 5.0"	"Remove all degradation"		
CLIP-IQA \uparrow	0.6190	0.6983	0.4074	0.4159	0.4831	0.4797	0.5818	0.5750	0.6460
MUSIQ \uparrow	63.65	69.69	35.82	39.87	49.98	48.81	61.68	60.27	67.54
MANIQA \uparrow	0.4035	0.2619	0.2299	0.2487	0.2970	0.2821	0.3694	0.3637	0.4355
									0.4474

Table 4: Quantitative results on real-world images. Our aesthetic quality is improved when increasing restoration strength or using a more accurate and detailed prompt.

test setting, our model has lower FID and LPIPS than DiffBIR. To fairly compare with StableSR, we finetune our model on their training set. Our finetuned model also outperforms StableSR in FID and CLIP-I.

Qualitative comparison with baselines is presented in Fig. 4, where the corresponding semantic prompt is provided below each row. Image-to-image baselines such as DiffBir and the retrained ControlNet-SR can easily generate artifacts (*e.g.*, hydrant in the first row) and blurry images (*e.g.*, hands example in third row) that may look close to an “average” estimation. Besides, naive combination with semantic prompt in the zero-shot approach results in heavy hallucinations and may fail to preserve the object identity in the input image (*e.g.*, giraffes on the second row). Unlike the existing methods, our full model considers semantic prompt, degradation image and restoration prompt in both training and test stages, which makes its results more aligned with all conditions.

4.3 Real-world restoration

Besides evaluation on synthetic degradations, we present an analysis on real-world images in Tab. 4 and Fig. 5. We choose RealPhoto60 [78] as our real-world test-set, and use non-reference metrics CLIP-IQA [66], MUSIQ [27] and MANIQA [76] scores to study the influence of the semantic and restoration prompts. First, we fix the semantic prompt to the empty string \emptyset and conduct an ablation study on the restoration prompt. As the *deblurring* strength increases from *0.6* to *2.0*, the image quality is improved consistently. A blind prompt “*Remove all degradation*” can provide better results than the manually controlled restoration prompts. This showcases that the proposed model is flexible to work with generic restoration prompts, especially when coming up with the perfect prompt is challenging. We also fix the blind restoration prompt and change the semantic prompts. Following [78], we feed degraded images to LLaVA [35] and generate corresponding semantic prompts. Our experiments show that detailed **full** descriptions (about 60 words) of the degraded image from



Fig. 6: Test-time semantic prompting. Our framework restores degraded images guided by flexible semantic prompts, while unrelated background elements and global tones remain aligned with the degraded input conditioning.

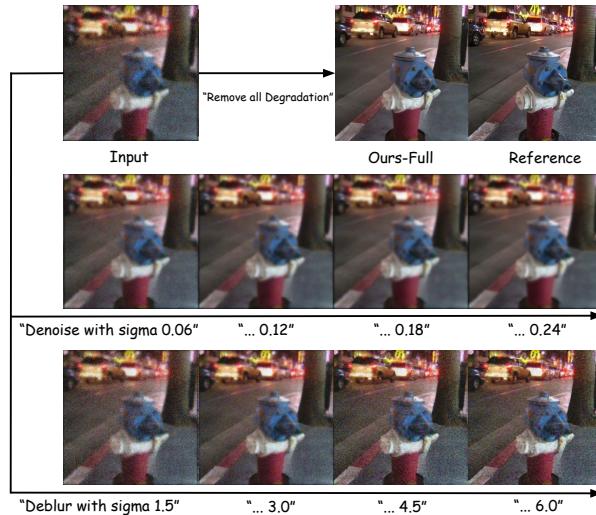


Fig. 7: Prompt space walking visualization for the restoration prompt. Given the same degraded input (upper left) and empty semantic prompt \emptyset , our method can decouple the restoration direction and strength via only prompting the **quantitative number in natural language**. An interesting finding is that our model learns a continuous range of restoration strength from discrete language tokens.

LLaVA can be more effective to improve the visual results than short prompts that have less than 20 words. Unmatched prompts simulate the case when a user provides wrong prompts by randomly re-sampling unmatched long prompts from the data set. An interesting observation is that unmatched long prompts also improve the aesthetic quality of images against a short prompt or a blind restoration. Nonetheless, it can lead to more “hallucination” when the semantic input prompt is unmatched with the real ground truth. In contrast to concurrent work [62] that leverages implicit clip embedding, our text-based formulation provides better control of hallucinations by adjusting the LLM-synthesized or manually-designed prompt at test-time. Note that many studies [6, 18] find current image restoration metrics are biased towards specific datasets and not fully aligned with human’s visual perception. The quantitative results of previous works [34, 71] are provided for reference. More results and discussions are provided in the supplements.

Method	Sem c_s	Res c_r	FID↓	LPIPS↓	PSNR↑	SSIM↑	CLIP-I↑	CLIP-T↑
ControlNet-SR	✗	✗	13.65	0.222	23.75	0.669	0.938	0.300
ControlNet-SR	✓	✓	12.14	0.223	23.50	0.660	0.940	0.306
Ours	✓	wo param	11.58	0.223	23.61	0.665	0.942	0.305
Ours	✓	✓	11.34	0.219	23.61	0.665	0.943	0.305

Table 5: Ablation of architecture and degradation strength in c_r

4.4 Prompting the SPIRE

Semantic Prompting. Our model can be driven by user-defined semantic prompts. As shown in Fig. 6, the blind restoration with empty string generates ambiguous and unrealistic identity. In contrast, our framework reconstructs sharp and realistic results. The object identity accurately follows user’s semantic prompt, while the global layout and color tone remain consistent with input.

Restoration Prompting. Users also have the freedom to adjust the degradation types and strengths in the restoration prompts. As shown in Fig. 7, the denoising and deblurring restoration are well decoupled. Using the task-specific prompts, our method can get a clean but blur (second row), or sharp but noisy (third row) result. In addition, our model learns the continuous space of restoration as the result becomes cleaner progressively if we modify the “Denoise with sigma 0.06” to “0.24.”. Beyond purely relying on numerical description, our model “understand” the difference of strength definition for tasks, such as the range difference of denoise sigma and deblur sigma, which makes it promising to be a universal representation for restoration strength control.

4.5 Ablation Study

We conduct several ablations in Tab. 5. Providing both prompts c_s , c_r to “ControlNet-SR” baseline improves the FID by 1.5. Adding the modulation fusion layer (“Ours”) further improves the generative quality (0.8 for FID) and pixel-level similarity (0.11 dB). Embedding degradation parameters in the restoration prompt, not only enables restoration strength prompting (Fig. 7), but also improves image quality (FID and LPIPS). More details are in the supplement.

5 Conclusion

We have presented a unified text-driven framework for instructed image restoration using semantic and restoration prompts. We design our model in a decoupled way to better preserve the semantic text-to-image generative prior while efficiently learning to control both the restoration direction and its strength. To the best of our knowledge, this is the first framework to support both semantic and parameter-embedded restoration instructions simultaneously, allowing users to flexibly prompt-tune the results up to their expectations. Our extensive experiments have shown that our method significantly outperforms prior works in terms of both quantitative and qualitative results. Our model and evaluation benchmark have established a new paradigm of instruction-based image restoration, paving the way for further multi-modality generative imaging applications.

References

1. Abuolaim, A., Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Learning to reduce defocus blur by realistically modeling dual-pixel data. In: ICCV. pp. 2289–2298 (2021)
2. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 1122–1131. IEEE Computer Society (2017). <https://doi.org/10.1109/CVPRW.2017.150>, <https://doi.org/10.1109/CVPRW.2017.150>
3. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (2023)
4. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR. pp. 18208–18218 (2022)
5. Bai, Y., Wang, C., Xie, S., Dong, C., Yuan, C., Wang, Z.: Textir: A simple framework for text-based editable image restoration. arXiv preprint arXiv:2302.14736 (2023)
6. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR. pp. 6228–6237 (2018)
7. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
8. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S.A., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: Pali: A jointly-scaled multilingual language-image model. In: ICLR (2023), <https://arxiv.org/abs/2209.06794>
9. Chen, Z., Zhang, Y., Gu, J., Yuan, X., Kong, L., Chen, G., Yang, X.: Image super-resolution with text prompt diffusion. arXiv preprint arXiv:2303.06373 (2023)
10. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: ICLR (2022)
11. Delbracio, M., Milanfar, P.: Inversion by direct iteration: An alternative to denoising diffusion for image restoration. Transactions on Machine Learning Research (2023), <https://openreview.net/forum?id=VmyFF51L3F>, featured Certification
12. Delbracio, M., Talebi, H., Milanfar, P.: Projected distribution loss for image enhancement. In: 2021 IEEE International Conference on Computational Photography (ICCP). pp. 1–12. IEEE (2021)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Neural Information Processing Systems (2021)
14. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016)
15. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023)
16. Galteri, L., Seidenari, L., Bertini, M., Bimbo, A.: Deep generative adversarial compression artifact removal. In: ICCV (2017)
17. Geng, Z., Yang, B., Hang, T., Li, C., Gu, S., Zhang, T., Bao, J., Zhang, Z., Hu, H., Chen, D., Guo, B.: Instructdiffusion: A generalist modeling interface for vision tasks. CoRR **abs/2309.03895** (2023). <https://doi.org/10.48550/arXiv.2309.03895>, <https://doi.org/10.48550/arXiv.2309.03895>

18. Gu, J., Cai, H., Dong, C., Ren, J.S., Timofte, R., Gong, Y., Lao, S., Shi, S., Wang, J., Yang, S., Wu, T., Xia, W., Yang, Y., Cao, M., Heng, C., Fu, L., Zhang, R., Zhang, Y., Wang, H., Song, H., Wang, J., Fan, H., Hou, X., Sun, M., Li, M., Zhao, K., Yuan, K., Kong, Z., Wu, M., Zheng, C., Conde, M.V., Burchi, M., Feng, L., Zhang, T., Li, Y., Xu, J., Wang, H., Liao, Y., Li, J., Xu, K., Sun, T., Xiong, Y., Keshari, A., Komal, Thakur, S., Jakhetiya, V., Subudhi, B.N., Yang, H.H., Chang, H.E., Huang, Z.K., Chen, W.T., Kuo, S.Y., Dutta, S., Das, S.D., Shah, N.A., Tiwari, A.K.: Ntire 2022 challenge on perceptual image quality assessment. In: CVPRW. pp. 951–967 (June 2022)
19. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2021)
20. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. In: ICCV. pp. 7323–7334 (October 2023)
21. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: ICLR (2023), https://openreview.net/forum?id=_CDixzkzeyb
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
24. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), <https://openreview.net/forum?id=qw8AKxfYbI>
25. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022), <https://openreview.net/forum?id=nZeVkeeFYf9>
26. Jiang, Y., Zhang, Z., Xue, T., Gu, J.: Autodir: Automatic all-in-one image restoration with latent diffusion. arXiv preprint arXiv:2310.10123 (2023)
27. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
28. Ke, J., Ye, K., Yu, J., Wu, Y., Milanfar, P., Yang, F.: Vila: Learning image aesthetics from user comments with vision-language pretraining. In: CVPR. pp. 10041–10051 (2023)
29. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR. pp. 1931–1941 (2023)
30. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of ICCV Workshops (2021)
31. Liang, Z., Li, C., Zhou, S., Feng, R., Loy, C.C.: Iterative prompt learning for unsupervised backlit image enhancement. In: ICCV. pp. 8094–8103 (2023)
32. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of CVPR Workshops (2017)
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, L.: Microsoft coco: Common objects in context. In: ECCV. ECCV (September 2014), <https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/>
34. Lin, X., He, J., Chen, Z., Lyu, Z., Fei, B., Dai, B., Ouyang, W., Qiao, Y., Dong, C.: Diffbir: Towards blind image restoration with generative diffusion prior. arXiv preprint arXiv:2308.15070 (2023)

35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
36. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for multi-task image restoration. In: The Twelfth International Conference on Learning Representations (2024), <https://openreview.net/forum?id=t3vnnLeajU>
37. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2022)
38. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR. pp. 6038–6047 (June 2023)
39. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
40. Nah, S., Timofte, R., Baik, S., Hong, S., Moon, G., Son, S., Lee, K.M., Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C., Fan, Y., Yu, J., Liu, D., Huang, T.S., Sim, H., Kim, M., Park, D., Kim, J., Chun, S.Y., Haris, M., Shakhnarovich, G., Ukita, N., Zamir, S.W., Arora, A., Khan, S.H., Khan, F.S., Shao, L., Gupta, R.K., Chudasama, V.M., Patel, H., Upla, K.P., Fan, H., Li, G., Zhang, Y., Li, X., Zhang, W., He, Q., Purohit, K., Rajagopalan, A.N., Kim, J., Tofighi, M., Guo, T., Monga, V.: NTIRE 2019 challenge on video deblurring: Methods and results. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16–20, 2019. pp. 1974–1984. Computer Vision Foundation / IEEE (2019). <https://doi.org/10.1109/CVPRW.2019.00249>, http://openaccess.thecvf.com/content_CVPRW_2019/html/NTIRE_Nah_NTIRE_2019_Challenge_on_Video_Deblurring_Methods_and_Results_CVPRW_2019_paper.html
41. OpenAI: Gpt-4 technical report (2023)
42. Paiss, R., Chefer, H., Wolf, L.: No token left behind: Explainability-aided image classification and generation. In: ECCV (2022)
43. Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., Dekel, T.: Teaching clip to count to ten. In: ICCV (2023)
44. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
45. Prakash, M., Delbracio, M., Milanfar, P., Jug, F.: Interpretable unsupervised diversity denoising and artefact removal. In: ICLR (2022), <https://openreview.net/forum?id=DfMq1BOPXjM>
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
47. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog 1(8), 9 (2019)
48. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
49. Ren, M., Delbracio, M., Talebi, H., Gerig, G., Milanfar, P.: Multiscale structure guided diffusion for image deblurring. In: ICCV. pp. 10721–10733 (2023)
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)

51. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
52. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
53. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)
54. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
55. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
56. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: ICLR. OpenReview.net (2022), <https://openreview.net/forum?id=TIdIXIpzhoI>
57. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
58. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS. pp. 11895–11907 (2019)
59. Song, Y., Shen, L., Xing, L., Ermon, S.: Solving inverse problems in medical imaging with score-based generative models. In: ICLR. OpenReview.net (2022)
60. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021), <https://openreview.net/forum?id=PxDIG12RRHS>
61. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: CVPR. pp. 1279–1288 (2017)
62. Sun, H., Li, W., Liu, J., Chen, H., Pei, R., Zou, X., Yan, Y., Yang, Y.: Coser: Bridging image and language for cognitive super-resolution. arXiv preprint arXiv:2311.16512 (2023)
63. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A.C., Li, Y.: MAXIM: multi-axis MLP for image processing. In: CVPR (2022)
64. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR. pp. 1921–1930 (2023)
65. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
66. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)
67. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015 (2023)
68. Wang, L., Wang, Y., Lin, Z., Yang, J., An, W., Guo, Y.: Learning a single network for scale-arbitrary super-resolution. In: ICCV. pp. 4801–4810 (2021)
69. Wang, X., Chen, X., Ni, B., Wang, H., Tong, Z., Liu, Y.: Deep arbitrary-scale image super-resolution via scale-equivariance pursuit. In: CVPR (2023)
70. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: CVPR (2021)
71. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCV. pp. 1905–1914 (2021)

72. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of ECCV Workshops (2018)
73. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: CVPR (2022)
74. Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A.G., Milanfar, P.: Deblurring via stochastic refinement. In: CVPR. pp. 16293–16303 (2022)
75. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: CVPR (2024)
76. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniq: Multi-dimension attention network for no-reference image quality assessment. In: CVPR. pp. 1191–1200 (2022)
77. Yang, T., Ren, P., Xie, X., Zhang, L.: Gan prior embedded network for blind face restoration in the wild. In: CVPR (2021)
78. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: CVPR (2024)
79. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021)
80. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV. pp. 4791–4800 (2021)
81. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36** (2024)
82. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing **26**(7), 3142–3155 (2017)
83. Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for CNN based image denoising. IEEE Transactions on Image Processing (2018)
84. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
85. Zhang, R., Fang, R., Gao, P., Zhang, W., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. In: ECCV (2022)
86. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

A Implementation Details

A.1 Derivation Details in Equation 2

We propose to compute the distribution of latents $p(\mathbf{z}_t|\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r)$ conditioned on degraded image \mathbf{y} , semantic prompt \mathbf{c}_s and restoration prompt \mathbf{c}_r . Using the Bayes' decomposition similar to score-based inverse problem [58, 59], we have

$$p(\mathbf{z}_t|\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r) = p(\mathbf{z}_t, \mathbf{y}, \mathbf{c}_s, \mathbf{c}_r)/p(\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r). \quad (3)$$

Then, we compute gradients with respect to \mathbf{z}_t , and remove the gradients of input condition $\nabla_{\mathbf{z}_t} \log p(\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r) = 0$ as:

$$\begin{aligned} & \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r) \\ &= \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t, \mathbf{y}, \mathbf{c}_s, \mathbf{c}_r) \end{aligned} \quad (4)$$

$$= \nabla_{\mathbf{z}_t} \log[p(\mathbf{c}_s) \cdot p(\mathbf{z}_t|\mathbf{c}_s) \cdot p(\mathbf{y}, \mathbf{c}_r|\mathbf{c}_s, \mathbf{z}_t)] \quad (5)$$

$$= \nabla_{\mathbf{z}_t} \log[p(\mathbf{z}_t|\mathbf{c}_s) \cdot p(\mathbf{y}, \mathbf{c}_r|\mathbf{c}_s, \mathbf{z}_t)] \quad (6)$$

$$= \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{c}_s) + \nabla_{\mathbf{z}_t} \log p(\mathbf{y}, \mathbf{c}_r|\mathbf{c}_s, \mathbf{z}_t). \quad (7)$$

We assume \mathbf{y} is generated through a degradation pipeline as $\mathbf{y} = \text{Deg}(\mathbf{x}, \mathbf{c}_r)$, thus it is independent of \mathbf{c}_s with \mathbf{x} and \mathbf{c}_r provided as condition. Removing redundant \mathbf{c}_s condition, the second term in the last equation can be approximated as:

$$\begin{aligned} & \nabla_{\mathbf{z}_t} \log p(\mathbf{y}, \mathbf{c}_r|\mathbf{c}_s, \mathbf{z}_t) \\ & \approx \nabla_{\mathbf{z}_t} \log p(\mathbf{y}, \mathbf{c}_r|\mathbf{z}_t) \end{aligned} \quad (8)$$

$$= \nabla_{\mathbf{z}_t} \log p(\mathbf{c}_r|\mathbf{z}_t) + \nabla_{\mathbf{z}_t} \log p(\mathbf{y}|\mathbf{z}_t, \mathbf{c}_r) \quad (9)$$

$$= \nabla_{\mathbf{z}_t} \log p(\mathbf{y}|\mathbf{z}_t, \mathbf{c}_r) \quad (10)$$

In summary of the above equations, we derive the Equation 2 in the main manuscript

$$\begin{aligned} & \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{y}, \mathbf{c}_s, \mathbf{c}_r) \\ & \approx \underbrace{\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{c}_s)}_{\text{Semantic-aware (frozen)}} + \underbrace{\nabla_{\mathbf{z}_t} \log p(\mathbf{y}|\mathbf{z}_t, \mathbf{c}_r)}_{\text{Restoration-aware (learnable)}}, \end{aligned} \quad (11)$$

where $\nabla_{\mathbf{z}_t} \log p(\mathbf{y}|\mathbf{z}_t, \mathbf{c}_r)$ is synthesized using stochastic degradation pipeline $\mathbf{y} = \text{Deg}(\mathbf{x}, \mathbf{c}_r)$ to train our ControlNet.

A.2 Pseudo Code for Degradation Synthesis

To support the learning of restoration-aware term $\nabla_{\mathbf{z}_t} \log p(\mathbf{y}|\mathbf{z}_t, \mathbf{c}_r)$, we synthesize the degradation image \mathbf{y} using clean image \mathbf{x} with the algorithm presented in Algorithm 1. First, we randomly choose one from Real-ESRGAN pipeline and our parameterized degradation. Then the degraded image from Real-ESRGAN pipeline is paired with restoration prompt $\mathbf{c}_r = \text{"Remove all degradation"}$. In our parameterized degradation, all processes are paired with restoration prompts \mathbf{c}_r listed in Table 2 of the main manuscript (e.g., *Deblur with sigma 3.0*).

Algorithm 1 SPIRE Degradation Pipeline in Training

Inputs: x : Clean image
Outputs: y : Degraded image; c_r : Restoration prompt

```

type ← RANDCHOICE(Real-ESRGAN, Param)
if type = Real-ESRGAN then // Real-ESRGAN degradation
     $y \leftarrow x$ 
    Deg ← RANDOM(Real-ESRGAN-Degradation)
    for PROCESS in Deg do:
         $y \leftarrow \text{PROCESS}(y)$ 
    end for
     $c_r \leftarrow \text{"Remove all degradation"}$ 
else// Parameterized degradation
     $c_r \leftarrow \emptyset$ 
     $y \leftarrow x$ 
    Deg ← RANDOM(Parameterized-Degradation)
    for PROCESS,  $c_{rp}$  in Deg do:
         $y \leftarrow \text{PROCESS}(y, c_{rp})$ 
         $c_r \leftarrow \text{CONCAT}(c_r, c_{rp})$ 
    end for
end if
return  $y, c_r$ 
```



Fig. 8: More semantic prompting for images with multiple objects.

B More Ablation Study

Tab. 6 provides more comprehensive ablations of text prompts by providing different information to our image-to-image baseline. Semantic prompts significantly improve image quality as shown in better FID and CLIP-Image, but reduce the similarity with ground truth image. Restoration types and parameters embedded in the restoration prompts both improve image quality and fidelity. Tab. 7 presents a comparison of our skip feature modulation f_{skip} with that in StableSR [67] which modulates both skip feature f_{skip} from encoder and upsampling feature f_{up} from decoder. We observe that modulating f_{up} does not bring obvious improvements. One possible reason is that γ and β of the middle layer adapts to the feature in the upsampling layers.

C Multiple Objects Semantic Prompting

Besides single semantic restoration, real applications may involve multiple objects with different semantic categories (*e.g.* Fig. 8). In each column, we guide

Method	Sem	Res Type	Res Param	FID↓	LPIPS↓	PSNR↑	SSIM↑	CLIP _{im} ↑	CLIP _{tx} ↑
Ours	✗	✗	✗	13.60	0.221	23.65	0.664	0.939	0.300
Ours	✓	✗	✗	11.71	0.226	23.55	0.663	0.941	0.305
Ours	✓	✓	✗	11.58	0.223	23.61	0.665	0.942	0.305
Ours	✓	✓	✓	11.34	0.219	23.61	0.665	0.943	0.306

Table 6: Ablation of prompts provided during both training and testing. We use an image-to-image model with our modulation fusion layer as our baseline. Providing semantic prompts significantly increases the image quality (1.9 lower FID) and semantic similarity (0.002 CLIP-Image), but results in worse pixel-level similarity. In contrast, degradation type information embedded in restoration prompts improves both pixel-level fidelity and image quality. Utilizing degradation parameters in the restoration instructions further improves these metrics.

Method	Modulate f_{skip}	Modulate f_{up}	Relative Param	FID↓	LPIPS↓
Ours w/ prompts	✗	✗	1	12.14	0.223
Ours w/ prompts	✓	✓	1.06	11.21	0.219
Ours w/ prompts	✓	✗	1.03	11.34	0.219

Table 7: Ablation of the architecture. Modulating the skip feature f_{skip} improves the fidelity of the restored image with 3% extra parameters in the adaptor, while further modulating the backbone features f_{up} does not bring obvious advantage.

the upper part of the image with *peppers*, *bananas* or *leaves*, while the lower part can be restored as *potatoes* or *stones*. Thanks to the cross attention mechanism, multiple semantics can be spatially decoupled and recombined following the user’s prompts, thus yielding better restoration for both objects.

D More Restoration Prompting

Fig. 9 shows the application of restoration prompt on images with different degradations and content, including Midjourney image and real-world cartoon. Since these images are not in our training data domain, a blind enhancement with prompt “*Remove all degradation*” can not achieve satisfying results. Utilizing restoration prompting (*e.g.*, “*Upsample to 6.0x; Deblur with sigma 2.9;*”) can successfully guide our model to improve the details and color tones of the Midjourney image. In the right half, a manually designed restoration prompt also reduces the jagged effect to smooth the lines in the cartoon image.

To study whether the model follows restoration instructions, a dense walking of restoration prompt is presented in Fig. 10. From left to right, we increase the strength of denoising in the restoration prompt. From top to the bottom, the strength of deblurring gets larger. The results demonstrate that our restoration framework refines the degraded image continuously following the restoration prompts



Fig. 9: Restoration prompting for images from internet.

E Real-world images

Although our model is trained on synthetic degradation, it generalizes to real-world data RealPhoto60 [78], as shown in Fig. 11. Compared to a model without semantic prompt, the synthetic semantic prompts from LLAVA [35] enhance fine-level details in Fig. 13 (e.g., grass under sheep in the upper left figure, and the staircase in the mountain in lower right photo). These results demonstrate an additional potential advantage of employing language prompts in real-world restoration: the ease of leveraging the logical reasoning capabilities in pre-trained large language models.

F Limitation

Although our framework can generate high-fidelity results following semantic and restoration prompts, it is prone to occasional hallucinations. As shown in Fig. 12, the image quality is degraded and the semantic is unclear when the input prompt ("Snow leopard") is misaligned with the ground truth ("Panda bear"). In-



Fig. 12: Hallucinations when the prompt is unmatched with input image restorations.

stead of relying on user input or frozen language models, one future direction can be fine-tuning multimodal language models to automatically provide more accurate instructions, thus reducing hallucinations. In addition, we plan to scale up our model parameters and extend it to more diverse and realistic degradation types in our future work.

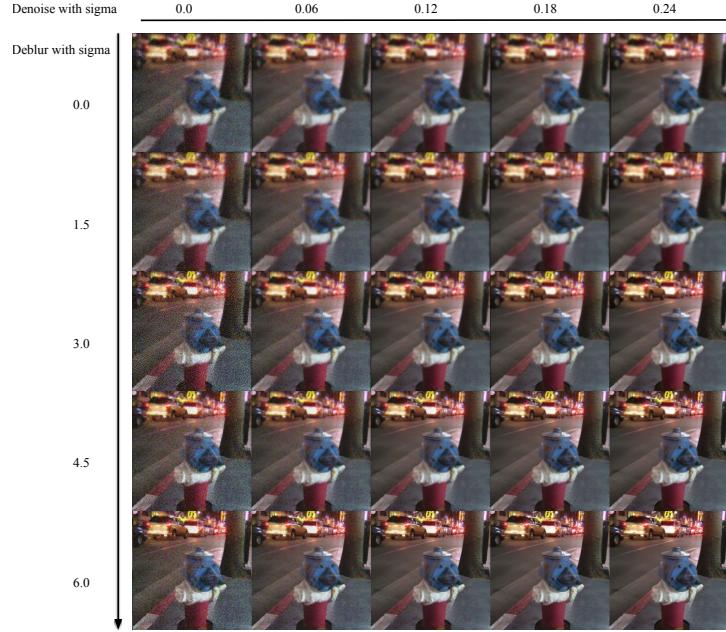


Fig. 10: Prompt space walking visualization for the restoration prompt. Given the same degraded input (upper left) and empty semantic prompt \emptyset , our method can decouple the restoration direction and strength via only prompting the **quantitative number in natural language**. An interesting finding is that our model learns a continuous range of restoration strength from discrete language tokens.



Fig. 14: Image restoration for unseen degradations.

G Mixed and universal degradation.

Our method can also restore mixed degradation (Figure 1 and Table 1 in the paper). For unseen degradations such as haze or rain, our pretrained model can still handle them properly, as shown in the figure below, since our pretrained prior and training data contains those concepts (*e.g.*, "A clear sky").

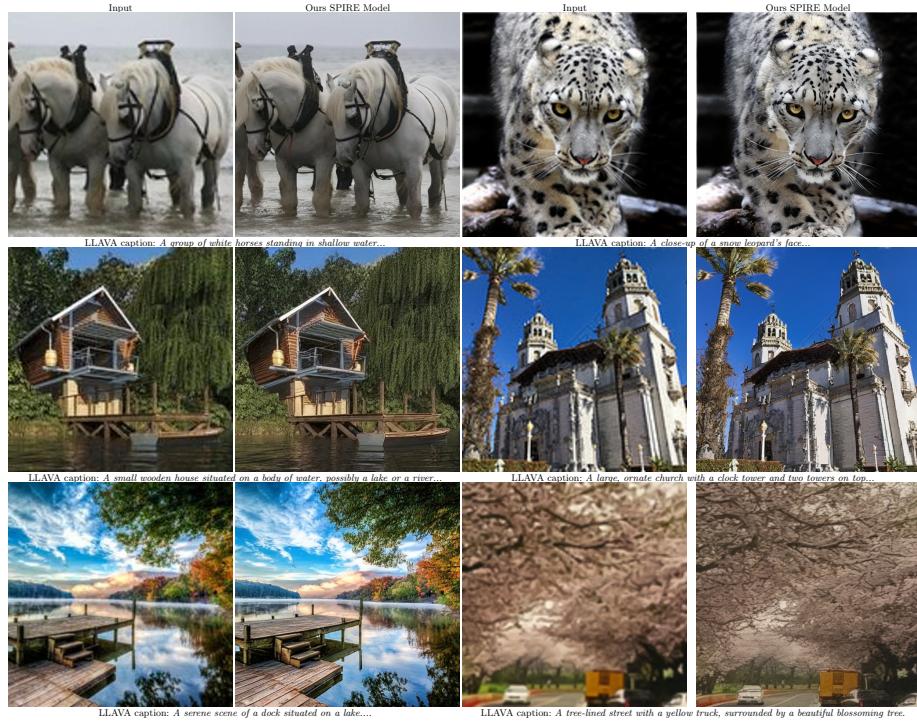


Fig. 11: Qualitative result on real-world images.

Model / Task	$\times 4$ super-resolution	denoising	de-jpeg
SwinIR (task-specific)	0.309	0.361	0.319
Ours	0.265	0.305	0.214

Table 8: Comparison with task-specific SwinIR.

H Comparison of inference cost

Our method takes 1.4s (50 DDIM steps) to run on TPUv4 and 1130 GFLOPS per step). Our UNet model has 1240 M (275 M trainable) parameters. The overall computation cost is comparable with 1203 M StableSR (19.3s on GPU) and 1510 M DiffBir (based on Stable Diffusion, 10.9s on GPU), and less than SUPIR (based on 2.6B SD-XL).

I Comparison with more models

We follow the test set design of task-specific SwinIR. In the table 8, our method outperforms the task-specific SwinIR and achieves lower LPIPS in evaluation. Following StableSR and Real-esrgan, our model is trained on large-scale open-domain images with ESRGAN degradations, which has a noticeable difference with the degradation in all-in-one restoration mentioned by reviewers (*e.g.*, DA-

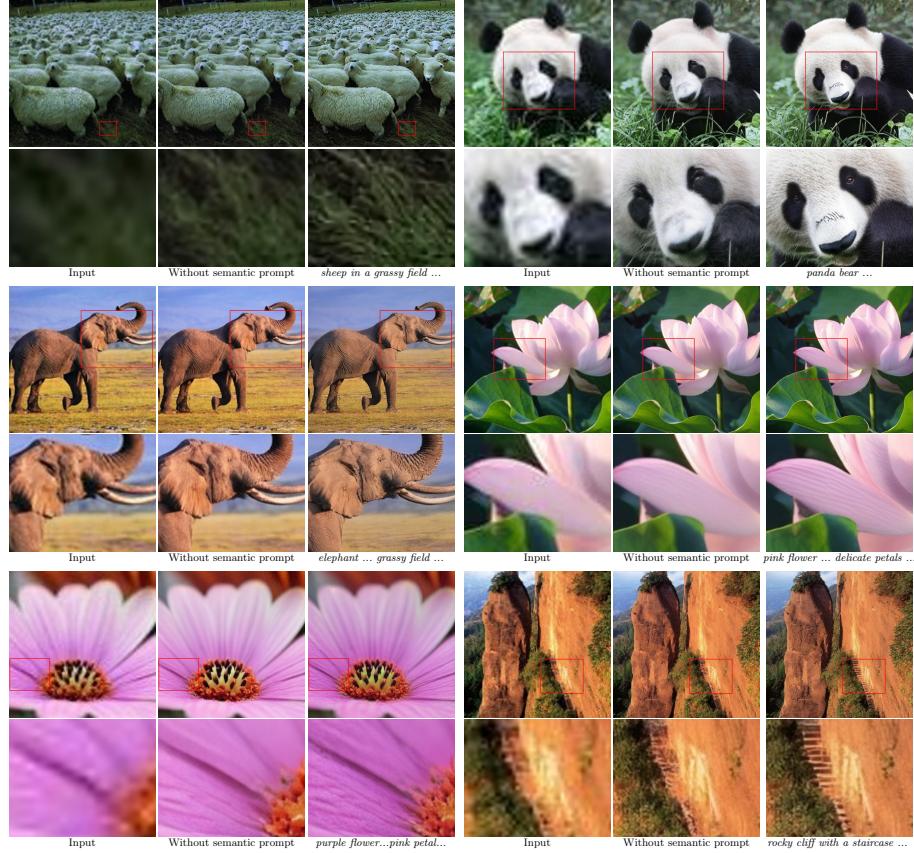


Fig. 13: Real-world examples showing the effect of semantic prompts.

CLIP considers raindrop but Real-esrgan does not). Thus, comparing our framework and other concurrent work (*e.g.*, DiffBir) to all-in-one restoration techniques proves difficult. To alleviate the concern, we evaluate our framework on the denoising testset of CBSD68 and our method achieves a comparable LPIPS (0.305) with DA-CLP (0.294).

J More Visual Comparison

More visual comparisons with baselines are provided in Fig. 15.

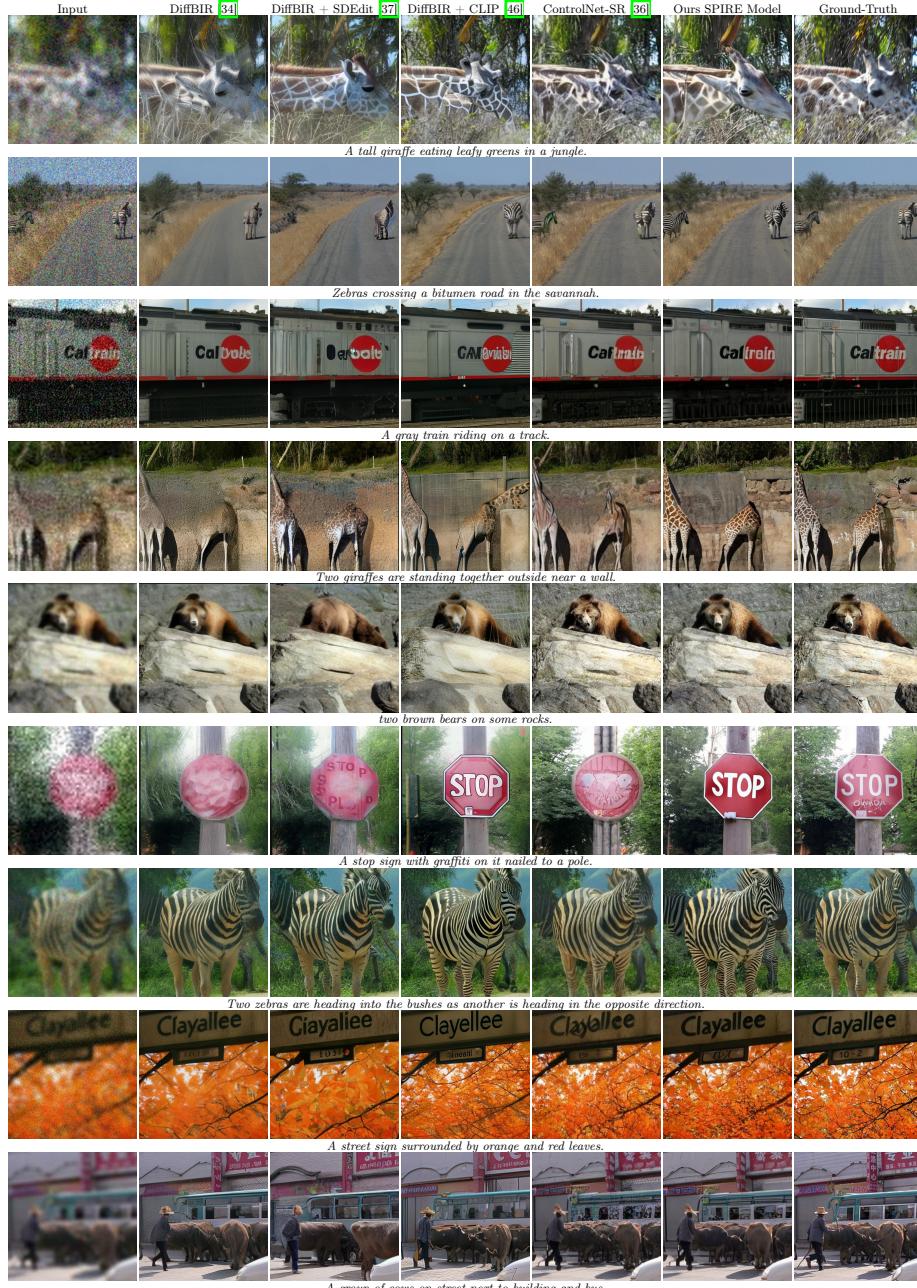


Fig. 15: Main visual comparison with baselines. (Zoom in for details)