Adapting Text-to-Image Generation with Feature Difference Instruction for Generic Image Restoration

Supplementary Material

1. More Dataset Details

In this section, we provide additional details about the mixed degradation dataset described in Section 4. The dataset comprises images for 10 distinct image restoration tasks, including rainstreak, raindrop, haze, snow, low-light, noise, motion blur, defocus blur, shadow, and JPEG compression. Detailed descriptions of these datasets are provided below:

- Rainstreak: For single-task setting, we use the Rain100H [88] dataset, which contains 1,800 images for training and 100 images for testing. We also conduct experiments on the Rain100L [88] and Rain1400 [19] datasets for different raining intensities and cross-dataset experiments. Rain100L also consists 1,800 training and 100 testing images with simpler scenes and sparser rainstreaks compared to Rain100H. For the Rain1400 dataset, we select 3,780 images for training with three different intensities of rain and 1,400 images for testing.
- Raindrop: we use RainDrop [57] dataset with 861 images for training and 58 images for testing.
- Haze: The dataset is derived from the RESIDE-6k dataset [58], which includes a mixture of indoor and outdoor images. It provides 6,000 images for training and 1,000 images for testing.
- Snow: The Snow100K [42] dataset is a large-scale benchmark designed for snow removal in image restoration tasks, consisting of 100,000 synthetic images with varying snow degradation levels. Following DA-CLIP [45], we utilize a subset comprising 1,872 images for training and 601 images for testing.
- Low-light: We use the LOL-v1 [83] dataset for single-task setting, which contains 485 images for training and 15 images for testing. We also use the MEF [47] dataset to evaluate the performance under different lighting conditions. Note that our FDI Adapter only uses LOL-v1 for training, and we select 17 images from MEF for further testing.
- Noise: We collect 3,440 training images from the DIV2K [2] and Flickr2K [75] datasets and use the CBSD68 [48] dataset for testing, which contains 68 images, and the noise level is set to 50.
- Motion blur: We use the GoPro [51] dataset consisting of 2,103 images for training and 1,111 images for testing.
- Defocus blur: We use the DPDD [1] dataset, which contains 350 training images and 150 testing images.
- Shadow: We use the SRD [59] dataset with 2,680 images for training and 408 images for testing.
- JPEG compression: the training dataset is the same as the denoising task, and we use 29 images from the LIVE1 [71] dataset for testing. The JPEG quality factor is set to 10.
- Mixed-degradation: for raindrop + rainstreak, we use the RainDS [60] dataset with 1,000 images for training and 200 images for testing. For rainstreak + haze + low-light, we use the Outdoor-Rain [77] dataset, which contains 750 images for training and 50 images for testing. We also use the BID [21] dataset for a more complex combination: rain streak + snow + haze + raindrop. We collect 3,975 training images and 500 images for testing.

2. More Implementation Details

We use the pretrained Stable Diffusion v2.1-512-base-ema [1] model and the pretrained BLIP-2-opt-2.7b [2] as backbones. Images are resized to 512 $\times$ 512 during our experiments. Following T2I Adapter [50], we evenly divide the 50-step DDIM sampling into 3 stages and only add guidance information to the first beginning stage. Besides the default training loss $\mathcal{L}_{\text{LDM}}$, we further introduce an $\mathcal{L}_1$ reconstruction loss to ensure better alignment between the latent space representation and the ground truth. Specifically, for the i-th elements of the respective latent vectors, the reconstruction loss is defined as:

$$\mathcal{L}_{\text{recon}} = \|z_t - z_{\text{GT}}\|_1 = \sum_i \left| z_t^{(i)} - z_{\text{GT}}^{(i)} \right|, \tag{6}$$

where z_t is the latent representation at timestep t, and z_{GT} is the latent representation of the ground truth image. This additional loss penalizes deviations in the latent space and enforces consistency during the denoising process.

The total training loss is thus formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LDM}} + \lambda_{\text{recon}} \mathcal{L}_{\text{recon}}, \tag{7}$$

where λ_{recon} is a hyperparameter that balances the contribution of the reconstruction loss to the overall objective, which is empirically set as 0.1. By incorporating $\mathcal{L}_1$ reconstruction loss, the model benefits from improved robustness and better fidelity in the generated outputs.

[1] https://huggingface.co/stabilityai/stable-diffusion-2-base

[2] https://huggingface.co/Salesforce/blip2-opt-2.7b

Figure 9. More examples of cross-dataset experiments: given a test input, we apply FDI from different datasets within the same task to perform restoration. Top row: simple synthetic images. Bottom row: real-world images. Please zoom in to see the details.

Figure 10. Examples of cross-task experiments: given a test input, we apply FDI from datasets for another task to perform restoration. Left: Denoising FDI for desnowing. Right: Dehazing FDI for underwater image restoration (IR). Please zoom in to see the details.

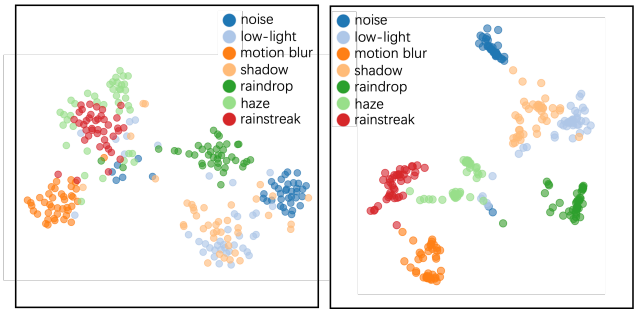

Figure 11. t-SNE visualization of FDI before and after the noise decoupling. Left: original FDI F. Right: refined FDI $\hat{F}$.

3. Discussion

3.1. Cross-dataset combinations

As mentioned in Section 4.1, there exists a certain level of compatibility between different datasets for the same task. This inspires us to evaluate the generalizability of existing training datasets to real-world degraded images by leveraging various FDI dataset sources. As shown in Figure 9, it can be observed that for simpler scenes (top row), there is a significant level of generalization across most datasets. However, for complex real-world scenes, more sophisticated datasets with richer scene or degradation modeling are required for effective FDI guidance.

3.2. Cross-task combinations

For some similar tasks, we observe a certain level of compatibility between different tasks due to similarities in the types of degradations. As illustrated in Figure 10, FDI derived from denoising datasets can be applied to desnowing, and FDI from dehazing datasets can guide underwater image restoration. Notably, denoising FDI may result in a loss of some image details, while FDI from dehazing datasets, influenced by their inherent image distributions, may introduce varying degrees of color distortion.

3.3. Analysis on FDI

Besides, we also visualize the t-SNE statistics of the original FDI F and the refined FDI $\hat{F}$ in Figure 11. (a) As shown in the left figure, FDI derived from datasets with different degradation types exhibits a natural degree of separability. However, certain overlaps between different FDIs are also observed, particularly for degradation types that frequently co-occur, such as rain streaks and haze. (b) After applying noise decoupling, degradation-independent noise is effectively eliminated (right figure). Moreover, it can be observed that FDIs for certain tasks still retain some level of similarity, such as those between low-light and shadow. This further validates the cross-task compatibility of FDIs.

4. More Experimental Results

More visual comparisons on other tasks are given in Figure 12 (single-degradation) and Figure 13 (mixed-

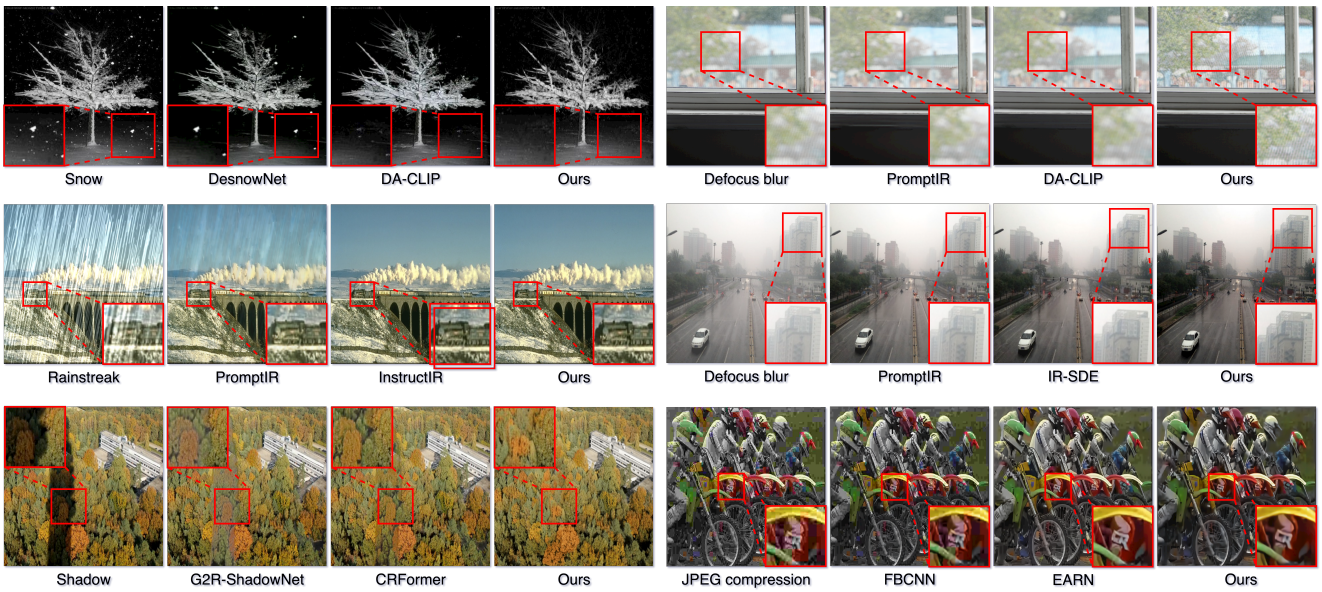

Figure 12. More visual comparison under different single-task degradations. Please zoom in to see the details.

Table 6. Ablation on different proportions of the training set.

Training Set	Deraining (Rain100H)			LLIE (LOL-v1)		
Proportion	30%	60%	whole	30%	60%	whole
PSNR ↑	29.57	31.88	34.55	23.19	24.28	24.55
LPIPS ↓	0.176	0.045	0.023	0.158	0.091	0.075

Table 7. Comparisons of single-type degradation efficiency. All models are tested under the same environment.

Method	**Prompt-based**	**T2I model-based**		
	PromptIR	DiffIR	DA-CLIP	**Ours**
Param (M)	102.50	375.81	88.25	45.10
Inference time (s)	0.88	9.97	5.11	3.04

degradation). It can be seen that our method stably produces well-structured results with finer details while remaining robust against different noise combinations.

Influence of image pair numbers. We further conducted comparisons by randomly selecting varying proportions of the training set, which can also be regarded as a few-shot setting. Tab. 6 shows that more reference pairs usually yield FDI better representing degradation characteristics. For tasks like low-light enhancement with similar degradations across images rather than various patterns like rain, fewer training pairs still achieve satisfactory results.

Computing cost. Table 7 further shows the comparisons in terms of computing efficiency. We use pre-trained BLIP-2 to obtain FDI, thus only the FDI adapter contributes to the parameter count, which is less than other prompt or T2I-based methods. Note that for multi-degradation, different task-specific adapters can be trained in parallel.

More discussions. ***1. CLIP vs. BLIP-2 encoders.*** Compared to the CLIP encoder, the Q-former can filter out more degradation-irrelevant information from image embeddings [17, 56], as shown in Figure 11. ***2. ControlNet vs. Adapter.*** ControlNet-based methods (*e.g.*, DiffBIR [40] and SUPIR [89]) usually result in a high parameter count due to the parameter replication [50]. Besides, these methods are better suited for blind image enhancement (*e.g.*, denoising and super-resolution), which can be restored using generative priors from diffusion models. For degradations like rain and fog that obstruct the original content, their reconstruction process without explicit degradation guidance may damage the content consistency. However, for blind image restoration, *e.g.*, image super-resolution or denoising, our semantic difference-based FDI is not well-suited, as such degradations often do not affect the image semantics, making them difficult to represent through BLIP-2 feature differences. We are experimenting to improve this by pertaining/fine-tuning the Q-Former to make it more sensitive to subtle degradations (future work).

5. More Applications

Benefiting from the reference-based mechanism, our method can also be applied to image-to-image translation tasks such as style transfer, as shown in Figure 14. In this task, the input reference image pair corresponds to images with different styles from the datasets. Additionally, our method enables the fusion of different styles through the combination of multiple FDI adapters.

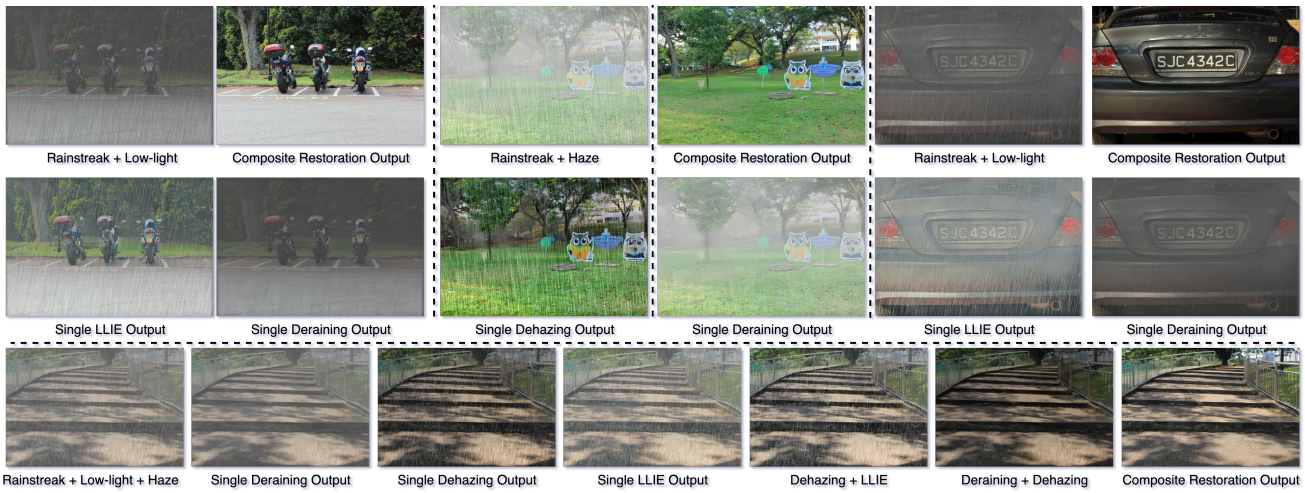

Figure 13. More visual results under different mixed-degradation combinations. Please zoom in to see the details.

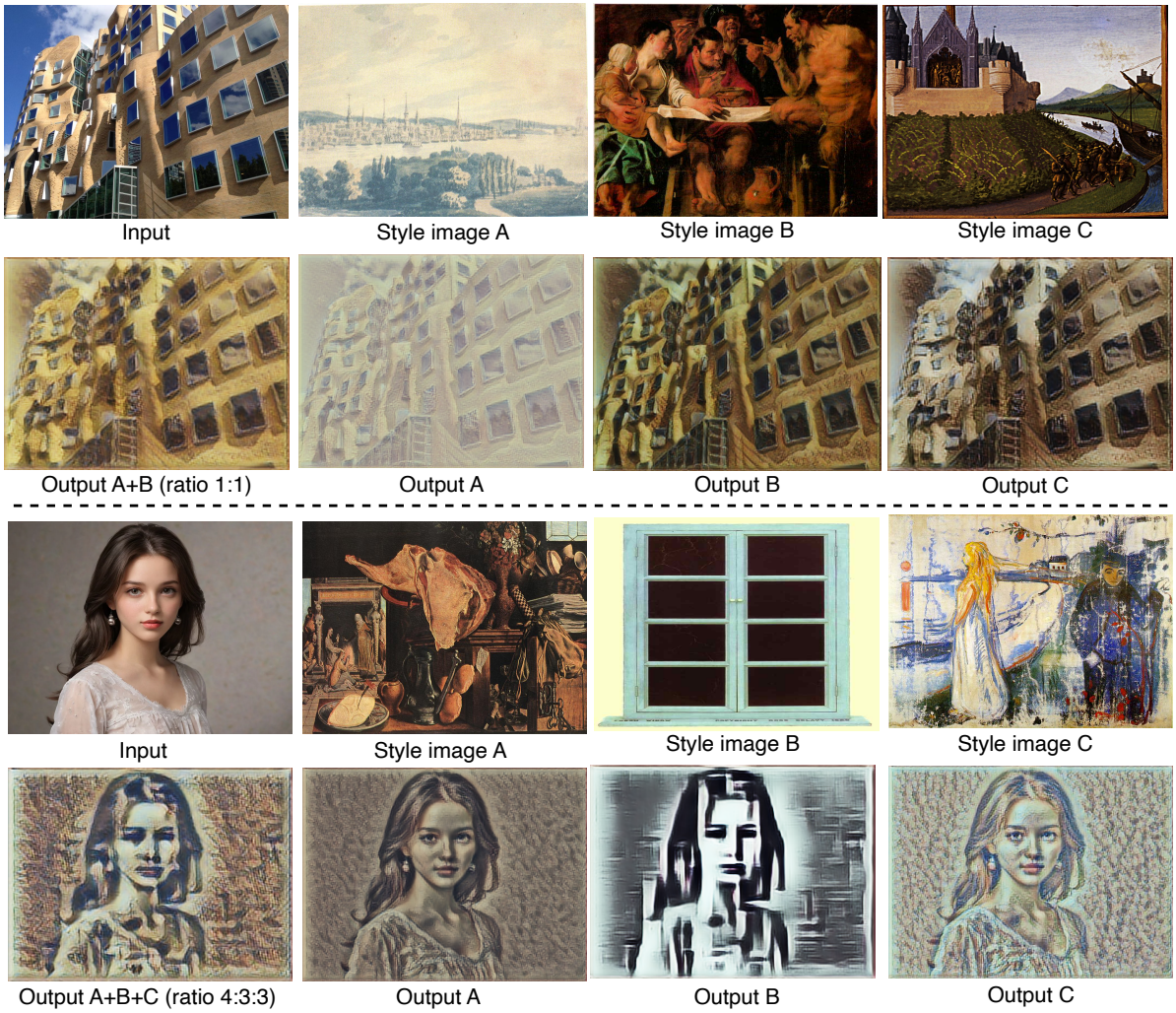

Figure 14. Examples of style transfer. Please zoom in to see the details.