JarvisIR: Elevating Autonomous Driving Perception with Intelligent Image Restoration

Supplementary Material

This is the supplementary material for the paper: "*JarvisIR: Elevating Autonomous Driving Perception with Intelligent Image Restoration.*" We provide the following materials in this manuscript:

- Sec.1 More implementation details.
  - Restoration tool settings.
  - Details of Model Setups.
- Sec.2 CleanBench dataset details.
  - Dataset statistics.
  - Details of degradation library.
- Sec.3 More ablation.
  - MRRHF vs. vanilla RRHF.
  - Sample generation strategy and entropy regularization.
  - Effectiveness of differentiated contrast weights.
  - Impact of reasoning for decision-making.
  - Impact of reward model.
- Sec.4 More visual results.
- Sec.5 Limitations, broader impacts and future work.

1. More implementation details

1.1. Restoration tool settings

Table 1 lists the task-specific restoration tools used in our implementation. Notably, some models lack weights corresponding to certain tasks but are inherently adaptable; we collect appropriate data to retrain them. For example, Img2img-turbo [18] is an image-to-image translation method based on SD-turbo that provides night-to-day and rainy-to-day weights but not snow-to-day weights. To enable Img2img-turbo to adapt to snow scenes, we retrain it using the CycleGAN paradigm on the snow subset of the ACDC dataset [22]. Additionally, it is important to note that we are not utilizing the latest state-of-the-art tools, suggesting considerable potential for enhancing our models.

1.2. Details of Model Setups

Model Architecture. In this study, JarvisIR primarily adopts the architecture from Llava-Llama3-8B [16]. Specifically, the input images and instruction texts are first tokenized, then fused, and finally processed by the Large Language Model (LLM) for response generation. (a) Tokenization of input images and instruction texts: We use a frozen CLIP pre-trained ViT-L/14 [21] as the image encoder to convert input images into visual tokens. The instruction texts are tokenized into textual tokens using the SentencePiece tokenizer [13]. To bridge the different embedding spaces of visual and textual tokens, we implement a trainable image projector to map visual tokens into the textual space, following [23, 44]. (b) Token Fusion: We integrate the visual tokens into predefined positions within the textual tokens to achieve token fusion. (c) Response Generation Using LLM: The fused tokens are fed into the LLM to generate the final response. In our experiments, we primarily use Llama3-8B [23]. Even with their advanced features, pre-trained LLMs lack the ability to furnish accurate responses, thorough reasoning regarding degradation, and precise restoration plans without dataset-specific fine-tuning. Therefore, we employ a full parameter fine-tuning technique that efficiently unleashes the potential of LLM to the maximum extent.

Model setup. Since the CLIP pre-trained ViT-L/14 [21] encodes each 14×14 image patch into a visual token, the input image dimensions must be integer multiples of 14. Therefore, we zero-pad the input images to meet this requirement. We encode the image patches into visual tokens using the CLIP pre-trained ViT-L/14 [21], where each token is a 1024-dimensional vector. These visual tokens are subsequently projected by the image projection layer into the LLM's hidden dimension of 4096.

Training setup. Both the SFT and MRRHF tuning phases utilize the Adam optimizer with learning rate 1e-5 with cosine decay. The warmup ratio is set to 0.03, the maximum sequence length is 2048, and the weight decay is 4. JarvisIR-SFT undergoes training for three epochs with a batch size of 128, while JarvisIR-MRRHF is trained for three epochs using a batch size of 2. During the MRRHF tuning phase, the diverse beam search settings include a size of 3, 5 beam groups, a diversity penalty of 2.0, and a sampling temperature of 0.8. Training is conducted on 8 GPUs (NVIDIA A100 80G).

2. CleanBench dataset details

2.1. Dataset statistics

CleanBench. In constructing the CleanBench process, we collected large-scale raw daytime images from various sources, including autonomous driving datasets [2, 2, 22, 38] and natural datasets [3, 10, 14, 15, 37, 43] .The CleanBench dataset contains a total of 150K degraded-clean image pairs. For the construction of CleanBench-Real, we gathered 80K real degraded images consisting of night scenes, fog scenes, snow scenes and rain scenes. These data

Table 1. Task-specific restoration tools with descriptions.

Task	Tools	Model Description
Super-resolution	StableSR-turbo [26]	Utilizes pre-trained diffusion models with a time-aware encoder for high-quality super-resolution, deblurring, and artifact removal.
	Real-ESRGAN [27]	Fast GAN for super-resolution, deblurring, and artifact removal, handling complex real-world degradations efficiently.
Denoising	SCUnet [40]	Hybrid UNet-based model combining convolution and transformer blocks, designed for robust denoising under diverse real-world noise conditions.
Compression artifact removal	StableSR-turbo [26]	Utilizes pre-trained diffusion models with a time-aware encoder for high-quality super-resolution, deblurring, and artifact removal.
	Real-ESRGAN [27]	Fast GAN for super-resolution, deblurring, and artifact removal, handling complex real-world degradations efficiently.
Deblurring	StableSR-turbo [26]	Utilizes pre-trained diffusion models with a time-aware encoder for high-quality super-resolution, deblurring, and artifact removal.
	Real-ESRGAN [27]	Fast GAN for super-resolution, deblurring, and artifact removal, handling complex real-world degradations efficiently..
Deraining	IDT [33]	Transformer-based model for de-raining and raindrop removal.
	UDR-S2Former [4]	An uncertainty-aware transformer model for rain streak removal.
	Img2img-turbo-rain [18]	Efficient model based on SD-turbo, designed for fast and effective rain removal in real-world images.
Raindrop removal	IDT [33]	Transformer-based model for de-raining and raindrop removal.
Dehazing	RIDCP [32]	Efficient dehazing model utilizing high-quality codebook priors to handle complex real-world haze.
	KANet [8]	Efficient dehazing network using a localization-and-removal pipeline to handle complex real-world hazy.
Desnowing	Img2img-turbo-snow [18]	Efficient model for removing snow artifacts while preserving natural scene details.
	Snowformer [5]	Transormer-based model for removing snowflakes while preserving natural scene details.
Low-light enhancement	Img2img-turbo-night [18]	Fast and efficient model based on SD-turbo, designed for low-light enhancement in real-world scenarios.
	HVI-CIDNet [34]	Lightweight transformer for low-light and exposure correction, enhancing both image quality and downstream vision tasks efficiently.
	LightenDiff [9]	Diffusion-based framework for low-light enhancement, leveraging Retinex theory and latent-space decomposition for high-quality unsupervised restoration.

come from diverse sources, including the aforementioned autonomous driving datasets. Additionally, to enhance the generalizability of JarvisIR in natural contexts, we incorporated natural adverse weather scenes from internet and public datasets [3, 10, 14, 15, 17, 20, 37, 43].

2.2. Details of degradation library

As described in Sec 3.1 of the manuscript, we simulate realistic adverse weather scenarios—rainy, nighttime, snowy, and foggy conditions—by customizing a degradation library developed with physical models and image transformation techniques to synthesize degraded images. In this section, we detail our degradation implementations, cov-

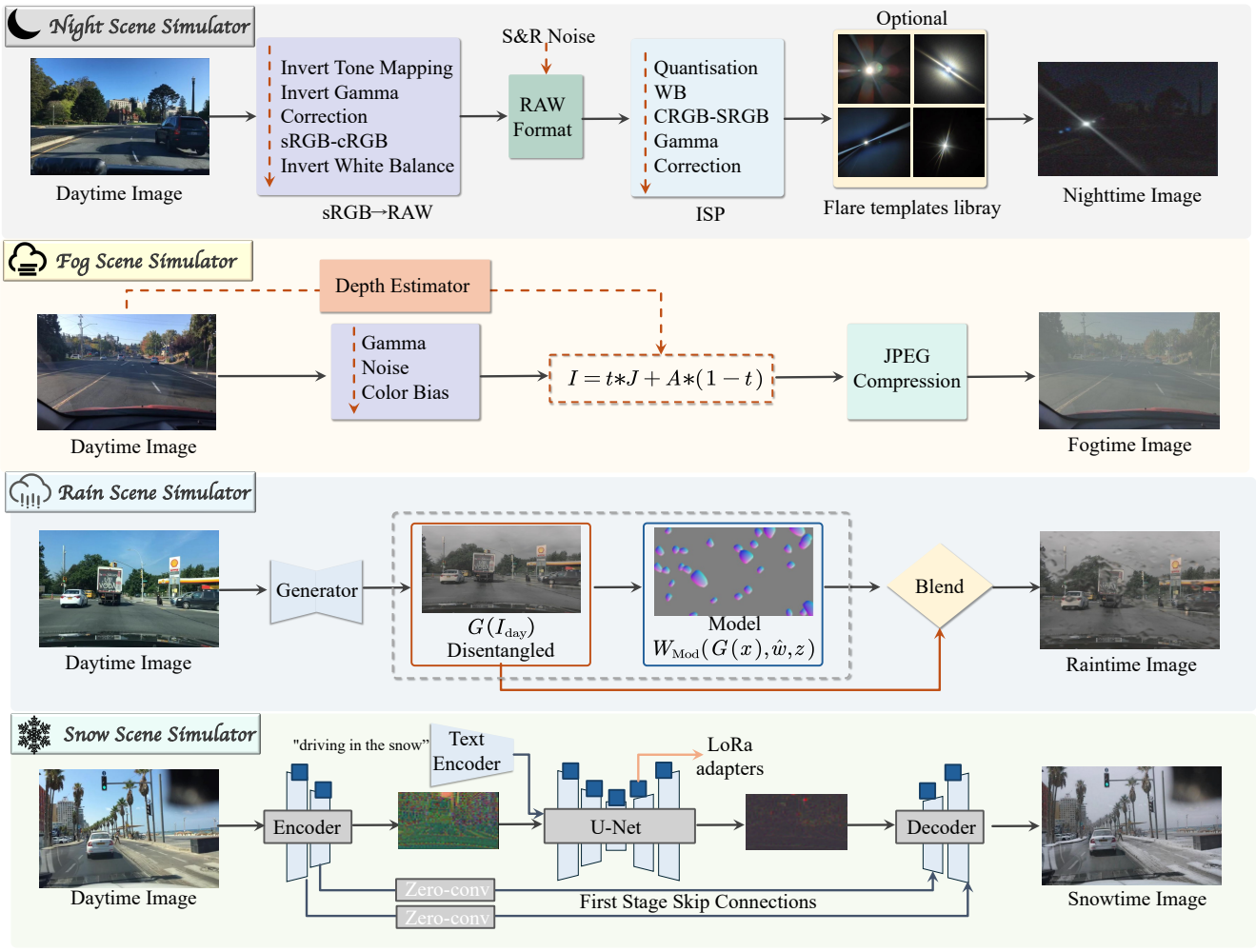

Figure 1. Adverse weather scene simulator. To simulate realistic adverse weather scenarios, including rainy, nighttime, snowy, and foggy, we customized the degradation library developed using physical models and image transformation techniques to synthesize degraded images.

ering the principles, formulas, and severity setups for the Night Scene Simulator, Fog Scene Simulator, Rain Scene Simulator, and Snow Scene Simulator. Examples for each implementation are provided in Figure 4.

Night Scene Simulator. Inspired by the work of [6], we employ a low-light degradation transform to synthesize realistic low-light images, denoted as T_{night}, as illustrated in Figure 1. Specifically, we first convert the daytime image I_{day} into RAW data using the sRGB→ RAW process [1]. Next, we linearly attenuate the RAW image and introduce Shot and Read (S&R) noise, which is commonly observed in camera imaging systems [1]. Finally, we apply the Image Signal Processing (ISP) pipeline to convert the low-light sensor data back into sRGB format. Additionally, we incorporate flare degradation using flare templates from the Flare7K++ [7] dataset. The complete low-light degradation transform T_{night} is given by:

$$T_{\text{night}}\left(\mathrm{I}_{\text{day}}\right) = T_{ISP}\left(T_{sRGB\rightarrow RAW}\left(\mathrm{I}_{\text{day}}\right) + \mathrm{I}_{\text{noise}}\right) + \mathrm{I}_{\text{flare}}, \tag{1}$$

which generates a degraded image I_{day} that closely resembles a dark nighttime scene. Furthermore, we use an online dynamic degradation process. It applies randomized parameter combinations, as defined in Equation 1, to simulate diverse nighttime driving conditions.

Fog Scene Simulator. Inspired by RIDCP [32], we design a foggy image degradation transform, denoted as T_{fog}, to synthesize realistic hazy images, as shown in Figure 1. Specifically, we simulate fog by introducing transmission maps $t(x)$ using depth estimation algorithms (e.g., Depth anything V2 [35]), combined with exponential attenuation $e^{\beta d(x)}$, where β controls haze density within the range $[0.3, 1.5]$. Additionally, poor lighting conditions are

modeled by applying a brightness adjustment factor $\gamma \in [1.5, 3.0]$, Gaussian noise $\mathcal{N}$, and atmospheric light variation $A + \Delta A$, where ΔA is sampled from $[-0.025, 0.025]$. To further enhance realism, JPEG compression artifacts are introduced by applying JPEG $(\cdot)$ to the degraded image. The complete foggy image synthesis process is defined as:

$$T_{\text{fog}}\ (I_{\text{day}}) = \text{JPEG}\left(\mathcal{P}\left(I_{\text{day}}^{\gamma} + \mathcal{N}, e^{\beta d(x)}, A + \Delta A\right)\right), \tag{2}$$

where $\mathcal{P}$ represents the hazy image formation process, I_{day} is the clean image, and $d(x)$ is the estimated depth map. The variable x refers to the spatial coordinates of the image. This dynamic degradation process is designed to operate online with randomized parameters, simulating diverse real-world foggy conditions.

Rain Scene Simulator. Inspired by PGDGN [19], we introduce a rain degradation transform, denoted as T_{rain}, to generate realistic rainy images (Figure 1). This transform synthesizes rainy images by combining a disentangled clean image with a physics-based rain rendering model. The degradation process is formulated as:

$$T_{\text{rain}}\ (I_{\text{day}}) = W_{\text{Mod}}(G(I_{\text{day}}), \hat{w}, z), \tag{3}$$

where I_{day} is the clean image, $G(I_{\text{day}})$ represents the disentangled base image, and W_{Mod} is the rain rendering model. W_{Mod} incorporates parameters $\hat{w} = \{\hat{w}_d, \hat{w}_{nd}\}$, with $\hat{w}_d$ controlling differentiable aspects such as raindrop size and streak density, and $\hat{w}_{nd}$ addressing nondifferentiable properties. The term z introduces stochastic noise for variability in rain effects. This process applies W_{Mod} to add realistic raindrop occlusions, rain streaks, and scene wetness to the disentangled image. $G(I_{\text{day}})$, generating a visually plausible rainy image $T_{\text{rain}}\ (I_{\text{day}})$ with controlled and diverse effects.

Snow Scene Simulator. Building on the img2img-turbo model [18], we introduce a snow transformation, denoted as T_{snow}, to generate realistic snowy images from daytime inputs. This process uses the SD-Turbo model with textual conditioning. It synthesizes snowy scenes by combining the input image with a latent diffusion-based generator and a textual prompt. The snow transformation is formulated as:

$$T_{\text{snow}}\ (I_{\text{day}}, C_{\text{snow}}) = G_{\text{snow}}\ (I_{\text{day}}, C_{\text{snow}}), \tag{4}$$

where I_{day} is the daytime input image, C_{snow} is the textual condition (e.g., "driving in the heavy snow"), and G_{snow} represents the generator. By employing LoRA adapters and skip connections, the generator enables precise control over scene characteristics while maintaining the structural integrity of the input image. This process applies G_{snow} to infuse the daytime image I_{day} with snowy features, guided by the contextual information in C_{snow}. The resulting synthetic image aligns closely with the visual expectations of a snowy environment while maintaining consistency with the original scene's structure.

3. More ablation

To thoroughly investigate the proposed JarvisIR, we conducted an extensive array of ablation studies on the CleanBench-Real dataset. Four non-reference metrics are used for assessment: MUSIQ [11], MANIQA [36], CLIP-IQA+ [25], LIQE [41]. The specific elements of these studies are further expounded in the sections that follow.

3.1. MRRHF vs. vanilla RRHF

We evaluate the effectiveness of our proposed MRRHF by comparing it with vanilla RRHF [39]. The reward and diversity metrics over training iterations are illustrated in Table 2. Fine-tuning JarvisIR with MRRHF significantly improves the average values of both reward and diversity by 0.19 and 3.43, respectively, compared to using RRHF. The degradation in diversity and reward when using vanilla RRHF results from its offline sample generation strategy. As discussed in Sec. 5 in the manuscript, this strategy confines its generated samples to the finite sample space created by the SFT model using diverse beam search [24]. In contrast, our MRRHF employs a hybrid sample generation strategy and entropy regularization, providing sufficient sample exploration space to achieve globally optimal results.

3.2. Sample generation strategy and entropy regularization

In our manuscript, we examine the effects of the sample generation strategy and entropy regularization on the MRRHF tuning process, focusing on reward scores and response diversity. This section provides further evidence of the effectiveness of our hybrid sample generation strategy and entropy regularization. Specifically, as shown in Table 2, we assess their impact on performance using the CleanBench-Real validation set. The results demonstrate that our hybrid sampling approach and entropy regularization not only enhance training stability and facilitate high-quality exploration of the optimization space but also significantly improve testing performance.

3.3. Effectiveness of differentiated contrast weights

In Equation 4 of our manuscript, we refine the original ranking loss [39] by introducing differentiated contrast weights, expressed as $L_{\text{rank}} = \sum_{s_i<s_j} (s_j - s_i) \max(0, p_i - p_j)$. The term $(s_j - s_i)$ represents the differentiated contrast weights. We compare this with the original ranking loss $\hat{L}_{\text{rank}} = \sum_{s_i<s_j} \max(0, p_i - p_j)$. Table 2 presents the reward and diversity metrics over training iterations. When the ranking loss is applied without differentiated contrast weights $\hat{L}_{\text{rank}}$ the average values of both reward and diversity decrease by 0.14 and 3.93, respectively, compared to using L_{rank}. We attribute this to the differentiated con-

Table 2. Ablation studies on tuning paradigm, differentiated contrast weights, different sample generation strategies, and entropy regularization. The "Reward" represents the average reward scores obtained during alignment tuning, spanning from -1 to 1. A negative score indicates a penalty, while a positive score represents a reward. The "Diversity" reflects the average number of unique responses produced during the training process. Additionally, we evaluate performance on the CleanBench-Real validation set using four non-reference metrics: MUSIQ, MANIQA, CLIP-IQA+, and LIQE. The reported values represent the average performance across all tested scenes.

Strategy	Reward	Diversity	MUSIQ ↑	MANIQA ↑	CLIP-IQA+ ↑	LIQE ↑
Vanilla RRHF	0.40	3.12	63.89	0.5090	0.5388	3.589
MRRHF (**Ours**)	**0.67**	**6.55**	**71.43**	**0.7099**	**0.7296**	**4.411**
w/o. differentiated contrast weights	0.53	2.62	63.22	0.5871	0.6130	3.597
w. differentiated contrast weights (**Ours**)	**0.67**	**6.55**	**71.43**	**0.7099**	**0.7296**	**4.411**
offline sample generation	0.43	3.63	64.12	0.5323	0.6012	3.620
online sample generation	-0.87	1.27	-	-	-	-
hybrid sample generation (**Ours**)	**0.67**	**6.55**	**71.43**	**0.7099**	**0.7296**	**4.411**
w/o. entropy regularization	0.50	4.56	65.06	0.6207	0.6915	3.867
w. entropy regularization (**Ours**)	**0.67**	**6.55**	**71.43**	**0.7099**	**0.7296**	**4.411**

trast weights enabling the VLM to recognize that some negative examples are neutral (with reward scores close to positive examples) and thus should not be excessively penalized, which helps prevent confusion during VLM training. Specifically, assuming the system uses diverse beam search to obtain multiple responses $r_1, \ldots, r_i, r_k, r_n$ the original RRHF algorithm treats the best response r_k as positive and the remaining responses $r_i < r_k$ as negative examples of r_k and applies the same penalty to them. However, this approach may not be reasonable, especially when the preference scores of different r_i are similar. For instance, when the preference of r_{k+1} is only slightly worse than r_k, while r_n is significantly worse than r_k, the model should differentiate and apply different penalty strengths, slightly penalizing r_{k+1} and heavily penalizing r_n compared to r_k. To address this, we propose using the score $\mathcal{S}(r_i)$ from a reward model $\mathcal{S}(\cdot)$ to indicate the numerical preference of r_i, i.e., the differentiated contrast weights $(s_j - s_i)$.

3.4. Impact of reasoning for decision-making

As the pioneering work [29] points out, Chain-of-Thought (CoT) is "a series of intermediate reasoning steps" that has proven effective in complex reasoning tasks [12, 29, 42]. The main idea of CoT is to prompt large language models (LLMs) to output not only the final answer but also the reasoning process leading to it, resembling human cognitive processes. Inspired by this approach, we enable JarvisIR to provide detailed degradation and reasoning insights about the degraded image before making decisions, specifically before producing the task sequence with model selection. To assess the impact of reasoning on final decision-making, we perform ablation experiments on the CleanBench-Real validation set by comparing two variants: (1) directly requesting JarvisIR to output the task sequences, and (2) providing detailed degradation and reasoning insights before outputting the task sequences. As shown in Table 3, providing detailed degradation and reasoning insights significantly enhances JarvisIR's decision-making, leading to notable improvements in the four non-reference metrics. By explicitly describing degradations and reasoning insights, the model can use in-context learning to align selected tasks and restoration experts with the specific degradations present. This strategy not only enhances interpretability but also introduces constraints that make the model's decisions more reliable in real-world scenarios.

Table 3. Ablation studies on the impact of reasoning for decision-making. We evaluate performance on the CleanBench-Real validation set using four non-reference metrics: MUSIQ, MANIQA, CLIP-IQA+, and LIQE. The reported values represent the average performance across all tested scenes.

Configurations	MUSIQ ↑	MANIQA ↑	CLIP-IQA+ ↑	LIQE ↑
w/o. reasoning	71.17	0.6942	0.7156	4.394
(Ours) w reasoning	**71.43**	**0.7099**	**0.7296**	**4.411**

3.5. Impact of reward model

To analyze how various reward model configurations affect model optimization, we conducted an ablation experiment exploring three distinct settings: (I) multiple VLM-based IQA models as a unified reward model (e.g., Q-instruct [31] and Q-align [30]). (II) using a single VLM-based IQA model (e.g., Q-Instruct [31] or Q-align [30]) or a traditional IQA model (e.g., MUSIQ [11] or MANIQA [36]). (III) multiple traditional IQA models as a unified model (e.g., MUSIQ [11] and MANIQA [36]). The results of JarvisIR-MRRHF trained with different reward models are summarized in Table 4. Based on the results, we make the following observations: (1) Using multiple VLM-based IQA models as the reward model significantly improves perception metrics, although it increases resource consumption during training. (2) Training with a single IQA model improves the corresponding metric sig-

Table 4. Ablation studies on the impact of different reward model configurations. We evaluate performance on the CleanBench-Real validation set using four non-reference metrics: MUSIQ, MANIQA, CLIP-IQA+, and LIQE. The reported values represent the average performance across all tested scenes.

Configurations	MUSIQ ↑	MANIQA ↑	CLIP-IQA+ ↑	LIQE ↑
(I) Q-align [30] + Q-Instruct [31]	71.41	0.7094	0.7308	4.419
(II) Q-align	71.35	0.7086	0.7288	4.409
(II) Q-Instruct [31]	71.37	0.7093	0.7257	4.402
(II) MUSIQ [11]	71.64	0.6932	0.6977	3.955
(II) MANIQA [36]	68.49	0.7126	0.6805	3.981
(III) MUSIQ [11] + MANIQA [36]	71.52	0.7118	0.7068	4.127
(**Ours**) Q-Instruct [31]+ MUSIQ [11] + MANIQA [36]	**71.43**	**0.7099**	**0.7296**	**4.411**

nificantly, but other metrics may experience some degradation. (3) Combining multiple traditional IQA models as the reward model enhances performance on certain metrics, but the improvements are asymmetrical—some traditional metrics exhibit very high performance while perception metrics are relatively low. Consequently, we opt to create the unified reward model by combining both VLM-based and non-VLM-based IQA models, such as Q-instruct [31], MUSIQ [11], and MANIQA [36]. This combination allows for a comprehensive evaluation of system responses while preserving training efficiency.

4. More visual results.

4.1. Perception restoration

Additional visual comparisons highlight the effectiveness of the proposed JarvisIR framework in real-world adverse weather conditions. Figure 2 illustrates the comprehensive workflow of JarvisIR, which begins by receiving user commands and degraded images. JarvisIR evaluates the image quality, identifies degradation factors, and formulates task sequences. It then selects appropriate models for tasks such as denoising, dehazing, and super-resolution. The outputs include evaluated inference insights, detailed restoration plans, and enhanced images, effectively bridging user instructions with image restoration plans.

Figure 3 illustrates the decision-making processes of both JarvisIR-MRRHF and JarvisIR-SFT. Experimental results indicate that the decision-making capability of JarvisIR-MRRHF surpasses that of JarvisIR-SFT. Specifically, JarvisIR-MRRHF makes correct decisions in cases where JarvisIR-SFT previously failed. For example, in coupled degraded real rain scenarios (the first row), JarvisIR-SFT yields a mediocre decision—"Enhancement (Img2img-turbo) → Dehaze (RIDCP) → DeRaindrop (IDT)"—which does not remove raindrops and blur the background. However, JarvisIR-MRRHF accurately identifies the appropriate restoration tasks and selects the optimal models to solve them: "Denoise (SCUNet) → DeRaindrop (IDT) → Deblur (StableSR-turbo)". This improvement confirms that MRRHF fine-tuning significantly enhances JarvisIR's decision-making ability under real-world conditions, reduces hallucination errors, and improves generalization performance.

Figures 5, 6, 7, and 8 illustrate visual comparisons of our method and the baseline methods across four different scenes on the CleanBench-Real test set. Our results demonstrate that JarvisIR outperforms the comparative methods in terms of color enhancement, detail preservation, and the elimination of degradations, achieving a superior balance among these aspects. Conversely, the baseline methods perform poorly in real-world environments. They struggle to handle coupled degradations that occur simultaneously in natural settings, such as low light combined with fog or a mixture of rain and fog. These limitations may arise from their heavy dependence on specific degradation priors and significant domain gaps due to mismatches between synthetic training data distributions and real-world data. Consequently, they often produce subpar recovery results featuring artifacts, overexposure, underexposure, and amplified noise.

5. Limitations, broader impacts and future work

The primary limitation of our research is that JarvisIR is unable to address all real-world restoration scenarios. While it demonstrates effectiveness in handling most degradation scenarios relevant to autonomous driving, it does not extend to tasks such as underwater image restoration, old photo enhancement, or blind face restoration. By incorporating appropriate data and tools, rapid adaptation could be achieved through the proposed training paradigm. Furthermore, the tools currently employed are limited in scope and capability. In our future work, we will incorporate more advanced and robust restoration tools that might further enhance JarvisIR's ability to address real-world coupled degradation challenges.

Another future work could focus on retaining the origi-

Figure 2. More examples of JarvisIR's perception restoration are presented. Initially, JarvisIR assesses the degradation of the input images and parses user instructions to formulate a task plan, selecting appropriate expert models for each subtask. The selected experts perform their designated tasks and return the results to JarvisIR, which integrates the outcomes and provides the final answer to the user.

Figure 3. Comparison of the decision-making processes of JarvisIR-MRRHF and JarvisIR-SFT. The results indicate that the MRRHF version accurately predicts the correct task sequence and selects appropriate restoration models. Conversely, the SFT version often fails to make suitable decisions in real-world scenarios due to the domain gap between training and real data distributions.

nal image resolution during training. Most current vision-language models (VLMs) resize input images to a fixed resolution, such as 336×336, which may degrade performance, as resolution variation may affect the model's perception of degradation. To mitigate this, future research could explore techniques to maintain original image resolutions. One approach involves adapting the position embeddings in CLIP [21] using bicubic interpolation to accommodate varying image dimensions.

This work focuses on building an autonomous, robust, intelligent restoration system tailored for real-world challenges. To enhance system robustness, reduce hallucinations, and improve generalizability, we introduce a novel two-stage framework that integrates supervised fine-tuning with human feedback alignment. By utilizing human feedback and large-scale real unlabeled data, our method allows the VLM to be fine-tuned in an unsupervised manner. We believe that this paradigm can inspire future work to build more powerful and versatile intelligent systems.

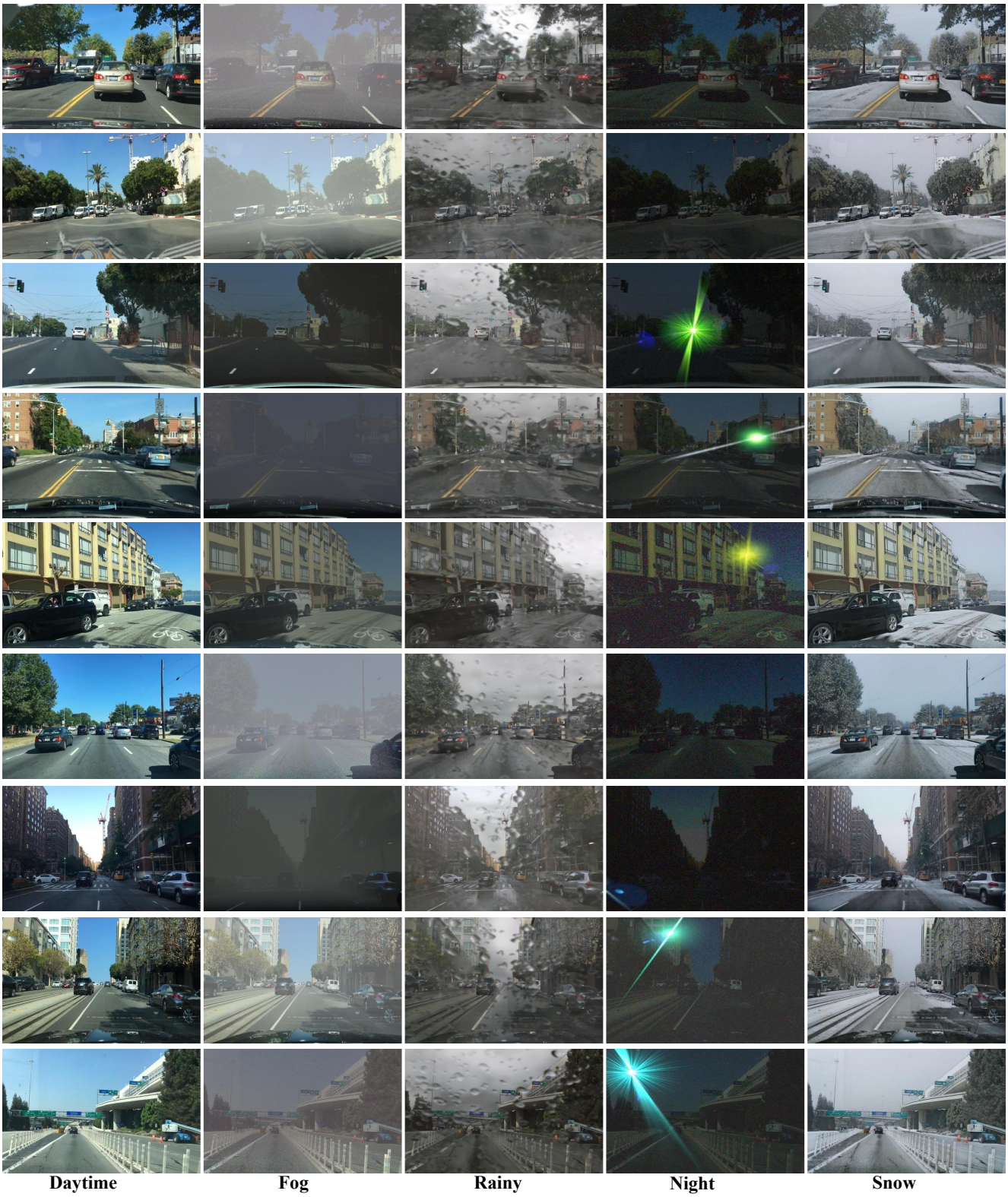

Figure 4. Examples of synthetic adverse weather scenarios in autonomous driving from the CleanBench dataset.

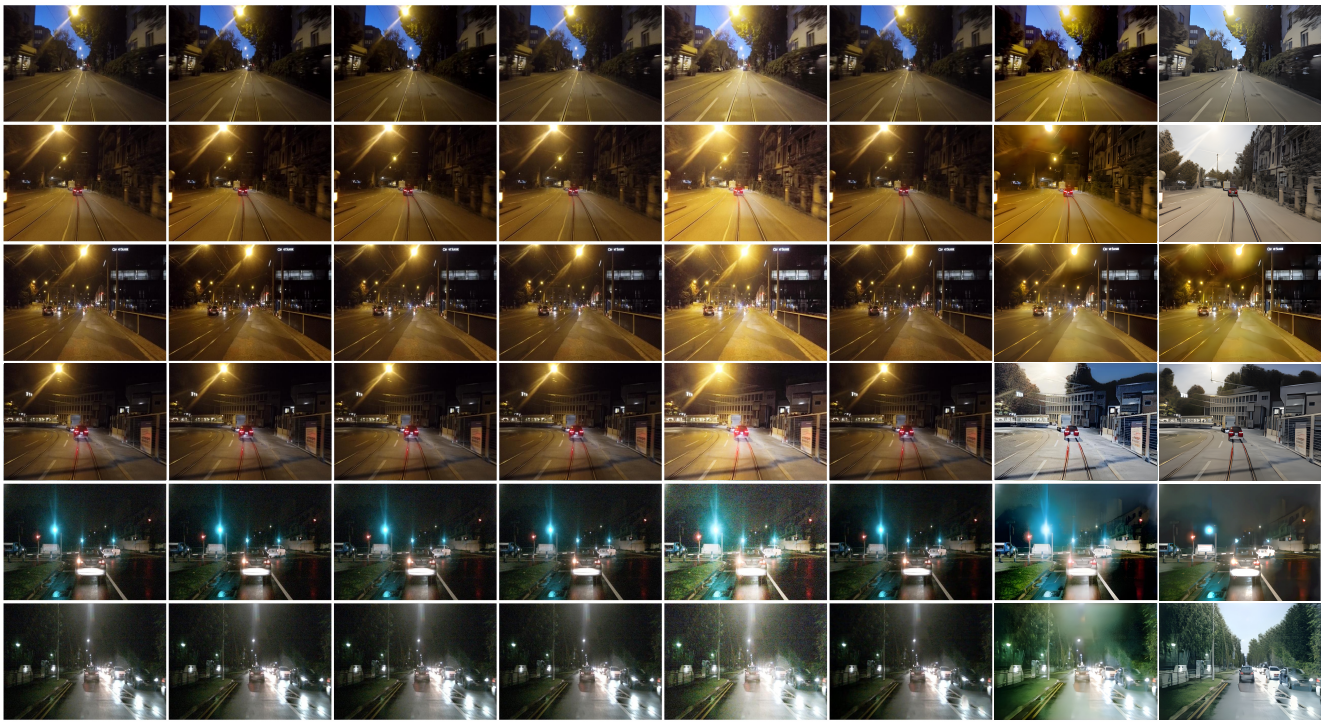

Figure 5. Visual comparisons among various methods on CleanBench-Real's night scene validation set.

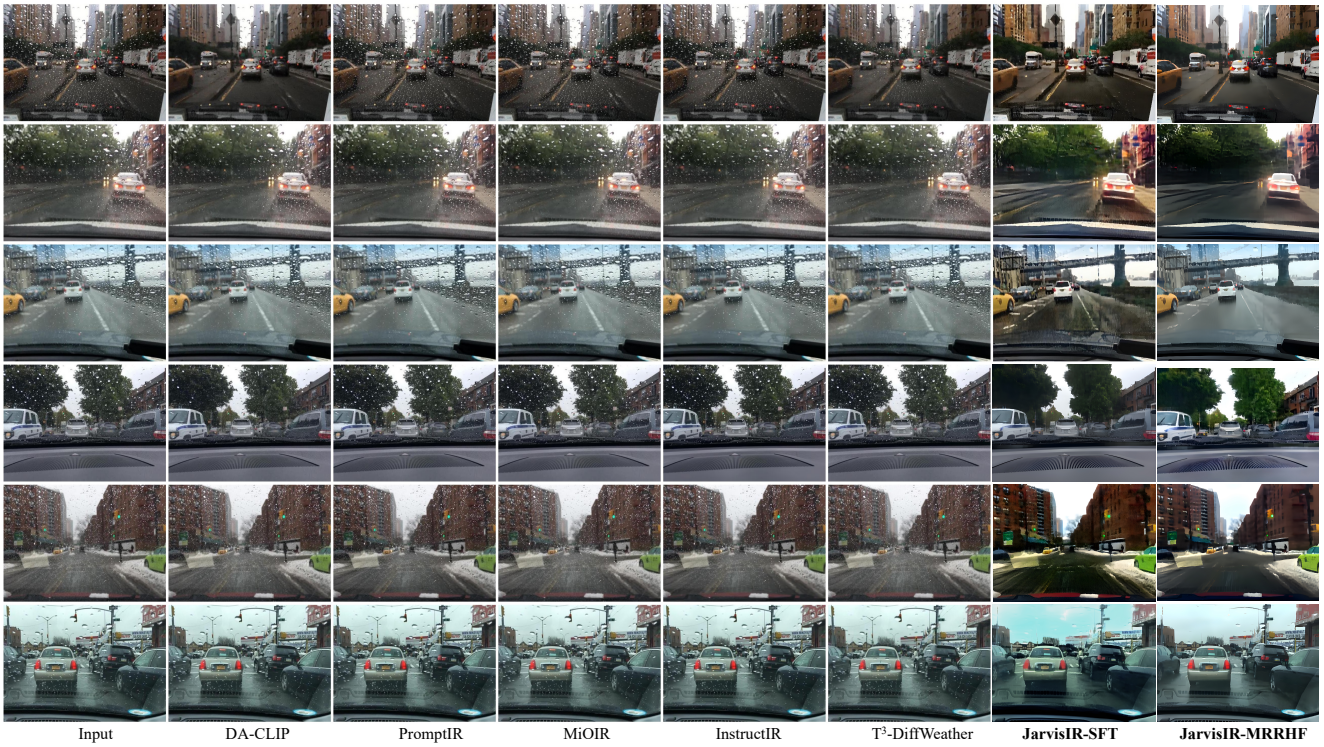

Figure 6. Visual comparisons among various methods on CleanBench-Real's rain scene validation set.

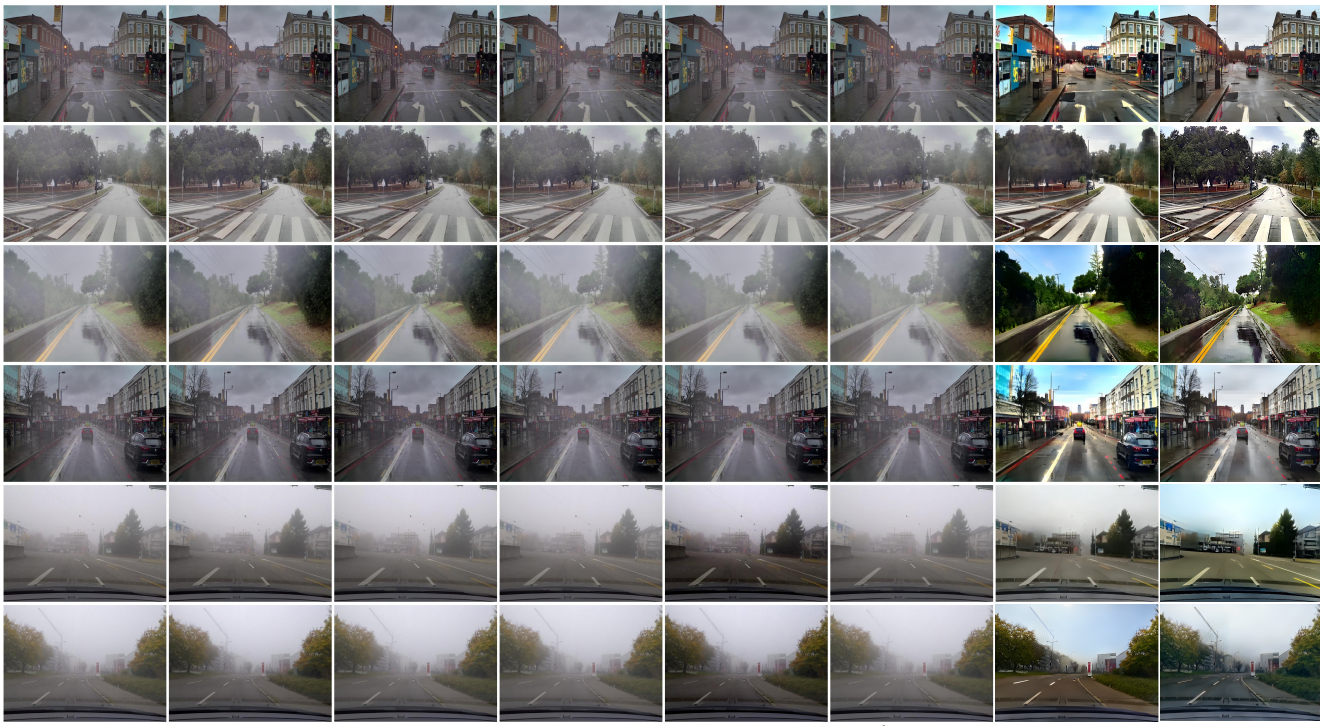

Figure 7. Visual comparisons among various methods on CleanBench-Real's fog scene validation set.

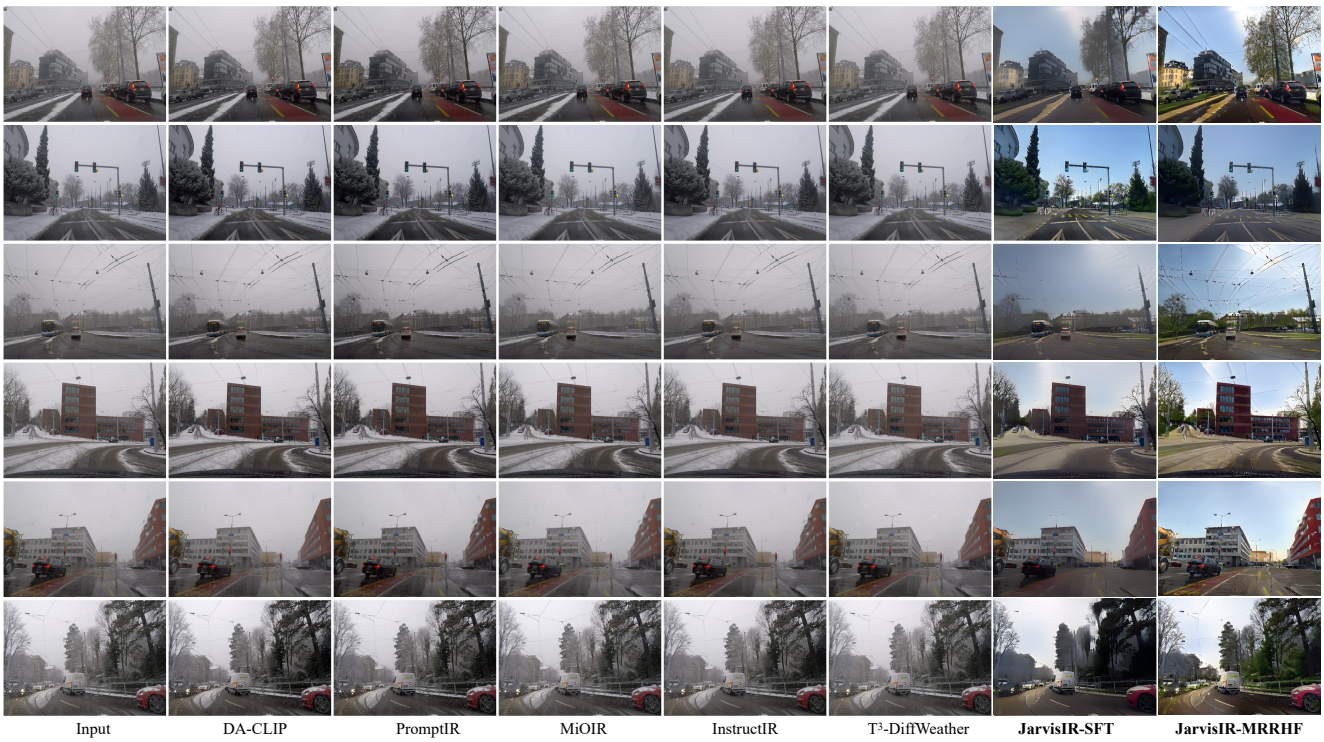

Figure 8. Visual comparisons among various methods on CleanBench-Real's snow scene validation set.

Table 5. Instruction generated by GPT-4V using the self-instruct strategy [28]

#	**Instruction**
1	Please evaluate the image's quality comprehensively and provide your insights. Additionally, outline a step-by-step restoration strategy, specifying the sequence of tasks and model choices for each step. The available tasks are super-resolution, denoising, dehazing,..., low-light enhancement. The available restoration models are StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
2	Analyze the quality of the image comprehensively and provide your insights. Furthermore, propose a restoration strategy by detailing each task and model choice sequentially. The available tasks include super-resolution, denoising, dehazing,..., low-light enhancement. The restoration models are StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
3	Assess the overall quality of the image and provide a detailed evaluation. Then, design a step-by-step restoration process, specifying tasks and model choices. Tasks available are super-resolution, denoising, dehazing,..., low-light enhancement, and models include StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
4	Perform a comprehensive evaluation of the image quality and explain your observations. Additionally, develop a step-by-step restoration plan, identifying tasks and model choices. Available tasks are super-resolution, denoising, dehazing,..., low-light enhancement, and models include StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
5	Conduct a thorough analysis of the image's quality and provide your insights. Subsequently, create a restoration strategy step by step, specifying the tasks and model choices. The tasks available are super-resolution, denoising, dehazing,..., low-light enhancement, and models are StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
6	Evaluate the quality of the image comprehensively and outline your findings. Moreover, formulate a sequential restoration plan, detailing tasks and model selections. Available tasks include super-resolution, denoising, dehazing,..., low-light enhancement, and models are StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
7	Provide a detailed assessment of the image's quality and share your observations. Then, create a restoration strategy in a step-by-step manner, specifying tasks and models. Available tasks are super-resolution, denoising, dehazing,..., low-light enhancement, and models include StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
8	Examine the image comprehensively and describe its quality in detail. Additionally, outline a step-wise restoration process, specifying the tasks and models for each stage. Tasks available are super-resolution, denoising, dehazing,..., low-light enhancement, and models are StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
9	Review the image quality comprehensively and present your findings. Furthermore, propose a restoration plan step by step, detailing the tasks and models used. Available tasks are super-resolution, denoising, dehazing,..., low-light enhancement, and restoration models are StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).
10	Perform a detailed evaluation of the image and provide your insights. Then, construct a step-by-step restoration strategy, specifying the sequence of tasks and the corresponding models. The available tasks include super-resolution, denoising, dehazing,..., low-light enhancement, and models are StableSR-turbo: (model description), SCUnet: (model description),..., RIDCP: (model description).

Table 6. Responses generated by GPT-4V using the self-instruct strategy [28]

#	**Response**
1	The image depicts an outdoor scene viewed through a wet windshield, with water droplets significantly obscuring the view. The evaluated image exhibits severe distortion caused by the water droplets, resulting in a loss of sharpness and clarity. Details of the background, such as the car ahead and surrounding structures, are blurred and difficult to discern. The image also suffers from uneven focus and lacks proper visibility.
2	This image portrays an outdoor scene as seen through a wet windshield, where water droplets heavily obstruct the view. The distortion caused by these droplets leads to a noticeable reduction in sharpness and clarity. Background elements, like the car in front and nearby structures, appear heavily blurred and challenging to distinguish. The image further suffers from uneven focus, reducing overall visibility.
3	An outdoor setting is captured in this image, viewed through a wet windshield. The water droplets significantly obscure the scene, causing severe distortion. Sharpness and clarity are greatly reduced, and background features, such as the car ahead and surrounding buildings, are blurred and hard to recognize. Uneven focus further detracts from the image's overall visibility.
4	This image showcases an outdoor scene as observed through a wet windshield, with water droplets causing significant obstruction. The distortion leads to substantial clarity and sharpness loss. Details in the background, including the car ahead and nearby structures, are blurred and hard to identify. Additionally, the image has uneven focus, which impairs its overall visibility.
5	The image illustrates an outdoor view observed through a wet windshield, where water droplets obscure much of the scene. The distortion caused by the droplets results in a notable reduction in sharpness and clarity. Background details, such as the car ahead and surrounding structures, appear blurry and indistinct. Uneven focus further degrades the visibility of the image.
6	In this image, an outdoor setting is depicted as seen through a wet windshield, with water droplets heavily obscuring the view. The distortion from the droplets results in a considerable loss of sharpness and clarity. Background elements, like the car ahead and nearby buildings, are blurred and hard to discern. The uneven focus of the image adds to its overall lack of visibility.
7	This image shows an outdoor scene viewed through a wet windshield, with water droplets creating significant visual obstruction. The distortion caused by the droplets reduces the sharpness and clarity of the image. Background features, including the car ahead and the surrounding structures, are blurred and difficult to distinguish. The image also suffers from uneven focus, limiting its visibility.
8	An outdoor view is displayed in this image, captured through a wet windshield with water droplets blocking much of the scene. The distortion caused by the droplets results in reduced clarity and sharpness. Background details, such as the car in front and nearby structures, are blurred and indistinct. Uneven focus further compromises the visibility of the image.
9	The image depicts an outdoor scene seen through a wet windshield, where water droplets obscure much of the view. This distortion leads to a significant loss of sharpness and clarity. Background elements like the car ahead and nearby structures are blurred and challenging to distinguish. Uneven focus further contributes to the poor visibility of the image.
10	This image represents an outdoor scene viewed through a wet windshield, with water droplets obscuring the visual field. The distortion caused by the droplets significantly affects sharpness and clarity, making background features like the car ahead and surrounding structures appear blurred and indistinct. The uneven focus further reduces the overall visibility of the image.

References

[1] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11036–11045, 2019. 3

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[3] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018. 1, 2

[4] Haoyu Chen, Jingjing Ren, Jinjin Gu, Hongtao Wu, Xuequan Lu, Haoming Cai, and Lei Zhu. Snow removal in video: A new dataset and a novel method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13211–13222, 2023. 2

[5] Sixiang Chen, Tian Ye, Yun Liu, and Erkang Chen. Snowformer: Context interaction transformer with scale-awareness for single image desnowing. *arXiv preprint arXiv:2208.09703*, 2022. 2

[6] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2553–2562, 2021. 3

[7] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Yihang Luo, and Chen Change Loy. Flare7k++: Mixing synthetic and real datasets for nighttime flare removal and beyond. *arXiv preprint arXiv:2306.04236*, 2023. 3

[8] Yuxin Feng, Long Ma, Xiaozhe Meng, Fan Zhou, Risheng Liu, and Zhuo Su. Advancing real-world image dehazing: perspective, modules, and training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[9] Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. *arXiv preprint arXiv:2407.08939*, 2024. 2

[10] Yeying Jin, Xin Li, Jiadong Wang, Yan Zhang, and Malu Zhang. Raindrop clarity: A dual-focused dataset for day and night raindrop removal. In *European Conference on Computer Vision*, pages 1–17. Springer, 2025. 1, 2

[11] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 4, 5, 6

[12] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 5

[13] T Kudo. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. 1

[14] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 1, 2

[15] Beibei Lin, Yeying Jin, Wending Yan, Wei Ye, Yuan Yuan, and Robby T Tan. Nighthaze: Nighttime image dehazing via self-prior learning. *arXiv preprint arXiv:2403.07408*, 2024. 1, 2

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[17] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6): 3064–3073, 2018. 2

[18] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 1, 2, 4

[19] Fabio Pizzati, Pietro Cerri, and Raoul de Charette. Physics-informed guided disentanglement in generative networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10300–10316, 2023. 4

[20] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9147–9156, 2021. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 7

[22] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[24] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016. 4

[25] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 4

[26] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. 2

[27] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with

pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 2

[28] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 11, 12

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 5

[30] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 5, 6

[31] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25490–25500, 2024. 5, 6

[32] Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22282–22291, 2023. 2, 3

[33] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12978–12995, 2022. 2

[34] Qingsen Yan, Yixu Feng, Cheng Zhang, Pei Wang, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. You only need one color space: An efficient network for low-light image enhancement. *arXiv preprint arXiv:2402.05809*, 2024. 2

[35] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 3

[36] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 4, 5, 6

[37] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 1, 2

[38] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1

[39] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023. 4

[40] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023. 2

[41] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 4

[42] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 5

[43] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, pages 573–589. Springer, 2022. 1, 2

[44] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1