JarvisIR: Elevating Autonomous Driving Perception with Intelligent Image Restoration

Yunlong Lin[1*♣] Zixu Lin[1*♣] Haoyu Chen[2*] Panwang Pan[3*] Chenxin Li[6]
Sixiang Chen[2] Yeying Jin[4] Wenbo Li[5†] Xinghao Ding[1†]
[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, Xiamen, Fujian, China
[2] The Hong Kong University of Science and Technology (Guangzhou)
[3] Bytedance's Pico [4] Tencent [5] Huawei Noah's Ark Lab
[6] The Chinese University of Hong Kong

Project page: https://cvpr2025-jarvisir.github.io/

Abstract

*Vision-centric perception systems struggle with unpredictable and coupled weather degradations in the wild. Current solutions are often limited, as they either depend on specific degradation priors or suffer from significant domain gaps. To enable robust and autonomous operation in real-world conditions, we propose JarvisIR, a VLM-powered agent that leverages the VLM as a controller to manage multiple expert restoration models. To further enhance system robustness, reduce hallucinations, and improve generalizability in real-world adverse weather, JarvisIR employs a novel two-stage framework consisting of supervised fine-tuning and human feedback alignment. Specifically, to address the lack of paired data in real-world scenarios, the human feedback alignment enables the VLM to be fine-tuned effectively on large-scale real-world data in an unsupervised manner. To support the training and evaluation of JarvisIR, we introduce CleanBench, a comprehensive dataset consisting of high-quality and large-scale instruction-responses pairs, including **150K** synthetic entries and **80K** real entries. Extensive experiments demonstrate that JarvisIR exhibits superior decision-making and restoration capabilities. Compared with existing methods, it achieves a **50%** improvement in the average of all perception metrics on CleanBench-Real.*

1. Introduction

Vision-centric perception systems often struggle in adverse weather, where images captured in real-world scenarios exhibit multiple and coupled degradations. Current adverse weather image restoration methods are primarily categorized into task-specific methods and all-in-one approaches. Both categories struggle with real-world coupled degradations, as shown in Figure 1. Task-specific methods [21, 28, 29, 40, 78] often require prior knowledge of specific degradation types, while real-world degradations are often unknown and coupled. All-in-one methods [12, 17, 27, 33, 44] trained on synthetic datasets in a supervised manner, suffer from a significant domain gap when applied to real-world data. One promising strategy to tackle multiple degradations in the wild is to integrate specialized models that excel in their domains. However, this strategy is highly sensitive to task order, and even minor changes in execution sequence can lead to significant performance degradation. Therefore, autonomously and efficiently coordinating expert models in real-world scenarios is essential for perceptual restoration.

Recently, large language models (LLMs) have exhibited remarkable proficiency in reasoning, decision-making and interaction with environments [23, 26, 48, 74, 86]. These advancements raise an important question: *Could vision-language models (VLMs) act as controllers, managing publicly available specialized restoration models, autonomously planning tasks, and selecting models to facilitate the development of comprehensive restoration systems?* The answer is affirmative, however, constructing such systems is non-trivial and typically requires extensive paired data. In real-world scenarios, while there exists extensive real degraded data, the lack of corresponding labels prevents the implementation of supervised fine-tuning approaches. To tackle this issue and harness large-scale unlabeled data, we design a fine-tuning framework based on human feedback, allowing the VLM to be trained in an unsupervised manner. With this approach, we could create a

∗ Authors Yunlong Lin and Zixu Lin contributed the most and led the study, while Authors Haoyu Chen and Panwang Pan also made significant contributions. † Corresponding author.

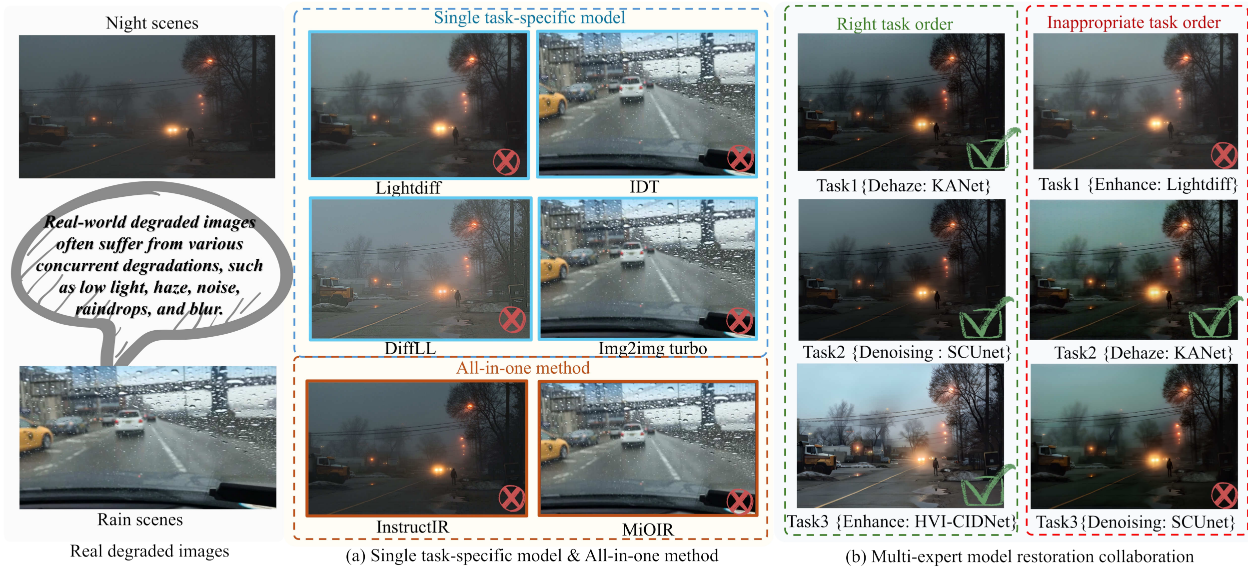

Figure 1. **Limitations of single-task methods, all-in-one methods, and inaccurate task order**. (a) Single-task specific and all-in-one methods fail to address coupled degradation in real-world scenarios. (b) Collaboration among multi-expert models effectively mitigates complex degradation, but is sensitive to the order of tasks. Unlike these approaches, JarvisIR can dynamically schedule different expert models in response to the rapidly changing scenarios and coupled degradation in the wild.

system that performs robustly and reliably in the wild.

In this work, we introduce JarvisIR, a VLM-powered agent integrating VLM (i.e., Llava-Llama3 [41]) with expert restoration models sourced from GitHub and Hugging Face. The development of this system involved two key components: 1) CleanBench, an instruction-following dataset constructed using the self-instruct strategy [65], which includes 150K synthetic and 80K real instruction-response pairs (CleanBench-Real), designed to support both training and evaluation. 2) A supervised fine-tuning (SFT) and human feedback alignment framework for training a VLM as an agent to be reliable and autonomous. Specifically, to enable the VLM to follow user instructions and perceive image degradation, we train it using the synthetic portion of CleanBench via SFT [45]. To enhance system robustness, reduce hallucinations, and improve generalizability in real-world adverse weather, we fine-tune JarvisIR on CleanBench-Real with human feedback. To ensure stability during training and improve overall performance, we propose the MRRHF algorithm, an extension of the ranking responses with human feedback (RRHF) approach [82]. Specifically, to expand the exploration space while maintaining a performance lower bound for JarvisIR, we introduce a hybrid sample generation strategy and regularization term. Furthermore, to comprehensively feedback the quality of system responses during training, we incorporate multiple VLM-based Image Quality Assessment (IQA) models as a unified reward model.

Our contributions can be summarized as follows:

- We introduce JarvisIR, a VLM-powered agent that autonomously manages and coordinates multiple expert restoration models to address coupled weather degradations in real-world environments.
- We present CleanBench, the first high-quality instruction-following dataset specifically curated for developing intelligent restoration systems, containing 150K synthetic and 80K real instruction-response pairs.
- We propose a novel two-stage framework combining supervised fine-tuning and human feedback alignment to enhance system robustness, reduce hallucinations and improve generalizability in the wild.
- Our experiments demonstrate that JarvisIR outperforms strong baselines in terms of decision-making and perception restoration.

2. Related Work

Tool-Augmented LLMs. Recent studies [5, 48, 50, 53–55, 85] highlight the growing potential of large language models (LLMs) for proficient tool usage and decision-making in complex settings. For example, Gorilla [48] facilitates LLMs' response to Tool calls through dataset construction and fine-tuning. ToolLLM [50] extends this concept to enable interaction with a large number of tools. ToolAlpaca [58] demonstrates the feasibility of generalized tool-use capabilities in smaller LLMs. Toolformer [53] constructs tool-use augmented data to train LLMs to select tools. In the realm of visual tools, various approaches have been proposed to enhance the capabilities of large language models in handling visual tasks [69, 77], augmented with Hugging Face models [54], Azure models [77], visual foundation models [69].

Alignment of LLMs. Reinforcement Learning from Human Feedback (RLHF) [1, 2, 25, 59] has emerged as a groundbreaking technique for aligning LLMs. The core idea is learning a reward function to reflect human preferences with human annotations and optimize LLMs by RL methods like proximal policy optimization (PPO). During PPO-based optimization, updating LLMs requires the likelihood of an entire generation. However, for LLM agents, human feedback is usually obtained only after the tool response is completed and the function is successfully invoked. Moreover, unlike typical LLM training, our two-stage fine-tuning process integrates both visual and linguistic modalities. Rank Responses to Align Human Feedback (RRHF) [82] has shown promise by using reward models to rank multiple responses, aligning LLMs effectively. This technique allows easy extension to fine-grained tool agents, thereby maximizing the utility of existing reward models.

Image Restoration. Single-task image restoration has achieved significant progress in addressing specific degradation types, such as dehazing [28, 37, 71], low-light enhancement [24, 30, 39], desnowing [7, 13, 16], deraining [11, 15, 72], denoising [6, 83] super-resolution [9, 57, 62, 64], image fusion [19, 20, 38, 66]. However, these task-specific approaches often lack generalizability and adaptability to complex, coupled degradations. To overcome this limitation, adverse weather restoration research aims to develop a unified framework capable of addressing multiple degradation types simultaneously [14, 22, 36, 46]. Another prevailing research line is dedicated to building more intelligent restoration systems. Clarity ChatGPT [68], integrated with advanced visual models, allows users to perform sophisticated image manipulation and enhancement through natural language interactions. RestoreAgent [8] and AgenticIR [88] are contemporaneous independent works that utilize MLLM as a task planner to coordinate multiple restoration tools. Specifically, RestoreAgent [8] involves fine-tuning a vision-language model (VLM) using synthetic datasets to directly produce an execution plan. AgenticIR [88] leverages two off-the-shelf LLMs and VLMs to achieve the scheduling of restoration tools on the synthetic experiment platform. Essentially, both studies focus on building intelligent restoration systems tailored for synthetic degradation conditions. *Conversely, our study aims to develop a robust system for real-world applications, incorporating human feedback to enhance robustness, reduce hallucinations and improve generalizability. Furthermore, our system is built in an unsupervised manner using large-scale, unlabeled real-world data.*

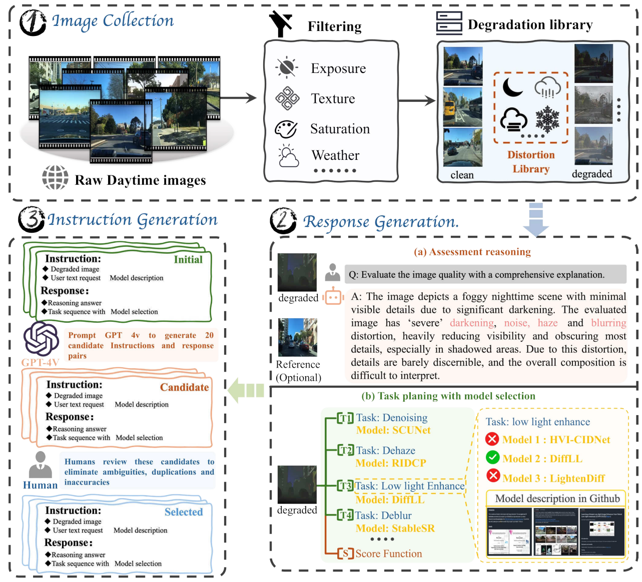

Figure 2. The dataset construction workflow consists of three main steps: 1) Synthesis of degraded images. 2) Generation of Assessment reasoning and the optimal task sequence. 3) Generation of instruction-response pairs for the system.

3. Methodology

In this section, we first describe CleanBench, a comprehensive benchmark consisting of extensive instruction-response pairs used for the training and evaluation of JarvisIR (Sec. 3.1). We then introduce JarvisIR, a VLM agent to call expert restoration models in response to intricate multiple degraded environments in the wild (Sec. 3.2). Finally, we describe the two-stage training framework for JarvisIR, comprising supervised fine-tuning and human feedback alignment.

3.1. CleanBench

High-quality and large-scale datasets are crucial for unleashing the full potential of VLMs. A multimodal instruction sample can be formally represented as a triplet: {*user instruction, degraded image, response*}, where "*user instruction*" specifies the task and describes the restoration tools, "*degraded image*" serves as the visual input to be processed, and the "*response*" provides the ground truth answer. In Figure 2, we outline the construction of our dataset, focusing on the generation of degraded images and the collection of task-specific instructions and responses.

Image Collection. We first collect raw daytime images from various sources, including autonomous driving datasets [3, 52, 81, 89] and natural scenes [4, 31, 34, 37, 42, 76, 87]. Then, Q-instruct [70] serves as a quality filter to extract high-quality samples. 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

Figure 3. Examples of CleanBench-Real dataset.

synthesize degraded images. More detail in supplementary material.

Response Generation. The response from JarvisIR consists of two components: "chain-of-thought" (COT) rationales and the optimal task sequence with model selection. (a) For COT rationales, we distill DepictQA-Wild's [80] knowledge, which excels in low-level quality reasoning assessment. Specifically, given a degraded image pair, we prompt DepictQA-Wild [80] to assess the quality of the degraded image in terms of clarity, colorfulness, and sharpness, generating detailed degradation and reasoning insights. (b) To determine the optimal task sequence with restoration model selection, we employ an exhaustive search strategy [8] to explore various task permutations and model combinations, scoring each sequence to identify the optimal restoration path.

Task-model Assignment. User instructions include descriptions of available tasks and models, sourced from GitHub or Hugging Face, to formulate task-model assignment as a single-choice problem. Presenting tasks and models as options within a context allows JarvisIR to more effectively identify the appropriate model for each sub-task.

Instruction Generation. Motivated by the self-instruct strategy [65], for each initial user instruction and response, GPT-4V is prompted to generate 20 candidate pairs. We then manually review these candidates to eliminate ambiguity, repetition, and inaccuracies, ultimately selecting 5 instruction-response pairs per degraded image (see supplementary material for details). Ultimately, CleanBench includes a total of *150K* instruction-response pairs, which are used in the initial instruction-tuning phase.

CleanBench-Real. To align and evaluate JarvisIR's performance in real-world scenarios, we introduce CleanBench-Real, comprising *80K* unlabeled real degraded images from internet and diverse sources [3, 30, 31, 42, 51, 63, 76, 81]. CleanBench-Real is categorized into four adverse weather scenarios: rainy, night, snowy, and foggy. The degradation in each scenario is complex and interwined. For example, as presented in Figure 3, an image captured in rain may experience multiple degradations concurrently, including rain, raindrops, defocus blur, and noise (more in supplementary material). For the division of the training and evaluation sets, we selected 500 images from each of the four CleanBench-Real scenarios to form the evaluation set (*2K*), while the remaining images are utilized for alignment tuning. Instruction-response pairs are generated in the same way as outlined in CleanBench.

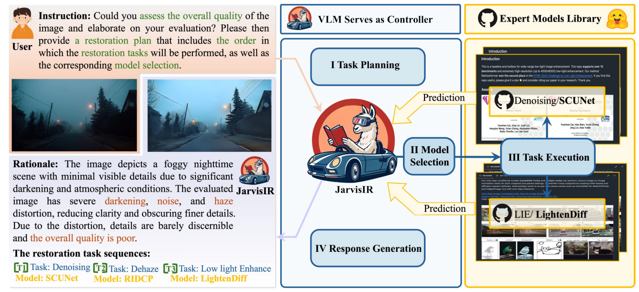

Figure 4. The workflow of JarvisIR. To address real-world coupled weather degradation, we develop JarvisIR, a VLM-powered intelligent system that dynamically schedules expert models for restoration. Initially, JarvisIR assesses the degradation of the input images and parses user instructions to formulate a task plan, selecting the 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. The design of the figure is inspired by [54].

3.2. JarvisIR

JarvisIR is a VLM-powered agent that coordinates multiple expert restoration models to address complex degradation. As illustrated in Figure 4, the workflow of JarvisIR consists of four steps: Task Planning, Model Selection, Task Execution, and Response Generation. To enhance the agent's decision-making and perception restoration capabilities in real-world scenarios, as depicted in Figure 5, we initially perform supervised fine-tuning (SFT) on CleanBench to obtain an initial version, termed JarvisIR-SFT. Subsequently, the JarvisIR-SFT is further fine-tuned utilizing the MRRHF algorithm on CleanBench-Real, yielding the JarvisIR-MRRHF model.

3.2.1. JarvisIR-SFT

We employ the standard SFT to get the JarvisIR-SFT model. Formally, the multimodal instruction sample can be denoted in a triplet form $(\mathcal{I}, \mathcal{M}, \mathcal{R})$, where $\mathcal{I}$, $\mathcal{M}$, $\mathcal{R}$ represent the user instruction, the degraded image, and the ground truth response, respectively. The VLM predicts an answer $\mathcal{A}$ given the instruction and the degraded image: $\mathcal{A} = f(\mathcal{I}, \mathcal{M}; \theta)$. The training objective is the original auto-regressive objective used to train LLMs [43, 79]:

$$L_{sft} = -\sum_{i=1}^{N} \log P_{\pi} \left(\mathcal{R}_i \mid \{\mathcal{I}_i, \mathcal{M}_i\}, \mathcal{R}_{<i}; \theta \right), \quad (1)$$

where N is the length of the ground-truth response.

3.2.2. JarvisIR-MRRHF

Intuitively, SFT allows JarvisIR-SFT to achieve favorable performance on synthetic data. Nevertheless, as previously noted, due to the distribution shift, transferring from synthetic training data to real test data, JarvisIR-SFT exhibits increased hallucination, i.e., degraded perception restoration performance and decision-making capability. To improve its generalizability, we further fine-tune JarvisIR on CleanBench-Real with refined ranking responses with human feedback algorithm (MRRHF).

Reward modeling. The reward model evaluates tool-calling outcomes and converts them into structured reward signals to guide the agent's optimization process. Therefore, selecting an appropriate reward model is crucial. Fortunately, in the image quality assessment (IQA) field, VLM-based IQA models have been developed [70], demonstrating strong performance in evaluating aesthetic quality and image distortion. These IQA models are inherently suitable for serving as reward models. To construct a comprehensive reward model $\mathcal{S}$, as well as an evaluation system, we integrated multiple IQA models. Specifically, we employ a z-score strategy [8] to standardize the scores assessed by each IQA model separately and then sum the standardized results:

$$\mathcal{S} = \sum_{i=1}^{k} \frac{s_i - \mu_i}{\sigma_i}, \tag{2}$$

where s_i represents the score assessed by i-th IQA model. μ_i and σ_i epresent the mean and standard deviation of s_i, respectively. k indicates the total number of IQA models.

Alignment with MRRHF. We propose an extension to the existing RRHF method that can be used for aligning JarvisIR in a cost-effective manner: 1) A hybrid sample generation strategy that combines offline and online approaches to expand the optimization exploration space while ensuring a performance lower bound. 2) Entropy regularization terms are integrated to foster diversity among agent responses, thereby facilitating exploration during training. Specifically, for a pair of user instruction $\mathcal{I}_i$ and degraded image $\mathcal{M}_i$, we first adopt offline diverse beam search [60] to get m_1 different responses $\mathcal{R}_{m_1} = \{r_1, r_2, \ldots, r_{m_1}\}$ from SFT model π. Similarly, we can obtain $\mathcal{R}_{m_2} = \{r_1, r_2, \ldots, r_{m_2}\}$ from policy model ρ (initialized from SFT model π) during training. The combined candidate m responses are denoted as $\mathcal{R}_m = \mathcal{R}_{m_1} \cup \mathcal{R}_{m_2}$. Subsequently, we execute the task sequences specified in candidate responses, calling multiple restoration models to generate restored images. These predictions are then assessed by the reward model $\mathcal{S}$, yielding scores for each r_i with $\mathcal{S}(r_i) = s_i$. To align with scores $\{s_i\}_m$, we use policy model ρ to give scores p_i for each r_i by:

$$p_i = \frac{\sum_t \log P_\rho \left(r_{i,t} \mid \{\mathcal{I}_i, \mathcal{M}_i\}, r_{i,<t}; \theta\right)}{\|r_i\|}, \tag{3}$$

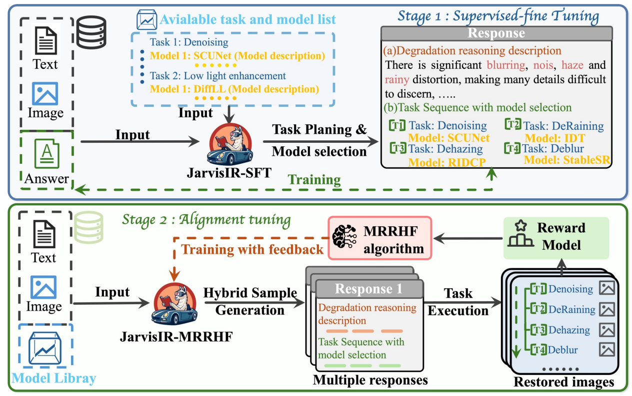

Figure 5. Two-stage training framework of JarvisIR. In the first stage, JarvisIR undergoes supervised fine-tuning on synthetic data from CleanBench to enable it to follow user instructions and recognize image degradation. In the second stage, we further fine-tune JarvisIR on CleanBench-Real using the MRRHF algorithm to improve system robustness, reduce hallucinations, and enhance generalizability under real-world adverse weather conditions.

where p_i is conditional log probability (length-normalized) of r_i under model ρ. The core idea is letting the policy model ρ give larger probabilities for better responses and give smaller probabilities for worse responses. Inspired by PRO [56], we refine the original ranking loss:

$$L_{\text{rank}} = \sum_{s_i < s_j} (s_j - s_i) \max(0, p_i - p_j), \tag{4}$$

and a cross-entropy loss like SFT process is added to learn the response with the highest reward s_i, $i' = \arg\max_i s_i$:

$$L_{ft} = -\sum_t \log P_\rho \left(r_{i',t} \mid \{\mathcal{I}_i, \mathcal{M}_i\}, r_{i',<t}; \theta\right). \tag{5}$$

Furthermore, we define the entropy regularization term as:

$$L_{er} = -\sum_a \rho(a \mid y) \log \rho(a \mid y), \tag{6}$$

where y represents the current state of the agent. The overall loss is utilized to optimize the JarvisIR-SFT to derive JarvisIR-MRRHF:

$$L = \lambda_1 L_{rank} + \lambda_2 L_{ft} + \lambda_3 L_{er}, \tag{7}$$

where λ_1, λ_2 and λ_3 are constants controlling the relative importance of the different losses, which are empirically set to 0.5, 0.5 and 0.1 in all experiments, respectively.

Discussion of RLHF and RRHF: The training of vanilla RLHF [45] necessitated a policy model, a value model, a reward model, and a reference model, which could be demanding on memory resources. Rank Responses to Align Human Feedback (RRHF) [82] can effectively alleviate

Figure 6. Visual comparisons of various methods on CleanBench-Real. Our approach delivers significant quality improvements, eliminating complex real-world degradation and preserving the most natural details.

the issues of resource-intensive and tedious hyperparameter tuning in RLHF. However, directly fine-tuning JarvisIR using RRHF yields limited improvement to its generalization in real-world scenarios. Although vanilla RRHF employs an off-policy learning strategy that could save time by avoiding the need to generate new responses during training, it has the drawback of relying on a static offline preference dataset for training the policy model. Consequently, the policy might over-optimize for reward on in-distribution data as the model cannot further query the preference oracle during the training process [67]. The RRHF incorporating online sampling like PPO might mitigate this issue, but it demands more GPU resources to store the reference model, thereby significantly decreasing the training speed [82].

Table 1. Comparison of JarvisIR with other strategies on the CleanBench-Real validation set. The "Score" represents the sum of the four normalized metrics. The "Ranking" indicates the given decision's percentage ranking among all possible decisions. We highlight the best and second-best results.

Strategy	Score	Ranking(%)
(I) Random Order and Model	1.12	43.2%
(II) Random Order + Predict Model	2.66	34.7%
(III) Random Model + Predict Order	3.08	23.4%
(IV) Pre-defined Order and Model	3.94	22.5%
(V) Human Expert	4.85	18.6%
★**JarvisIR-SFT**	5.17	14.3%
★**JarvisIR-MRRHF**	6.21	4.8%

4. Experiments

4.1. Experimental Settings

Training Setup. Llava-Llama3-8b [41] serves as the base model for JarvisIR, which undergoes full parameter fine-tuning using the Adam optimizer. During the SFT phase, we fine-tune JarvisIR for 3 epochs with a batch size of 128 and a learning rate of 1e-5. In the MRRHF tuning phase, we set the diverse beam search size to 3, the diverse beam group to 5, the diversity penalty to 2.0, and the sampling temperature to 0.8. Alignment tuning is performed over 3 epochs with a batch size of 1 and a learning rate of 1e-5. To speed up training, we select three IQA models—Q-instruct [70], MUSIQ [32] and MANIQA [75]—to construct the unifed reward model (Eq. 2). All experiments are conducted on 8 NVIDIA A100 80G GPUs.

Dataset Setings & Metrics. The CleanBench is fully utilized for supervised fine-tuning of Llava-Llama3-8b [41] to obtain JarvisIR-SFT. The training set of CleanBench-Real is used for alignment tuning, yielding JarvisIR-MRRHF. Additionally, JarvisIR's evaluation is conducted on the validation set of CleanBench-Real, focusing on 1) decision-making ability and 2) perception restoration capability in real-world scenarios. Due to the lack of paired clean-degraded data in the real scenarios. Four image quality assessment metrics are used for evaluation: MUSIQ [32], MANIQA [75], CLIP-IQA+ [61], LIQE [84].

Tool Settings. We present the task-specific restoration tools employed in our implementation, including denoising (SCUnet [83]), super-resolution & deblur & compression artifact removal (StableSR-turbo [62] and Real-ESRGAN [64]), deraining (IDT [72], UDR-S2Former [7] and Img2img-turbo [47]), dehazing (RIDCP [71] and KANet [18]), low-light enhancement (Img2img-turbo [47], HVI-CIDNet [73] and LightenDiff [24]) and desnowing (Img2img-turbo [47] and Snowformer [10]). More details are in the supplementary material. Notably, we select lightweight and efficient models instead of the latest state-of-the-art models to simplify the validation process of our proposed paradigm. Incorporating more advanced models could further enhance performance.

4.2. Decision Making Capability

Compared Baselines. We conducted a comparative analysis of JarvisIR against several alternative approaches: (I) Random selection of both the task order and the models, as-

Table 2. Comparison of JarvisIR with All-in-One methods for multi-degraded perception restoration on CleanBench-Real. We highlight the best, second-best and third-best results. Notably, all scenes represent multiple degraded weather conditions, such as haze, low light and blur.

Method	Night Scenes				Rain Scenes			
	MUSIQ ↑	MANIQA ↑	CLIP-IQA+ ↑	LIQE ↑	MUSIQ ↑	MANIQA ↑	CLIP-IQA+ ↑	LIQE ↑
AirNet [35]	44.26	0.1889	0.4429	1.313	62.61	0.3871	0.5867	3.136
AutoDIR [27]	47.30	0.1885	0.4341	1.403	63.93	0.4002	0.6082	3.312
DA-CLIP [44]	45.86	0.2010	0.4544	1.427	63.28	0.3993	0.5959	3.194
PromptIR [49]	45.45	0.2010	0.4473	1.408	62.85	0.3926	0.5941	3.161
MiOIR [33]	46.93	0.2013	0.4403	1.408	63.07	0.3779	0.5841	3.055
InstructIR [17]	44.03	0.1533	0.3689	1.257	62.93	0.3657	0.5609	3.055
T^3-DiffWeather [12]	46.79	0.1964	0.4547	1.413	62.67	0.3689	0.5823	3.011
⋆**JarvisIR-SFT**	60.77	0.5048	0.5239	3.224	65.03	0.5339	0.6290	4.005
⋆**JarvisIR-MRRHF**	67.25	0.5876	0.6336	3.613	70.38	0.7004	0.7127	4.435

Method	Fog Scenes				Snow Scenes			
	MUSIQ ↑	MANIQA ↑	CLIP-IQA+ ↑	LIQE ↑	MUSIQ ↑	MANIQA ↑	CLIP-IQA+ ↑	LIQE ↑
AirNet [35]	64.23	0.3829	0.6173	2.686	67.32	0.4320	0.6379	3.794
AutoDIR [27]	64.84	0.3966	0.6443	2.928	67.62	0.4305	0.6453	3.824
DA-CLIP [44]	64.78	0.3880	0.6540	2.793	67.71	0.4294	0.6426	3.817
PromptIR [49]	64.54	0.3810	0.6417	2.557	67.34	0.4292	0.6435	3.776
MiOIR [33]	64.93	0.3501	0.5969	2.415	67.28	0.4187	0.6404	3.702
InstructIR [17]	64.82	0.3904	0.6449	2.919	67.98	0.4038	0.6052	3.715
T^3-DiffWeather [12]	64.58	0.3715	0.6163	2.497	67.72	0.4129	0.6268	3.713
⋆**JarvisIR-SFT**	70.45	0.4855	0.6560	3.977	70.24	0.7133	0.7127	4.086
⋆**JarvisIR-MRRHF**	74.22	0.7502	0.7805	4.714	73.87	0.8014	0.7918	4.881

suming that task types are accurately determined. (II) Random task order, but models predicted by JarvisIR. (III) Random model selection, but task orders predicted by JarvisIR. (IV) Using a human expert's predefined order and models for different scenes, assuming the approximate scene degradation can be determined. (V) A human expert manually generates a solution case by case for each image, determining both the task sequence and the appropriate models.

Results. As indicated in Table 1, strategies that involve human expert participation—specifically settings (IV) and (V)—demonstrate strong performance compared to random strategies, ranking within the top 22.5% and 18.6% of all possible strategies, respectively. These results indicate the effectiveness of human experts' experience in complex decision-making processes. Interestingly, however, our JarvisIR model achieves the highest performance, surpassing even the expert-driven customization strategies. Furthermore, JarvisIR-MRRHF (4.8%) outperforms JarvisIR-SFT (14.3%) in both score and ranking, highlighting that the MRRHF stage in our training framework effectively mitigates hallucination errors in system responses, thereby enabling the generation of more optimal decisions.

4.3. Perception Restoration Ability

Compared All-in-One Methods. We compare JarvisIR with existing advanced all-in-one methods: AirNet [35], AutoDIR [27], DA-CLIP [44], PromptIR [49], MiOIR [33], InstructIR [17], T^3-DiffWeather [12]. For a fair comparison, we repeatedly run these compared methods multiple times to fully leverage their capabilities. Additionally, we supply InstructIR and AutoDIR with explicit prompts detailing degradation scenarios to optimize their performance.

Results. As shown in Table 2 and Figure 6, JarvisIR outperforms existing All-in-One approaches across all metrics. In Night Scenes, JarvisIR-MRRHF achieves a MUSIQ score of 67.25, which is 42.2% higher than AutoDIR's score of 47.30. In MANIQA, JarvisIR-MRRHF scores 0.5876, much better than DA-CLIP (0.2010) and MiOIR (0.2013). These results show that JarvisIR autonomously selects optimal task sequences and models, outperforming methods with predefined or random sequences. Additionally, JarvisIR-MRRHF also exceeds the SFT version in all scenes, with notable gains in Rain (70.38 vs. 65.03 MUSIQ) and Fog (74.22 vs. 70.45 MUSIQ). These results demonstrate that JarvisIR fine-tuned with MRRHF can improve generalizability, fewer hallucination errors, and better decision-making ability.

5. Ablation Study

Sample generation strategy. To assess the effectiveness of the hybrid sample generation strategy, we compared it with two variations of the original setting: 1) offline sample gen-

Table 3. Ablation studies on different sample generation strategies and entropy regularization. The "Reward" represents the average reward scores obtained during MRRHF training, 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.

Strategy	Reward	Diversity
offline sample generation	0.43	3.63
online sample generation	-0.87	1.27
hybrid sample generation (ours)	**0.67**	**6.55**
w/o. entropy regularization	0.50	4.56
w. entropy regularization (ours)	**0.67**	**6.55**

eration strategy. 2) online sample generation strategy. The results in Table 3 and Figure 7 yield the following observations: 1) The offline sample generation strategy yields limited performance gains, with a reward score of 0.43 and a diversity score of 3.63. This limitation arises because the sample distribution is restricted to the finite dataset generated by the SFT model using diverse beam search [60]. Consequently, the policy model may over-optimize for in-distribution data, thereby limiting its ability to generalize and achieve higher reward scores. 2) The online sampling strategy initially yields higher reward scores and diversity. However, as training progresses, the model encounters a collapse, leading to a significantly low reward score (-0.87) and decreased diversity (1.27). This instability may result from an excessively large optimization space without adequate constraints during training. When the model reaches a local minimum, it struggles to escape, as the candidate responses generated using diverse beam search [60] are of poor quality, causing the model to produce repetitive and invalid responses. Our hybrid sampling approach combines both online and offline samples, resulting in superior performance with a reward score of 0.67 and the highest diversity score of 6.55. This balanced strategy leverages the advantages of both online and offline sampling, ensuring stable training by providing sufficient exploration space while avoiding the pitfalls associated with purely online sampling. As a result, the hybrid strategy maintains high reward scores and diversity throughout training, outperforming both online and offline strategies.

Entropy regularization. As discussed in Sec. 3.2.2, entropy regularization significantly affects the diversity of system responses during training. The results in Table 3 and Figure 7 show that without this regularization, the reward decreases from 0.67 to 0.50, while the diversity drops from 6.55 to 4.56. This highlights the role of entropy regularization in fostering greater exploration and producing more diverse, high-quality responses.

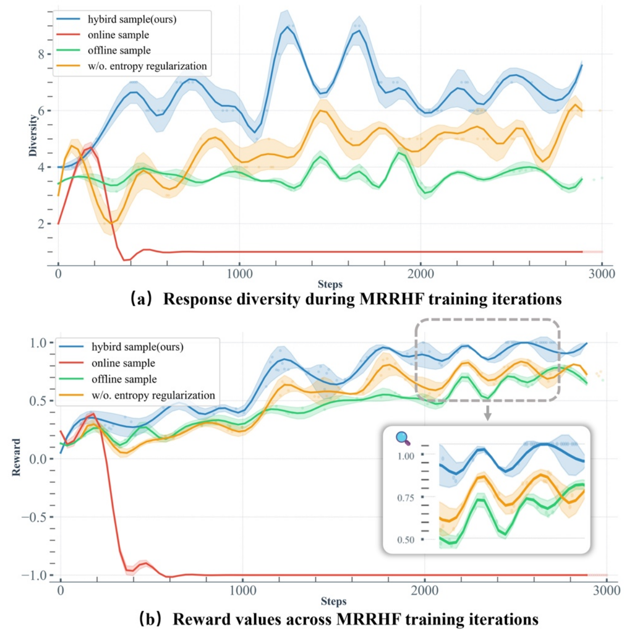

Figure 7. Ablation studies on different sample generation strategies and entropy regularization. (a) Response diversity during MRRHF training iterations. (b) Reward values across MRRHF training iterations.

6. Conclusions

This paper introduces JarvisIR, a VLM-powered intelligent system that leverages Llava-Llama3 to connect distinct restoration expert models. JarvisIR can autonomously schedule different expert models in response to the rapidly changing scenarios and coupled degradation in autonomous driving and natural environments. To enhance system robustness, minimize hallucinations, and improve generalizability, we propose a novel two-stage framework comprising supervised fine-tuning and human feedback alignment. Specifically, we design the human feedback alignment to effectively tune the VLM in an unsupervised manner, leveraging large-scale unlabeled real-world data. To support the training and evaluation of JarvisIR, we present CleanBench, a high-quality, large-scale dataset containing 150K synthetic and 80K real instruction-response pairs. Experiments show that JarvisIR outperforms existing methods, achieving a 50% improvement in the average of all perception metrics on CleanBench-Real.

7. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 82172033, Grant U19B2031, Grant 61971369, Grant 52105126, Grant 82272071, and Grant 62271430; and in part by the Dreams Foundation of Jianghuai Advance Technology Center; and in part by the Open Fund of the National Key Laboratory of Infrared Detection Technologies.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1, 2024. 3

[3] 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. 3, 4

[4] 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. 3

[5] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 2

[6] Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1703, 2023. 3

[7] 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. 3, 6

[8] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. *arXiv preprint arXiv:2407.18035*, 2024. 3, 4, 5

[9] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Haoze Sun, Xueyi Zou, Zhensong Zhang, Youliang Yan, and Lei Zhu. Low-res leads the way: Improving generalization for super-resolution by self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25857–25867, 2024. 3

[10] 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. 6

[11] Sixiang Chen, Tian Ye, Jinbin Bai, Erkang Chen, Jun Shi, and Lei Zhu. Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13106–13117, 2023. 3

[12] Sixiang Chen, Tian Ye, Kai Zhang, Zhaohu Xing, Yunlong Lin, and Lei Zhu. Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint. In *European Conference on Computer Vision*, pages 95–115. Springer, 2025. 1, 7

[13] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4196–4205, 2021. 3

[14] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17653–17662, 2022. 3

[15] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5896–5905, 2023. 3

[16] Bodong Cheng, Juncheng Li, Ying Chen, and Tieyong Zeng. Snow mask guided adaptive residual network for image snow removal. *Computer Vision and Image Understanding*, 236: 103819, 2023. 3

[17] Marcos V Conde, Gregor Geigle, and Radu Timofte. High-quality image restoration following human instructions. *arXiv preprint arXiv:2401.16468*, 2024. 1, 7

[18] 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. 6

[19] Chunming He, Kai Li, Guoxia Xu, Jiangpeng Yan, Longxiang Tang, Yulun Zhang, Yaowei Wang, and Xiu Li. Hqg-net: Unpaired medical image enhancement with high-quality guidance. *TNNLS*, 2023. 3

[20] Chunming He, Kai Li, Guoxia Xu, Yulun Zhang, Runze Hu, Zhenhua Guo, and Xiu Li. Degradation-resistant unfolding network for heterogeneous image fusion. In *ICCV*, pages 12611–12621, 2023. 3

[21] Chunming He, Chengyu Fang, Yulun Zhang, Kai Li, Longxiang Tang, Chenyu You, Fengyang Xiao, Zhenhua Guo, and Xiu Li. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. *ICLR*, 2025. 1

[22] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *TPAMI*, 2025. 3

[23] Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer, 2022. 1

[24] 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. 3, 6

[25] Songtao Jiang, Yan Zhang, Ruizhe Chen, Yeying Jin, and Zuozhu Liu. Modality-fair preference optimization for trust-

worthy mllm alignment. *arXiv preprint arXiv:2410.15334*, 2024. 3
[26] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3843–3860, 2024. 1
[27] Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. *arXiv preprint arXiv:2310.10123*, 2023. 1, 7
[28] Yeying Jin, Wending Yan, Wenhan Yang, and Robby T Tan. Structure representation network and uncertainty feedback learning for dense non-uniform fog removal. In *Proceedings of the Asian Conference on Computer Vision*, pages 2041–2058, 2022. 1, 3
[29] Yeying Jin, Wenhan Yang, and Robby T Tan. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 1
[30] Yeying Jin, Beibei Lin, Wending Yan, Yuan Yuan, Wei Ye, and Robby T Tan. Enhancing visibility in nighttime haze images using guided apsf and gradient adaptive convolution. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2446–2457, 2023. 3, 4
[31] 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. 3, 4
[32] 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. 6
[33] Xiangtao Kong, Chao Dong, and Lei Zhang. Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy. *arXiv preprint arXiv:2401.03379*, 2024. 1, 7
[34] 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. 3
[35] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17452–17462, 2022. 7
[36] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3175–3185, 2020. 3
[37] 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. 3
[38] Yunlong Lin, Zhenqi Fu, Ge Meng, Yingying Wang, Yuhang Dong, Linyu Fan, Hedeng Yu, and Xinghao Ding. Domain-irrelevant feature learning for generalizable pan-sharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3287–3296, 2023. 3
[39] Yunlong Lin, Zhenqi Fu, Kairun Wen, Tian Ye, Sixiang Chen, Ge Meng, Yingying Wang, Yue Huang, Xiaotong Tu, and Xinghao Ding. Unsupervised low-light image enhancement with lookup tables and diffusion priors. *arXiv preprint arXiv:2409.18899*, 2024. 3
[40] Yunlong Lin, Tian Ye, Sixiang Chen, Zhenqi Fu, Yingying Wang, Wenhao Chai, Zhaohu Xing, Lei Zhu, and Xinghao Ding. Aglldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement. *arXiv preprint arXiv:2407.14900*, 2024. 1
[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 6
[42] 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. 3, 4
[43] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 4
[44] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for multi-task image restoration. *arXiv preprint arXiv:2310.01018*, 2023. 1, 7
[45] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2, 5
[46] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10346–10357, 2023. 3
[47] 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. 6
[48] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023. 1, 2
[49] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36, 2024. 7
[50] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023. 2
[51] 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. 4
[52] 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. 3

[53] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[54] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4

[55] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 2

[56] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18990–18998, 2024. 5

[57] Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25868–25878, 2024. 3

[58] Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023. 2

[59] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[60] 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. 5, 8

[61] 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. 6

[62] 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. 3, 6

[63] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013. 4

[64] 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. 3, 6

[65] 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. 2, 4

[66] Yingying Wang, Yunlong Lin, Xuanhua He, Hui Zheng, Keyu Yan, Linyu Fan, Yue Huang, and Xinghao Ding. Learning diffusion high-quality priors for pan-sharpening: A two-stage approach with time-aware adapter fine-tuning. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 3

[67] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024. 6

[68] Yanyan Wei, Zhao Zhang, Jiahuan Ren, Xiaogang Xu, Richang Hong, Yi Yang, Shuicheng Yan, and Meng Wang. Clarity chatgpt: An interactive and adaptive processing system for image restoration and enhancement. *arXiv preprint arXiv:2311.11695*, 2023. 3

[69] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3

[70] 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. 3, 5, 6

[71] 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. 3, 6

[72] 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. 3, 6

[73] 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. 6

[74] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[75] 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. 6

[76] 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. 3, 4

[77] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 3

[78] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *European conference on computer vision*, pages 130–145. Springer, 2022. 1

[79] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 4

[80] Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Tianfan Xue, and Chao Dong. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*, 2024. 4

[81] 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. 3, 4

[82] 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. 2, 3, 5, 6

[83] 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. 3, 6

[84] 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. 6

[85] Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, and Rongrong Ji. Diffagent: Fast and accurate text-to-image api selection with large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6390–6399, 2024. 2

[86] Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pages 187–204. Springer, 2025. 1

[87] 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. 3

[88] Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. *arXiv preprint arXiv:2410.17809*, 2024. 3

[89] Jannik Zürn, Paul Gladkov, Sofía Dudas, Fergal Cotter, Sofi Toteva, Jamie Shotton, Vasiliki Simaiaki, and Nikhil Mohan. Wayvescenes101: A dataset and benchmark for novel view synthesis in autonomous driving. *arXiv preprint arXiv:2407.08280*, 2024. 3