Review

Not peer-reviewed version

Language-Driven Image Restoration and Semantic-Aware Quality Assessment: A Survey

Mingyu Liu , Haozhan Shu , Yuning Cui * , Xingcheng Zhou , Hu Cao , Wenqi Ren , Boxin Shi , Alois C. Knoll

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2366.v1

Keywords: image restoration; vision-language models; multimodal large language models; image quality assessment

 doi:10.20944/preprints202603.2366.v1

Review

Language-Driven Image Restoration and Semantic-Aware Quality Assessment: A Survey

Mingyu Liu [1], Haozhan Shu [2], Yuning Cui [1,*], Xingcheng Zhou [1], Hu Cao [1,3], Wenqi Ren [4], Boxin Shi [5] and Alois C. Knoll [1]

[1] Robotics, Artificial Intelligence and Real-Time Systems, Technical University of Munich, Germany
[2] School of Engineering and Design, Technical University of Munich
[3] School of Automation, Southeast University, Nanjing, China
[4] School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[5] State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China
* Correspondence: yuning.cui@in.tum.de

Abstract

Image restoration aims to recover a high-quality image from its degraded counterpart by mitigating distortions introduced during acquisition, transmission, or environmental interaction. Despite the remarkable progress of deep learning–based restoration models, most conventional approaches remain tightly coupled to predefined degradation assumptions and pixel-level supervision, limiting their capability to handle complex and diverse scenarios or user-dependent restoration targets. Recent advances in multimodal large language models (MLLMs) and vision–language models (VLMs) have introduced a new paradigm in which restoration systems can incorporate semantic reasoning, language-driven interaction, and cross-modal knowledge. In these frameworks, language models extend restoration beyond purely visual reconstruction by enabling degradation interpretation, perceptual alignment, and high-level control. In this survey, we present a systematic review of language-integrated restoration frameworks, organizing existing studies through an interaction-centric taxonomy that captures distinct modes of interaction between language models and restoration networks. We investigate how semantic priors, textual guidance, perceptual supervision, and decision-centric mechanisms recast restoration behavior, and analyze the implications of these developments for model design and training strategies. In parallel, we review emerging language-driven image quality assessment approaches that complement traditional evaluation metrics. Finally, we identify unresolved challenges and outline potential research directions toward more robust, efficient, and trustworthy restoration techniques.

Keywords: image restoration; vision-language models; multimodal large language models; image quality assessment

1. Introduction

Image restoration (IR), as a fundamental problem in low-level computer vision, aims to recover high-quality images from degraded counterparts. Over the past decades, numerous IR methods have been extensively studied for a wide range of tasks, including denoising [1–3], deraining [4–6], dehazing [7–9], desnowing [10–12], deblurring [13–15], low-light enhancement (LLIE) [16–18], super-resolution [19–21]. Beyond these tasks, IR techniques have also been widely applied to domain-specific scenarios, such as underwater enhancement [22–24], medical IR [25,26].

IR methodologies have evolved from model-driven approaches based on handcrafted priors to data-driven paradigms powered by deep neural networks, including CNNs [27], Transformers [28], and more recent architectures [29]. In parallel, restoration frameworks have progressed from task-specific designs, where each degradation type is modeled independently [5,9,30], to unified frameworks

 doi:10.20944/preprints202603.2366.v1

such as all-in-one (AiO) restoration, which aim to handle multiple degradation types within a single model [31–34].

Despite these advances, existing methods remain largely constrained by predefined degradation assumptions and are typically optimized with pixel-level supervision, limiting their ability to generalize to complex or user-dependent restoration scenarios. These limitations motivate the exploration of new paradigms beyond conventional restoration frameworks.

Recent breakthroughs in multimodal large language models (MLLMs) and vision–language models (VLMs) have opened new opportunities to address these challenges. By encoding rich semantic priors and exhibiting strong reasoning capabilities beyond conventional visual representations, MLLMs and VLMs have been increasingly introduced into IR pipelines [34–37]. Rather than directly performing pixel-level reconstruction, these models are typically employed as auxiliary components to provide high-level information, such as degradation interpretation, semantic guidance, and adaptive control. This emerging paradigm fundamentally reshapes IR from a purely visual mapping problem into a multimodal, semantically informed, and interactive framework.

This paradigm shift introduces a critical challenge in image quality assessment (IQA). Traditional IQA metrics [38–40] are effective for measuring pixel-level fidelity or perceptual similarity. However, they are not designed to assess semantic alignment or instruction consistency, leading to a mismatch between the restoration targets and the evaluation criteria. Recent language-driven IQA methods [41–47] address this gap by leveraging multimodal representations to assess semantic coherence and cross-modal consistency.

Although integrating language for IR and IQA has developed rapidly [34,37,48], a comprehensive survey of this emerging field remains absent, to the best of our knowledge. Existing surveys on IR primarily focus on architectural designs [49] or task-specific learning strategies [4,31,50], while recent reviews on multimodal models [51,52] seldom address low-level vision problems in depth.

In this survey, we provide the first systematic review of VLM-/MLLM-based image restoration and language-driven IQA. Figure 1 summarizes the taxonomy of this survey in a hierarchically structured way. We review advances in language-driven restoration frameworks, summarize representative model designs and training strategies, and discuss open challenges and potential future research directions.

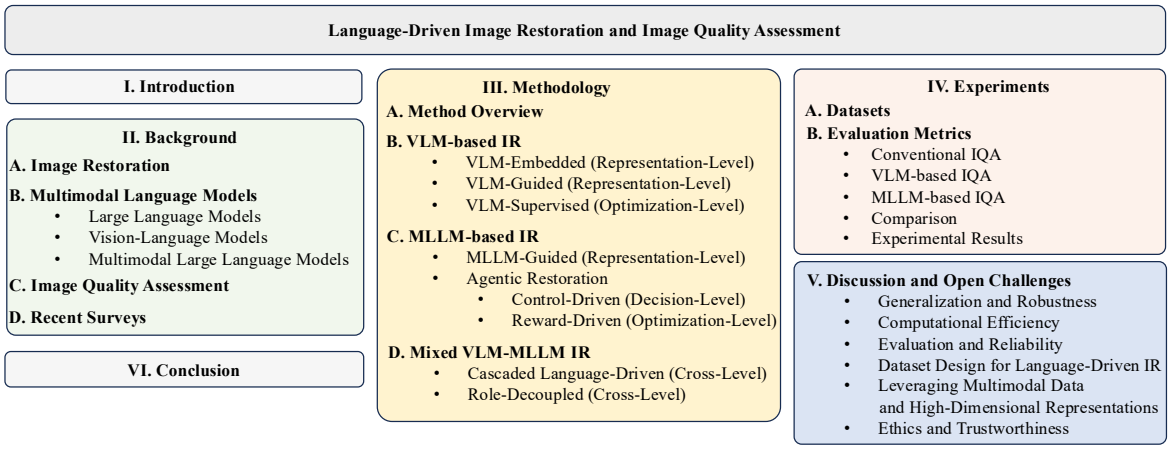

Figure 1. Unified taxonomy and survey structure of language-driven image restoration and image quality assessment.

The main contributions of this survey are summarized as follows:

- We introduce a unified conceptual framework that interprets language-driven IR as an interaction-centric paradigm, revealing how language models impact restoration behavior beyond architectural modifications.

 doi:10.20944/preprints202603.2366.v1

- We systematically categorize VLM-/MLLM-based IR methods through a functional role-oriented taxonomy, clarifying distinct interaction mechanisms including representation-level integration, optimization-level coupling, decision-driven control, and cross-level paradigms.
- We critically analyze VLM-/MLLM-based IQA, highlighting its conceptual distinctions from conventional fidelity metrics and clarifying challenges related to evaluation reliability, calibration stability, and semantic bias.
- We summarize the restoration datasets used in VLM-/MLLM-based frameworks and analyze their limitations and emerging requirements from a language-driven perspective, emphasizing the need for semantically enriched, language-aware benchmarks. We also provide comparisons between conventional frameworks and VLM-/MLLM-based methods across different settings.
- We investigate open challenges posed by language-integrated restoration systems and outline promising directions for future research that bridge multimodal reasoning, visual perception, and restoration optimization.

The rest of the work is organized as follows. Section 2 introduces the necessary preliminaries, including IR fundamentals, multimodal language models, and IQA, providing an overview of the core concepts relevant to this work. Section 3 reviews representative VLM-/MLLM-based IR approaches, and the analysis is guided by the proposed interaction-centric taxonomy with a focus on their model architectures and key technical innovations. Section 4 summarizes commonly used datasets for different IR tasks and discusses corresponding evaluation metrics, covering both traditional criteria and recent VLM-/MLLM-based assessment methods. Finally, Section 5 analyzes open challenges in current studies and outlines potential solutions and future research directions.

2. Background

In this section, we introduce the fundamental concepts and taxonomy of VLM-/MLLM-based image restoration, aiming to provide a structured background for understanding recent language-driven restoration frameworks.

2.1. Image Restoration

Image restoration aims to recover a clean image $\mathcal{I}(x)$ from its degraded observation $\mathcal{D}(x)$ by mitigating various types of distortions:

$$\mathcal{D}(x) = H(\mathcal{I}(x)) + \mathcal{N}, \tag{1}$$

where $H(\cdot)$ denotes the degradation operator and $\mathcal{N}$ represents additive noise.

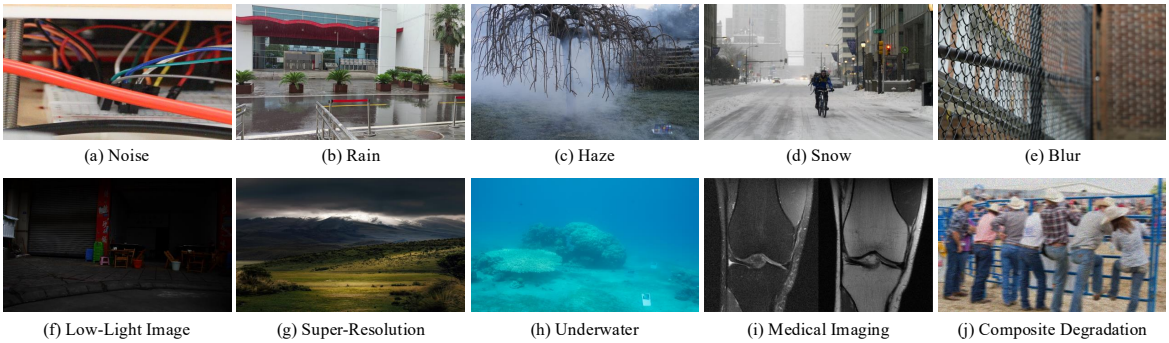

Figure 2. Examples of different degradation types and domains. The images correspond to the following open-source datasets: (a) PolyU [53], (b) LHP [54], (c) NH-HAZE [55], (d) RealSnow10K [56], (e) DPDblur [14], (f) LSRW [57], (g) DIV4K50 [58], (h) Squid [59], (i) FastMRI [60], and (j) MiO100 [61].

Existing IR methods can be broadly categorized into task-specific models and unified frameworks. Task-specific IR methods [8,17] are trained to handle a single degradation type. While such models often achieve strong performance within their targeted tasks, their specialization inherently restricts

 doi:10.20944/preprints202603.2366.v1

generalization to unseen or mixed degradations. This limitation motivates the development of general IR frameworks [62,63], which employ unified architectures capable of addressing multiple restoration tasks. Nevertheless, these approaches typically still rely on task-wise training or fine-tuning procedures. More recently, AiO restoration frameworks [32,33,64,65] have attracted increasing attention. By enabling multiple restoration tasks within a single model without explicit retraining, AiO methods aim to further improve generalization. This is often achieved through degradation-aware representations, prompt-based mechanisms, or multi-branch architectures that explicitly distinguish different degradation types.

Despite these advances, conventional deep learning–based restoration paradigms remain largely constrained by predefined task formulations and limited adaptability to diverse restoration requirements. To address these limitations, recent studies have started to explore the use of language models for IR [34,36,66–68]. Leveraging their strong semantic understanding, reasoning capability, and cross-modal alignment, language-driven approaches enable more flexible and interactive restoration pipelines, such as instruction-guided or context-aware restoration. This emerging paradigm marks a shift from task-centric restoration models toward more general, adaptive, and semantically informed IR frameworks.

2.2. Multimodal Language Models

Recent studies on multimodal language models involve multiple closely related model families, including large language models (LLMs), vision-language models (VLMs), and multimodal large language models (MLLMs). These models differ in their input modalities and output objectives, yet collectively form the foundation of language-driven IR methods reviewed in this survey. While capability boundaries between VLMs and MLLMs continue to develop, we adopt a functional distinction based on their dominant usage patterns in restoration frameworks.

Large language models. LLMs are foundation models trained on large-scale text datasets and operate exclusively on language inputs and outputs. Representative models include GPT [69], LLaMA [70], Qwen [71], and DeepSeek [72]. These models have achieved remarkable progress in natural language understanding, reasoning, and generation, demonstrating strong generalization across a wide range of language-centric tasks. LLMs typically do not process images directly. Instead, they receive visual information as textual descriptions, such as captions and image attribute descriptions.

Vision-language models. VLMs focus on aligning visual and textual representations through cross-modal embedding spaces [73–75]. By jointly modeling visual perception and language semantics, VLMs support cross-modal understanding and semantic grounding. Typical VLMs include CLIP-like [73,75] and ALIGN-like [74] architectures. While recent VLM models demonstrate limited reasoning ability, they can process images directly and are typically optimized for representation alignment.

Multimodal large language models. Building upon the success of LLMs, recent studies have extended LLMs to multimodal settings by giving them the capability to process non-textual inputs [76–80]. In particular, by integrating visual perception modules, MLLMs can directly adopt images as input and perform high-level reasoning. At the same time, these models preserve the ability to follow instructions and make stepwise decisions, making them suitable for tasks that need both image understanding and logical decision-making.

In this survey, we adopt the term MLLM to denote language-centric foundation models equipped with visual perception modules, regardless of whether they are implemented via explicit multimodal pretraining or modular integration. When referring to language-driven IR methods without emphasizing a specific model category, we adopt the umbrella term VLM-/MLLM-based IR.

2.3. Image Quality Assessment

IQA plays a critical role in image processing by enabling the evaluation and optimization of visual content quality [81]. Although subjective assessment based on the human visual system (HVS) is generally regarded as the most reliable criterion, it is costly and time-consuming, motivating extensive research into objective IQA metrics.

 doi:10.20944/preprints202603.2366.v1

Depending on the availability of reference images, IQA methods are broadly categorized into full-reference (FR-IQA) and no-reference (NR-IQA). FR-IQA methods compare distorted images with their high-quality references using metrics such as PSNR [39], SSIM [38], and MSE, as well as learning-based perceptual metrics [40,82–85]. In contrast, NR-IQA predicts image quality directly from distorted inputs without reference images. Early methods rely on handcrafted natural scene statistics, such as BRISQUE [86], NIQE [87], and PIQE [88], while recent approaches adopt deep learning to model complex perceptual patterns [89–93].

Most recently, VLM-/MLLM-based IQA methods [41,42,45–47,94–96] have emerged as a new paradigm. By leveraging multimodal reasoning capabilities and foundational knowledge, these approaches aim to narrow the gap between objective metrics and subjective human assessment. Some methods evaluate image quality using natural languages [96–98], while others directly produce quantitative quality scores [45,95,99].

2.4. Relevant Surveys

A number of surveys have reviewed IR from different perspectives. Many of these works organize the literature according to specific degradation types or application domains [1,4,7,13,16,19,22,31,50]. Along the recent trend of AiO restoration, Jiang *et al.* [31] provided a systematic overview of AiO restoration frameworks. Beyond task-oriented surveys, several studies have examined IR from a model-centric perspective. For example, Su *et al.* [49] reviewed deep learning architectures widely adopted in IR, while Li *et al.* [100] offered an in-depth review of diffusion-based IR methods. Despite these efforts, the rapid emergence of VLM-/MLLM-based IR methodologies has not yet been systematically reviewed. Differently, in this work, we present a comprehensive review of VLM-/MLLM-based IR from the following three aspects: 1) recent advances in VLM-/MLLM-based IR methods, 2) datasets and VLM-/MLLM-assisted evaluation protocols, and 3) benchmarking and comparative evaluation of VLM-/MLLM-based approaches.

3. Methodology

This section first categorizes VLM-/MLLM-based method prototypes and then analyzes each category in detail through representative approaches. Figure 3 summarizes the representative paradigms of existing VLM-/MLLM-based IR methods.

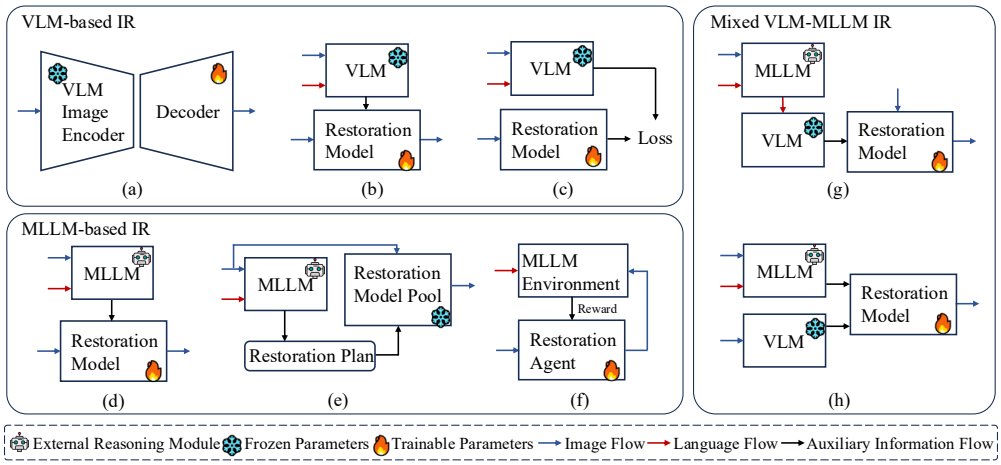

Figure 3. Categories of VLM-/MLLM-based IR paradigms. We summarize representative existing frameworks into seven prototypes. Specifically, VLM-based methods can be categorized into three paradigms: (a) VLM-Embedded IR, (b) VLM-Guided IR, and (c) VLM-Supervised IR. Similarly, MLLM-based approaches are grouped into three categories: (d) MLLM-Guided IR, (e) Control-Driven IR, and (f) Reward-Driven IR. Finally, (g) Cascaded Language-Guided IR and (h) Role-Decoupled IR represent cross-level paradigms that combine MLLMs and VLMs within a unified restoration pipeline.

 doi:10.20944/preprints202603.2366.v1

3.1. Overview of Language-Driven Restoration and Taxonomy Definition

Recent advances in MLLMs and VLMs have significantly influenced the design of IR systems. Rather than solely improving restoration architectures, an increasing number of approaches leverage language-driven semantic priors, cross-modal reasoning, and high-level decision-making to enhance restoration robustness, flexibility, and controllability. As a result, clarifying the roles of language models and how they interact with restoration pipelines becomes critical for understanding existing frameworks and guiding future research.

Existing surveys on IR mainly organize the literature from architecture-driven (e.g., CNN or transformer) [49] or task-driven [4,31] perspectives. However, these categorizations become insufficient for language-driven restoration systems, where the primary methodological distinctions arise from interaction mechanisms between language models and restoration networks. Specifically, language-driven restoration systems reshape restoration behavior by modifying information flow structures, supervision signals, control mechanisms, and optimization objectives. Therefore, to provide a structured understanding of this rapidly growing body of work, we adopt an *interaction-centric* taxonomy that categorizes methods according to the functional role of language models within restoration pipelines. The definition is described below:

Definition 1: Interaction Interface of Language Models in Restoration

We characterize the involvement of VLMs or MLLMs in IR by an interaction interface tuple:

$$\mathcal{T} = (\mathcal{I}, \mathcal{O}, \mathcal{G}),$$

where $\mathcal{T}$ represents the interaction interface between the language model (LM) and the restoration pipeline.
(1) Input space: $\mathcal{I}$ denotes the modalities consumed by the LM, including degraded images, textual instructions, intermediate features, or evaluation scores.
(2) Output space: $\mathcal{O}$ denotes the signals produced by the LM that influence the restoration pipeline, such as aligned embeddings, conditioning prompts, differentiable loss terms, restoration plans, or scalar rewards.
(3) Coupling function: $\mathcal{G}$ specifies how $\mathcal{O}$ modifies the restoration process, including feature conditioning, perceptual supervision, control-based execution, or reward-driven policy optimization.

Based on this interface characterization, we distinguish different paradigms according to the dominant interaction type:

- **Representation-Level Coupling:** LM outputs modify forward feature representations without altering the optimization objective or execution logic. This includes embedding replacement and semantic conditioning.
- **Optimization-Level Coupling:** LM outputs reshape the training objective by defining differentiable supervision signals or scalar reward functions, thereby altering optimization dynamics.
- **Decision-Level Coupling:** LM outputs regulate the execution structure of the restoration pipeline, such as task decomposition, module scheduling, or control policies.
- **Cross-Level Coupling:** Systems that simultaneously integrate multiple interaction depths within a unified framework, combining representation, optimization, and/or decision mechanisms.

On the other hand, from the perspective of model families, existing approaches are naturally organized into VLM-based, MLLM-based, and hybrid VLM–MLLM frameworks. This organization reflects practical architectural distinctions and facilitates systematic presentation. However, the model family alone does not fully explain how the introduction of language information influences restoration behavior. Within each family, language models may intervene at different depths of the restoration pipeline, from representation modulation to decision-level orchestration. Therefore, the interaction-

 doi:10.20944/preprints202603.2366.v1

centric taxonomy introduced above serves as a complementary analytical dimension that abstracts the coupling mechanism between language reasoning and pixel reconstruction, independent of whether VLMs or MLLMs are employed. In other words, the remaining parts are organized by model families for clarity, while the interaction-centric taxonomy provides a unified theoretical interpretation across these categories by revealing how semantic information affects representation, optimization, and decision-making processes.

VLM-based image restoration operates mainly at the representation and supervision levels of interaction depth. These methods exploit pretrained VLMs, such as CLIP [73], to extract semantically aligned embeddings from textual and visual inputs and integrate them into conventional restoration networks. Specifically, the representation-level coupling includes: (a) VLM-embedded image restoration and (b) VLM-guided image restoration. In contrast, (c) VLM-supervised image restoration corresponds to optimization-level coupling.

MLLM-based image restoration extends IR frameworks by introducing multimodal large language models as external agents. MLLM-based systems often not only modulate at representations but also intervene at optimization or decision levels, enabling higher-level control and policy adjustment. Accordingly, existing MLLM-based approaches can be grouped into three categories: (d) MLLM-guided image restoration, (e) control-based agentic image restoration, and (f) reward-driven agentic image restoration.

Cross-level hybrid systems combine multiple interaction depths within a unified framework. Two representative formats are: (g) cascaded language-guided image restoration, representing a hybrid paradigm that hierarchically combines MLLMs and VLMs within a unified pipeline. The second one is (h) role decoupled image restoration, where VLMs and MLLMs operate as independent but complementary modules.

Although the proposed interaction-centric framework provides a principled interpretation of language-driven restoration systems, practical implementations may exhibit hybrid characteristics. To reduce ambiguity, we clarify the categorization criteria based on the primary functional influence of language models. Figure 4 illustrates these paradigms along a temporal axis, highlighting the rapid development and variety of language-driven IR approaches since 2023. In the following subsections, we review each paradigm in detail, summarizing representative methods, core design improvements, as well as their respective strengths and limitations.

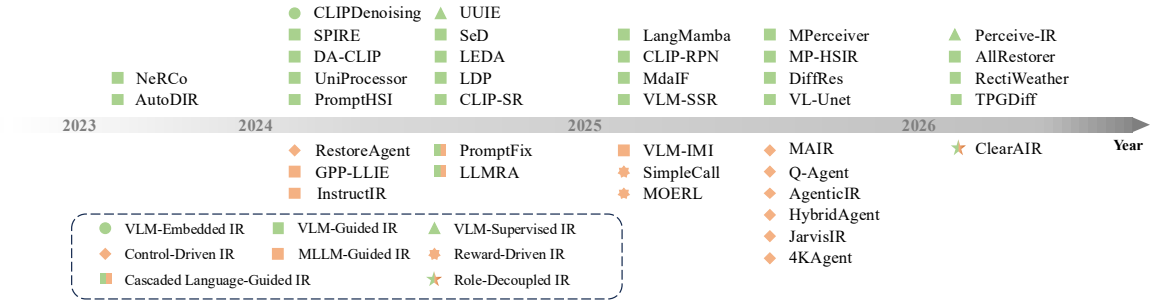

Figure 4. Timeline of representative LLM/VLM-based IR approaches from 2023 to Jan. 2026, where colors encode model families (VLM, MLLM, or hybrid) and marker shapes indicate different interaction paradigms between language models and restoration models.

3.2. VLM-Based Image Restoration

VLM-Embedded Image Restoration. As one subtype of the representation-based interaction, VLM-Embedded IR directly integrates representations from pretrained VLMs into restoration backbones, typically by replacing the visual encoding stage. In this paradigm, VLMs act as fixed feature extractors that inject robust and semantically aligned priors into conventional IR networks. During CLIP pretraining, visual representations are optimized to align with text semantics, which are inherently insensitive to high-frequency details. As a result, CLIP features tend to emphasize semantic content while being less sensitive to degradation-specific perturbations, thereby improving robustness.

 doi:10.20944/preprints202603.2366.v1

CLIPDenoising [101] incorporates a frozen CLIP [102] visual encoder together with a learnable denoising decoder. By leveraging the distortion-invariant and content-aware properties of CLIP's dense visual features, the method demonstrates enhanced out-of-distribution (OOD) generalization under unseen degradation conditions.

VLM-Guided Image Restoration. In this paradigm, VLM-derived signals, such as semantic cues or degradation-aware representations, are incorporated into restoration networks through feature modulation, conditioning mechanisms, or representation alignment.

For example, CLIP-RPN [35] leverages CLIP features to characterize rain patterns and intensities, enabling adaptive deraining via dynamic cross-attention modulation. Similarly, LDP [103] utilizes text-aligned embeddings to describe blur attributes, allowing restoration models to adjust internal feature representations. In structured medical imaging, LEDA [104] enforces consistency across perceptual representations and semantic tokens guided by a language-informed codebook, while LangMamba [105] integrates language-aligned features into a Mamba-based backbone for efficient CT denoising. In addition, CLIP-SR [106] couples linguistic guidance with image feature processing to refine structure and textures for super-resolution. In AiO multi-degradation restoration, UniProcessor [107], VL-Unet [67], MdaIF [108], and AllRestorer [109] exploit textual priors to identify and handle mixed degradations. In hyperspectral IR, MP-HSIR [110] combines spectral prompts with language-visual prompts, where spectral representations provide low-rank priors and language guidance conveys degradation semantics. PromptHSI [111] further represents degradation factors through prompts to realize a universal hyperspectral restoration model.

Beyond feed-forward restoration networks, VLM guidance has been widely integrated into diffusion-based frameworks. In these approaches, VLM-derived embeddings or perceptual priors serve as conditioning signals that constrain the diffusion process. Concretely, DA-CLIP [112] explores controllable adaptation of VLM representations for multi-task restoration. Methods such as AutoDIR [113], SPIRE [114], and MPerceiver [115] encode degradation descriptors into language-aligned embeddings for diffusion conditioning. TPGDiff [116] further extends this idea by incorporating VLM-derived degradation priors into a hierarchical multi-prior formulation, constraining the diffusion trajectory at multiple stages. VLM-SSR [117] introduces CLIP-based semantic features to guide real-world super-resolution diffusion.

VLM-guided strategies have also been explored in adversarial learning settings. By conditioning discriminators on VLM-derived representations, methods such as NeRCo [118] and SeD [119] introduce semantic-aware discriminative signals that complement adversarial objectives.

VLM-Supervised Image Restoration. VLM-Supervised IR employs VLMs as perceptual supervisors to guide the training of restoration networks by constructing semantic-aware supervision, typically in the form of differentiable perceptual losses. By measuring the similarity between restored images and textual descriptions of desirable visual attributes, restoration models are optimized to produce outputs that are semantically aligned with target concepts.

This paradigm is particularly effective in scenarios where paired ground-truth data are scarce or unavailable. A representative application is underwater image enhancement [120], where a FLIP [121]-pretrained text encoder is used to encode high-level quality concepts through textual prompts, and the resulting similarity scores are incorporated as perceptual loss terms alongside traditional reconstruction objectives. Furthermore, Perceive-IR [122] enables fine-grained quality control through CLIP-aligned prompt learning and difficulty-adaptive supervision. However, since the supervision from VLMs is typically global, this paradigm may be less sensitive to fine-grained local artifacts, and its effectiveness depends on prompt design and the expressive capacity of the underlying VLM.

Methodological Synthesis. Despite leveraging different coupling mechanisms, VLM-based IR methods exhibit several shared methodological characteristics. Usually, VLMs act as semantic regularizers in many restoration frameworks. By leveraging language-aligned representations that emphasize content-level invariance, these models introduce features that are inherently less sensitive to degradation-specific perturbations. This property improves restoration robustness, particularly under

 doi:10.20944/preprints202603.2366.v1

distribution shifts and unseen degradations. As a result, most frameworks treat pretrained VLMs as fixed modules, leveraging their generalizable cross-modal representations without task-specific retraining.

3.3. MLLM-Based Image Restoration

MLLM-Guided Image Restoration. MLLM-guided IR refers to a class of methods that use MLLMs to interpret images or assess perceptual quality and, based on the resulting outputs, generate auxiliary guidance signals for IR networks. Typical guidance includes instruction interpretation, perceptual attribute assessment, or preference feedback that conditions restoration models in a soft and interpretable manner.

Representative works include instruction-driven restoration methods such as InstructIR [34], which uses natural language instructions to model a unified all-in-one restoration network across multiple degradation types. VLM-IMI [123] further explores instruction-based guidance by generating textural descriptions from the input images and user prompts. Beyond instruction following, perceptual guidance has also been investigated. For example, GPP-LLIE [124] introduces a generative perceptual prior for low-light enhancement by extracting high-level perceptual attributes from LLaVA [44]. Additionally, SnowMaster [56] leverages MLLM-derived perceptual preference feedback to rank candidate desnowing results and uses the selected outputs as pseudo-labels to support semi-supervised training on real-world snowy images.

Agentic Restoration. In these frameworks, MLLMs operate as high-level decision-making modules that regulate restoration behavior through planning or optimization mechanisms. Instead of directly performing pixel reconstruction, language models govern how restoration actions are selected, scheduled, or evaluated. From a system perspective, existing approaches can be broadly categorized into control-driven and optimization-driven paradigms.

Control-Driven Image Restoration. In contrast to guidance-based approaches that provide external semantic cues or conditioning signals, control-driven restoration assigns MLLMs the role of central controllers responsible for structuring the restoration procedure. The restoration problem is interpreted as a sequential decision process in which degradations are analyzed, tasks are decomposed, and restoration operations are scheduled. Language models regulate execution logic, while task-specific restoration networks remain responsible for pixel-level recovery.

Early works such as RestoreAgent [125] and AgenticIR [37] demonstrate the feasibility of this paradigm by enabling MLLMs to autonomously determine restoration tasks and execution sequences. Subsequent works focus on improving planning stability and computational efficiency. Q-Agent [126] employs chain-of-thought [127] (CoT) reasoning to decompose multi-degradation perception and adopts quality-driven greedy planning based on no-reference image quality assessment (IQA), effectively mitigating unnecessary rollbacks. To enhance scalability and robustness, several studies extend MLLM-planned restoration to multi-agent systems. MAIR [128] consists of a scheduler–expert hierarchy to decouple degradation perception and restoration execution, while HybridAgent [129] adopts collaborative agents to balance planning accuracy and computational efficiency. Specialized frameworks such as 4KAgent [58] further adapt the planning paradigm to ultra-high-definition (UHD) IR. System-level frameworks, including Clarity ChatGPT [130] and JarvisIR [36], emphasize interactive refinement and robustness in real-world environments.

Reward-Driven Image Restoration. Reward-driven restoration as an optimization-level coupling treats language models primarily as evaluators rather than controllers. Instead of structuring restoration procedures, MLLMs define perceptual objectives that guide policy learning. Restoration behavior is shaped through reward signals, preference feedback, or evaluator-based scoring, shifting the learning target from explicit reconstruction supervision toward perceptual optimization.

MOERL [131] formulates IR as a reinforcement learning (RL) problem, where a Mixture-of-Experts (MoE) model is dynamically refined based on perceptual rewards. A reward-driven policy is learned to adaptively route restoration actions without requiring explicit degradation labels, enabling robust handling of complex and mixed weather conditions. In contrast, SimpleCall [132] adopts a label-free

 doi:10.20944/preprints202603.2366.v1

decision process in which a lightweight agent is trained to sequentially invoke restoration tools using MLLM-derived perceptual feedback. By optimizing restoration policies through evaluator-based rewards, the framework enables effective restoration without ground-truth supervision.

Methodological Synthesis. MLLM-based restoration paradigms extend coupling from representation modules to optimization and decision levels. Specifically, rather than serving as auxiliary feature providers, these models participate in degradation interpretation, task decomposition, and restoration planning. Across guided, planned, and agentic formulations, MLLMs operate at a higher semantic level, representing a structural reformulation of restoration pipelines rather than conventional multimodal enhancement.

3.4. Cross-Level Hybrid Systems

Cascaded Language-Guided Image Restoration. Cascaded language-guided IR adopts a multi-stage pipeline in which language understanding and visual restoration are decomposed into hierarchically connected modules. In this paradigm, high-level semantic understanding is first performed by MLLMs, and the resulting linguistic representations are subsequently propagated to VLM components to bridge language semantics with visual embedding spaces. The transformed semantic information is then utilized to support downstream pixel-level restoration.

For instance, LLMRA [68] employs an MLLM [133] to perform high-level degradation understanding. The generated language descriptions are then mapped into an embedding space via a pretrained CLIP text encoder, forming structured degradation representations that condition a restoration network. Similarly, PromptFix [134] utilizes a combination of LLaVA [44] together with a pretrained CLIP visual encoder to extract semantic cues, which are used as auxiliary prompts to guide an instruction-driven diffusion restoration model. While both LLMRA and PromptFix adopt cascaded language-guided pipelines, they differ in the functional criticality of language guidance. In LLMRA, language-generated degradation representations constitute a structural prerequisite for restoration, whereas in PromptFix, language guidance from the MLLM serves as an auxiliary enhancement that complements a generative prior.

Role-Decoupled Image Restoration. In contrast to cascaded language-guided pipelines that integrate VLMs and MLLMs into a hierarchical semantic processing chain, the role-decoupled interaction pattern employs them as parallel and functionally independent modules. In this paradigm, VLMs and MLLMs are utilized for different roles, such as task identification or perceptual assessment, and jointly support the restoration process.

Following this paradigm, ClearAIR [66] employs a VLM-based task identifier [112] to recognize degradation types and determine appropriate restoration branches, while an MLLM-based IQA module [45] provides global perceptual quality scores that guide restoration refinement. Such designs allow different language models to specialize in heterogeneous subtasks, improving system interpretability while avoiding error accumulation caused by hierarchical semantic dependencies.

Methodological Synthesis. Mixed systems demonstrate cross-level coupling and typically exploit functional complementarity, where VLMs provide stable representation-level priors, and MLLMs contribute higher-level reasoning capabilities. Such designs mitigate limitations of single-model systems while promoting modular restoration architectures. This transition suggests that language-driven restoration may evolve toward heterogeneous multimodal systems, in which restoration behavior emerges from coordinated interactions among specialized semantic and visual components.

4. Experiments

In this section, we summarize representative datasets widely used in language-driven IR frameworks across different tasks. We then review commonly adopted IQA metrics, covering both classical and recent VLM-/MLLM-based evaluation paradigms. Together, these datasets and metrics provide a comprehensive view of current benchmarks and evaluation practices for language-driven IR.

 doi:10.20944/preprints202603.2366.v1

4.1. Datasets

Existing VLM-/MLLM-based IR studies commonly adopt subsets of established restoration benchmarks, as summarized in Table 1. These datasets vary substantially in scale, spatial resolution, scene diversity, and data-acquisition protocols, spanning both real-world captures and synthetically generated degradations. While synthetic datasets offer controllability and scalability, they may inadequately capture the ambiguity, complexity, and stochastic characteristics of real-world degradations. This discrepancy becomes particularly critical in language-driven frameworks, where degraded understanding relies on semantic representations that are inherently sensitive to contextual variations.

Table 1. Summary of datasets used in VLM- and MLLM-based image restoration. *r* and *s* denote real and synthetic data, respectively; "–" indicates unavailable train/test splits. LR and HR refer to low- and high-resolution.

Task	Dataset	Year	Type	Domain	Training/Testing	Description
Denoising	Kodak24 [135]	1999	r	Natural	-/24	Clean color images
	McMaster [136]	2011	r	Natural	-/18	18 high quality color images
	CBSD68 [137]	2001	r	Natural	-/68	68 clean natural images with different noisy levels
	Urban100 [138]	2015	r	Natural	-/100	100 high resolution urban scenes with repetitive structures
	DIV2K [139]	2017	r	Natural	800/100	1000 high resolution images
	SIDD [140]	2018	r	Natural	-/160	Real-noise image pairs with clean ground truth
	PolyU [53]	2018	r	Natural	-/40	Real-noise paired dataset with 40 scenes
	WED [141]	2016	r&s	Natural	4744/-	Waterloo Exploration Database
	BSD400 [142]	2010	r	Natural	400/-	Training subset from BSD500
	Mayo-2016 [26]	2016	r	Medical	4800/1136	Paired normal-dose and simulated quarter-dose abdominal CT
Deraining	Rain100L [143]	2017	s	Natural	200/100	Images with light rain effect
	Rain100H [143]	2017	s	Natural	1800/100	Images with heavy rain conditions
	Rain800 [144]	2019	s	Natural	700/100	Images with diverse rain patterns
	Rain1400 [145]	2017	s	Natural	12600/1400	14 rain streak types
	Raindrop [146]	2018	r	Natural	1069/58	A paired raindrop dataset captured using dual identical glass setups
	Outdoor-Rain [147]	2019	r&s	Natural	9000/1500	A synthetic outdoor rain dataset with streak and accumulation effects
	RainDS [148]	2021	r&s	Natural	-/5800	Paired deraining dataset organized as a 4-image set
	SSID [149]	2022	r&s	Natural	47600/200	Semi-supervised image deraining sets
	LHP [54]	2023	r	Natural	2100/300	Largest paired real rain dataset with 1920 × 1080 image resolution
Dehazing	FoggyCityscapes [150]	2018	s	Natural	2975/1525	Paired foggy and clear images
	ACDC [151]	2021	r	Natural	1600/2400	Real-world images captured under adverse conditions
	RESIDE [152]	2018	r	Natural	86125/4842	Real and synthetic data across indoor and outdoor scenarios
	NH-HAZE [55]	2020	r	Natural	45/5	A real paired outdoor dehazing set with non-homogeneous haze
	Dense-Haze [153]	2019	r	Natural	45/5	A real paired dehazing dataset for dense, homogeneous haze
Desnowing	RealSnow10K [56]	2025	r	Natural	6406/1047	Real-world snow removal dataset
	Snow100K-L [12]	2018	s	Natural	1872/601	A single-image snow removal benchmark
Deblurring	DPD-blur [14]	2020	r	Natural	350/150	500 real defocus blur image pairs
	DPD-disp [154]	2020	r	Natural	-/350	Reuse the checkpoints trained on the DPD-blur dataset
	DDD-syn [155]	2021	s	Natural	10000/1000	Synthetic deblurring dataset with paired blurry and sharp images
	RDPD [156]	2021	s	Natural	18000/1000	Images captured using a dual-pixel camera
	GoPro [15]	2017	r&s	Natural	2103/1111	Paired images generated from real high-frame-rate GoPro videos

 doi:10.20944/preprints202603.2366.v1

Table 1. *Cont.*

Task	Dataset	Year	Type	Domain	Training/Testing	Description
LLIE	LOL-v1 [17]	2018	r	Natural	485/15	Paired low-light and normal-light under controlled conditions
	LSRW [57]	2023	r	Natural	445/50	paired low-light LR with normal-light HR
	DICM [157]	2013	r	Natural	-/64	Low light images without ground truth for visual comparison
	NPE [158]	2013	r	Natural	-/85	Unpaired low light images
	VV [159]	2018	r	Natural	-/24	24 real-world unpaired low light images
	LOL-v2-real [18]	2021	r	Natural	689/100	Real paired low-light sets
	LOL-v2-syn [18]	2021	s	Natural	900/100	Synthetic paired low-light sets
	MEF [160]	2015	r	Natural	-/17	Multiple images with different exposure levels for the same scene
	SICE [161]	2018	r	Natural	360/229	Multiple reference images of different enhancement levels
	LIME [162]	2016	r	Natural	-/10	10 images without ground truth
Underwater	EUVP [24]	2019	r	Underwater	20000/-	Include both paired and unpaired samples
	UIEB [23]	2019	r	Underwater	800/90	Underwater image enhancement benchmark
	RUIE [163]	2020	r	Underwater	-/4230	Real-world underwater image enhancement
Super Resolution	Set5 [164]	2021	r	Natural	-/5	5 real-world natural images
	Set14 [165]	2010	r	Natural	-/14	14 real-world natural images
	Manga109 [166]	2017	r	Natural	-/109	109 real-world manga images
	CelebA [167]	2015	r	Natural	162770/19867	Images with 40 binary attributes
	RealSR [168]	2019	r	Natural	-/35	Real-world low-and high-resolution image pairs
	DrealSR [169]	2020	r	Natural	-/93	93 aligned LR-HR image pairs
	DIV2K-Val [170]	2024	r	Natural	-/100	3K patches from the DIV2K validation set
	RealSRSet [171]	2021	r	Natural	-/20	comprising images captured in practical scenarios
	DIV4K-50 [58]	2024	r	Natural	-/50	256 × 256 distorted images paired with 4096 × 4096 counterparts
	DiffusionDB [172]	2023	s	Natural	-/100	Text-to-image prompt gallery sets
	AID [173]	2017	r	Natural	-/135	Aerial image dataset
	DIOR [174]	2019	r	Natural	-/154	Object detection in optical remote sensing images
	DOTA [175]	2018	r	Natural	-/183	Dataset for object detection in aerial images
	bcSR [176]	2023	r	Medical	-/200	Pathology images patches from breast cancer whole slide images
	US-Case [177]	2025	r	Medical	-/111	Ultrasound cases
Composite	PromptFix [134]	2024	r&s	Natural	101320/-	Paired input–goal–instruction triplets spanning 7 tasks
	MiO100 [61]	2024	r&s	Natural	-/700	Each image is degraded with 7 single degradation types
	AgenticIR [37]	2025	r&s	Natural	-/1440	16 mixed-degradation combinations (2–3 types)
	CleanBench [36]	2025	r&s	Natural	150000/80000	A large-scale, high-quality instruction-response
	MSRS [178]	2022	r	Natural	1163/361	A multi-spectral IR-VIS paired set
	FMB [179]	2023	r	Natural	1220/280	1500 aligned pairs
	CDD-11 [180]	2024	r&s	Natural	13013/2200	1080 × 720 images selected from the RAISE dataset
	TOLED [181]	2021	r	Natural	240/30	A real paired under-display camera restoration set
	AVIRIS [182]	2024	r	HSI	1678/200	Airborne visible/infrared imaging spectrometer
	ARAD [183]	2022	r	HSI	1000/-	A large natural spectral image set

On the other hand, recent language-driven restoration frameworks increasingly require datasets that explicitly model image-language interactions. However, as shown in Figure 5, most datasets consist primarily of degraded-clean image pairs while rarely providing structured linguistic annotations describing degradation characteristics or perceptual attributes. The absence of structured language annotations may introduce inconsistencies when training or evaluating language-driven restoration models. To bridge this gap, PromptFix [134] constructs a large-scale instruction-following dataset that contains approximately 1.01 million input-goal-instruction triplets across diverse low-level tasks. Furthermore, in safety-critical applications such as autonomous driving, CleanBench [36] defines an instruction sample as a triplet, consisting of a user instruction, a degraded image, and a response. Such datasets reflect a paradigm shift in which restoration is formulated not only as pixel reconstruction but also as a language-conditioned generation problem. Despite these initial explorations, restoration datasets with captions are still underexplored.

 doi:10.20944/preprints202603.2366.v1

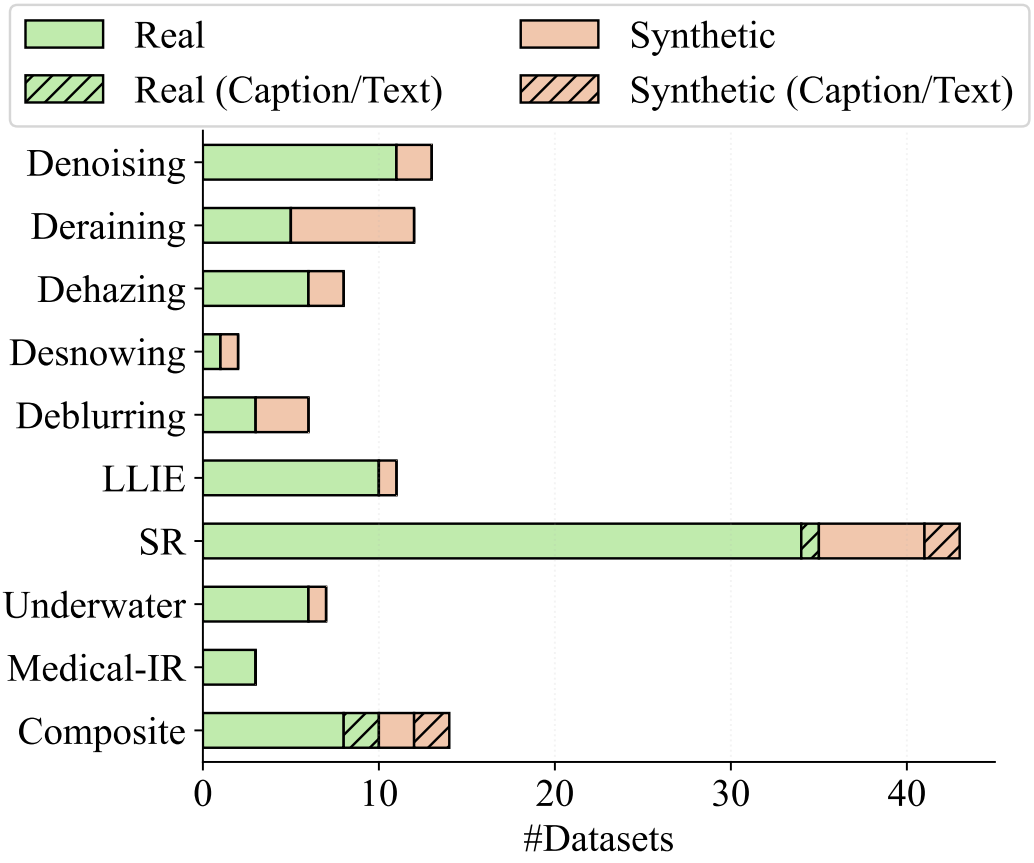

Figure 5. Distribution of datasets across restoration tasks. Bars indicate the number of datasets per task, distinguishing real-world and synthetic datasets. Hatched segments denote datasets accompanied by textual or caption annotations. SR denotes super-resolution.

4.2. Evaluation Metrics

Existing IQA metrics are summarized in Table 2, including FR-IQA, NR-IQA, and other evaluation paradigms.

Conventional IQA metrics. Image restoration performance is traditionally evaluated using IQA metrics that measure fidelity or perceptual similarity between restored images and references. Full-reference IQA metrics, such as PSNR [39], SSIM [38], FSIM [184], and MAE, quantify pixel-level or structural consistency with ground-truth images. Learning-based metrics (e.g., LPIPS [40], DISTS [82], CKDN [83], and AHIQ [84]) assess perceptual similarity in deep feature spaces and exhibit improved correlation with human judgments. When reference images are unavailable, no-reference IQA metrics are commonly adopted, ranging from hand-crafted statistical measures (e.g., BRISQUE [86], NIQE [87], and PIQE [88]) to learning-based models such as MUSIQ [89] and MANIQA [91]. In addition, distribution-based metrics like FID [185] are often used to evaluate feature distribution alignment between restored and real images, particularly in generative restoration scenarios.

doi:10.20944/preprints202603.2366.v1

Table 2. Taxonomy of Image Quality Assessment Methods and Evaluation Paradigms in Image Restoration. Metrics are categorized according to supervision requirements, evaluation paradigms, and application scope. GT indicates whether ground-truth images are required.

Category	Sub-category	Representative Metrics	GT	Usage
Full-Reference	Non-learning-based	PSNR, SSIM [38], FSIM [184], MAE, MSE, RMSE, ERGAS [186]	✓	Pixel-level fidelity or structural consistency evaluation
	Learning-based	LPIPS [40], DISTS [82], CKDN [83], AHIQ [84], TOPIQ-FR [85]	✓	Feature-based perceptual similarity
	Distribution-based	FID [185]	✓	Feature-space distribution alignment
No-Reference	Hand-crafted	BRISQUE [86], NIQE [87], PIQE [88], LOE [158], PI [187]	×	Blind perceptual quality estimation
	Learning-based	MUSIQ [89], MANIQA [91], NIMA [90], HyperIQA [92], PAQ2-PIQ [188], DBCNN [189], TOPIQ-NR [85], CNNIQA [93]	×	Learning-based NR-IQA
	VLM-based	CLIP-IQA [41], QualiCLIP [43], UNIQA [190], LIQE [191], GRMP-IQA [192], PromptIQA [193], ATTIQA [194], SFD [195]	×	Vision-language-aligned perceptual quality evaluation
	MLLM-based	DeQA-Score [45], Q-Align [46], Compare2Score [95], Dog-IQA [196], Q-Insight [197], DepictQA [96], DepictQA-Wild [198], IQAGPT [199], Q-Hawkeye [200], Q-Ground [98], Q-Scorer [99], AgenticIQA [48], SEAGULL [97], SCUIA [201], LEAF [202], Q-Ponder [203], CAP-IQA [204], Co-Instruct [205], Q-Instruct [94], Q-Bench [206], Q-Bench^{+} [207], RALI [208]	×	Language-grounded perceptual reasoning, preference modeling, and quality scoring
Evaluation Paradigms	Human-aligned	PLCC, SRCC, KRCC [209], Weighted Kappa [210], Percent Agreement	×	Correlation with human subjective quality perception
	Task-oriented	Precision, Recall, F1, mIoU, Accuracy [96]	✓	Downstream task performance
	Text-based	BLEU-N [211], ROUGE-L [212], METEOR [213], CIDEr [214]	✓	Textual or semantic fidelity evaluation

 doi:10.20944/preprints202603.2366.v1

VLM-based IQA. VLMs have recently been widely employed for NR-IQA, leveraging multimodal representations learned through large-scale image–text pretraining, such as CLIP [102]. Unlike conventional IQA approaches that rely on explicit distortion modeling or perceptual regression networks, VLM-based methods exploit the implicit perceptual priors encoded within cross-modal alignment spaces.

Early studies, such as CLIP-IQA [41], demonstrate that quality perception can be formulated as a prompt-driven similarity-comparison problem using antonym prompt pairs (e.g., *good* vs. *bad*). Subsequent works extend this principle in various directions. QualiClip [43] proposes a self-supervised, opinion-unaware approach via quality-aware image–text ranking to learn distortion-sensitive representations. PromptIQA [193] further enables requirement-adaptive assessment by incorporating a small set of image–score pairs as prompts. In addition, GRMP-IQA [192] improves data efficiency through meta-prompt learning with gradient regularization, and ATTIQA [194] addresses annotation scarcity by introducing attribute-aware pretraining that leverages VLMs to generate quality-related attribute pseudo-labels. In summary, these approaches highlight that perceptual quality can be effectively inferred from language-aligned representations, often exhibiting strong cross-dataset generalization without explicit mean opinion scores (MOS) supervision. In special domains, SCUIA [201] explicitly models semantic context via contrastive learning for underwater IQA.

Beyond prompt-based mechanisms, LIQE [191] introduces a vision–language correspondence framework that jointly models semantics, distortions, and quality, whereas UNIQA [190] proposes a unified multimodal pretraining strategy bridging image quality and aesthetic assessment. Meanwhile, SFD [195] explores semantic feature discrimination as a proxy for quality estimation, demonstrating that perceptual quality can be approximated via semantic consistency and feature-space discrimination.

In summary, VLM-based IQA treats IR as a multimodal alignment problem, offering advantages in training efficiency and generalization, yet remaining constrained by prompt dependency, semantic bias, and imperfect calibration with human subjective judgments.

MLLM-based IQA. MLLM-based IQA naturally supports both language-based quality representation and quantitative scoring paradigms. Description-oriented approaches formulate IQA as a language-grounded reasoning task, where quality is expressed through structured linguistic outputs that improve interpretability and human alignment. Representative frameworks like DepictQA [96] and DepictQA-Wild [198] formulate IQA as descriptive and comparative language-based quality understanding to achieve human-like assessment. IQAGPT [199] proposes a caption-driven pipeline to generate quality scores and natural-language explanations. Beyond global quality understanding, SEAGULL [97] highlights region-aware quality reasoning through SAM-guided [215] feature modeling, while Q-Ground [98] introduces a detailed visual quality analysis through visual grounding. Additionally, Co-Instruct [205] explores an open-ended comparative reasoning framework, facilitating MLLMs to produce quality comparisons with detailed reasoning.

Score-based MLLM-IQA methods focus on predicting continuous quality scores while leveraging reasoning mechanisms for robustness. In particular, Q-Align [46] discretizes subjective scores to a one-hot label to emulate the human judgment process. In contrast, DeQA-Score [45] improves quality prediction by modeling score distributions as soft labels, and Q-Scorer [99] explicitly adapts multimodal representations for direct scalar quality prediction. In addition, Dog-IQA [196] introduces a standard-guided discrete scoring mechanism combined with mix-grained global–local quality aggregation. To address label efficiency, LEAF [202] decouples perceptual knowledge from MOS calibration via distillation.

On the other hand, the comparison-derived scoring framework Compare2Score [95] achieves quality scoring by inferring continuous scores from adaptive pairwise comparisons. Q-Insight [197] integrates reinforcement learning with preference modeling, whereas Q-Ponder [203] introduces explicit joint optimization objectives for both scoring and reasoning. Furthermore, RALI [208] treats reinforcement learning-based IQA as a reasoning-driven representation compression process and proposes a lightweight framework that preserves scoring accuracy and cross-domain generalization without requiring LLM inference. Reliability-aware modeling is explored by Q-Hawkeye [200], incorporating

 doi:10.20944/preprints202603.2366.v1

uncertainty-aware policy optimization. In domain-specific settings, CAP-IQA [204] integrates text priors and image-specific context for task-aware quality scoring, demonstrating reliable performance.

In parallel, Q-Instruct [94] and Q-Bench [206,207] focus on improving and evaluating low-level perceptual reasoning capabilities of multimodal language models, underscoring the importance of instruction tuning and benchmarking. Additionally, the AgenticIQA framework [48] represents a paradigm that treats perceptual quality assessment as a structured decision-making process rather than a single-pass prediction task. While MLLM-based IQA offers enhanced interpretability, flexibility, and perceptual reasoning capacity, challenges remain in score calibration, reasoning consistency, and computational efficiency.

Comparison: Conventional vs. VLM-/MLLM-based Evaluation. Conventional IQA metrics and language-driven evaluation reflect different yet complementary assessment paradigms. Classical full-reference measures (e.g., PSNR, SSIM) quantify pixel fidelity and structural consistency, providing deterministic and reproducible criteria. These metrics remain essential for benchmarking reconstruction accuracy and optimization stability.

By contrast, VLM-/MLLM-based evaluation estimates quality through multimodal similarity modeling, preference reasoning, and language-conditioned quality interpretation. Instead of directly measuring pixel fidelity, these approaches estimate quality through mechanisms such as prompt-driven image–text alignment, pairwise comparison, distribution-aware score modeling, and reasoning-guided scoring. Such evaluation strategies are particularly relevant for generative restoration, mixed degradations, and no-reference scenarios. Unlike conventional metrics, language-driven approaches may favor semantic plausibility or contextual consistency.

Importantly, language-driven evaluation should not be interpreted as a direct replacement for conventional IQA metrics, but rather as a complementary perceptual assessment mechanism. Conventional metrics ensure numerical stability and comparability, whereas language-driven evaluators capture complementary perceptual and semantic cues. Recent studies [37,114,132] increasingly adopt hybrid evaluation protocols that combine deterministic fidelity measures with language-driven perceptual assessment to achieve more comprehensive performance evaluation.

Evaluation Criteria and Auxiliary Paradigms. In addition to direct quality prediction metrics, correlation-based criteria such as PLCC, SRCC, and KRCC [209] are widely adopted to measure the consistency between predicted scores and human subjective ratings. These statistics do not evaluate image quality directly but instead quantify the reliability of IQA models with respect to human perception. Additionally, text-based metrics such as BLEU-N [211] and ROUGE-L [212] are often used to evaluate the semantic fidelity of generated descriptions and reference captions, especially in instruction-following or explanation-based restoration frameworks.

Furthermore, several studies adopt task-oriented evaluation paradigms that assess image quality indirectly through downstream performance [98]. In this setting, restored images are treated as inputs to high-level vision systems, such as detection, segmentation, or recognition models, where improvements in task performance serve as proxies for quality enhancement. The underlying assumption is that degradations often impair feature extraction and perception reliability, and thus higher downstream accuracy may indicate improved visual quality. Nevertheless, this assumption is not universally valid. Prior studies [216] have shown that perceptual quality and downstream performance are not strictly causally linked. Despite these limitations, task-driven evaluation remains particularly relevant in application-critical domains, including medical imaging and autonomous driving, where restoration quality is ultimately defined by its functional impact on subsequent perception or decision-making processes.

Metrics for Language-driven IQA Ability. Leveraging the strong semantic understanding capabilities of MLLMs and VLMs for IQA has emerged as a promising and rapidly evolving research direction. Usually, correlation-based metrics, such as PLCC and SRCC, are widely used, which quantify agreement between predicted scores and human subjective ratings. However, the systematic evaluation of the IQA capability of such models remains limited. 2AFC [217] represents an early attempt by

 doi:10.20944/preprints202603.2366.v1

proposing consistency, accuracy, and correlation metrics to analyze judgment robustness and alignment with human opinion scores. Despite this initial effort, the evaluation of VLM-/MLLM-based IQA remains largely underexplored, leaving substantial room for further investigation.

4.3. Experimental Results

To better understand the effectiveness of different design choices, we report experimental results of several VLM-/MLLM-based all-in-one restoration frameworks under three commonly used experimental settings. Table 3 presents results on three degradations: deraining, denoising ($\sigma = 15, 25, 50$), and dehazing. Meanwhile, Table 4 extends the evaluation to five degradations by additionally including deblurring and LLIE. Table 5 further reports performance on the single LLIE task. All results are reported in terms of PSNR and SSIM.

Under the three-degradation setting (Table 3), language-driven restoration frameworks exhibit comparable overall performance. For instance, VLU-Net [67] achieves the highest deraining PSNR (38.93 dB), whereas ClearAIR [66] exhibits more consistent cross-task behavior, maintaining stable performance across deraining, denoising, and dehazing. A similar trend is observed in the five-degradation setting (Table 4), where ClearAIR achieves the best average performance compared to other language-driven approaches. In contrast, on the single-task LLIE benchmark (Table 5), task-specific non-language-driven methods such as RetinexFormer [30] (25.16 dB) and RetinexDiff++ [218] (24.67 dB) outperform most language-driven models in terms of PSNR.

Table 3. Comparison with state-of-the-art methods on three image restoration tasks. The top rows correspond to non-language-driven approaches, while the bottom rows represent VLM-/MLLM-based methods. Performance is reported in terms of PSNR and SSIM for each dataset.

Method	Venue	Params	Deraining		Denoising (BSD68 [137])						Dehazing		Average	
			Rain100L [143]		$\sigma = 15$		$\sigma = 25$		$\sigma = 50$		SOTS [152]			
AirNet [219]	CVPR'22	9M	34.90	0.967	33.92	0.933	31.26	0.888	28.00	0.797	27.94	0.962	31.20	0.910
IDR [220]	CVPR'23	15M	36.03	0.971	33.89	0.931	31.32	0.884	28.04	0.798	29.87	0.970	31.83	0.911
PromptIR [33]	NeurIPS'23	33M	36.37	0.972	33.98	0.933	31.31	0.888	28.06	0.799	30.58	0.974	32.06	0.913
AdaIR [221]	ICLR'25	29M	38.64	0.983	34.12	0.934	31.45	0.892	28.19	0.802	31.06	0.980	32.69	0.918
DSwinIR [222]	T-PAMI'25	24M	37.73	0.983	34.12	0.933	31.59	0.890	28.31	0.803	31.86	0.980	32.72	0.917
VIVNet [223]	T-PAMI'26	7.42M	38.47	0.983	34.16	0.936	31.50	0.893	28.24	0.806	32.19	0.982	32.91	0.920
InstructIR-3D [34]	ECCV'24	16M	37.98	0.978	34.15	0.933	31.52	0.890	28.30	0.803	30.22	0.959	32.43	0.913
VLU-Net [67]	CVPR'25	35M	38.93	0.984	34.13	0.935	31.48	0.892	28.23	0.804	30.71	0.980	32.70	0.919
Perceive-IR [122]	T-IP'25	42M	38.29	0.980	34.13	0.934	31.53	0.890	28.31	0.804	30.87	0.975	32.63	0.917
ClearAIR [66]	AAAI'26	31M	38.61	0.984	34.18	0.935	31.50	0.891	28.31	0.804	31.08	0.981	32.74	0.919

Table 4. Comparison with state-of-the-art methods on five image restoration tasks. The top rows correspond to non-language-driven approaches, while the bottom rows represent VLM-/MLLM-based methods. Performance is reported in terms of PSNR and SSIM for each dataset.

Method	Venue	Params	Dehazing		Deraining		Denoising		Deblurring		LLIE		Average	
			SOTS [152]		Rain100L [143]		$BSD68_{\sigma=25}$ [137]		GoPro [15]		LOL [17]			
AirNet [219]	CVPR'22	9M	21.04	0.884	32.98	0.951	30.91	0.882	24.35	0.781	18.18	0.735	25.49	0.847
IDR [220]	CVPR'23	15M	25.24	0.943	35.63	0.965	31.60	0.887	27.87	0.846	21.34	0.826	28.34	0.893
PromptIR [33]	NeurIPS'23	33M	26.54	0.949	36.37	0.970	31.47	0.886	28.71	0.881	22.68	0.832	29.15	0.904
AdaIR [221]	ICLR'25	29M	30.53	0.978	38.02	0.981	31.35	0.888	28.12	0.858	23.00	0.845	30.20	0.910
DSwinIR [222]	T-PAMI'25	24M	30.09	0.975	37.77	0.982	31.34	0.885	29.17	0.879	22.64	0.843	30.20	0.913
VIVNet [223]	T-PAMI'26	7.42M	31.85	0.982	38.67	0.984	31.46	0.892	28.50	0.866	23.03	0.857	30.70	0.916
DA-CLIP [112]	ICLR'24	125M	26.28	0.939	35.91	0.972	25.77	0.653	28.81	0.882	22.57	0.832	29.23	0.898
DiffRes [224]	CVPR'25	45M	27.23	0.958	37.25	0.979	32.07	0.890	29.33	0.883	23.13	0.843	29.78	0.911
InstructIR-5D [34]	ECCV'24	16M	27.10	0.956	36.84	0.973	31.40	0.887	29.40	0.886	23.00	0.836	29.55	0.907
VLU-Net [67]	CVPR'25	35M	30.84	0.980	38.54	0.982	31.43	0.891	27.46	0.840	22.29	0.833	30.11	0.905
Perceive-IR [122]	T-IP'25	42M	28.19	0.964	37.25	0.977	31.44	0.887	29.46	0.886	22.88	0.833	29.84	0.909
ClearAIR [66]	AAAI'26	31M	30.12	0.978	38.20	0.982	31.53	0.888	29.67	0.887	22.83	0.846	30.45	0.916

 doi:10.20944/preprints202603.2366.v1

Table 5. Performance comparison with state-of-the-art approaches on the LOL-v1 [17] dataset.

Method	Venue	PSNR	SSIM
RetinexFormer [30]	ICCV'23	25.16	0.845
LLFormer [225]	AAAI'23	23.65	0.8163
CWNet [226]	ICCV'25	23.60	0.8496
RetinexDiff++ [218]	T-PAMI'25	24.67	0.867
LLMRA [68]	ECCV'24	23.30	0.846
DA-CLIP [112]	ICLR'24	23.40	0.811
DiffRes [224]	CVPR'25	24.55	0.839
Perceive-IR [122]	T-IP'25	23.79	0.841

These observations suggest that, although language-driven frameworks enhance flexibility and generalization in multi-degradation scenarios, their advantages in specialized tasks remain limited. Moreover, gains in pixel-level fidelity (e.g., PSNR) remain limited. This is because language-driven methods focus more on semantic alignment and user intent than on distortion minimization. This leads to a mismatch between current evaluation protocols and the goals of language-driven restoration. Moreover, language-driven IQA metrics are rarely included in existing benchmarks, highlighting the need for evaluation frameworks that jointly consider fidelity, semantic correctness, and user-oriented restoration quality.

5. Discussion and Open Challenges

Although the integration of MLLMs and VLMs has driven the development of IR frameworks, it has introduced new capabilities while simultaneously giving rise to additional challenges. In this section, we analyze key open challenges in VLM-/MLLM-based methods, focusing on generalization, computational efficiency, cross-paradigm trade-offs, evaluation reliability, dataset design, high-dimensional representations, and trustworthiness. We further discuss potential research directions that may help address these challenges and inform future developments in language-driven restoration systems.

5.1. Generalization and Robustness

In real-world scenarios, degradations are often complex, mixed, or poorly defined, making generalization and robustness persistent challenges for IR, even when incorporating semantic priors and degradation-aware guidance via MLLMs and VLMs. Specifically, while VLM- or MLLM-derived representations provide high-level semantic context, they do not fully eliminate sensitivity to degradation variations. Second, language-driven frameworks frequently rely on textual descriptions, prompt formulations, or semantic interpretations to characterize degradation types [68,123,134]. Such dependencies may lead to inconsistent restoration behaviors due to linguistic variability and prompt sensitivity. Moreover, language models themselves may exhibit hallucinations and domain biases inherited from large-scale pretraining, which can affect degradation understanding and guidance reliability.

One potential solution involves prompt-invariant modeling, such as template-based or structured prompts, to reduce instability caused by diverse linguistic formulations. Another direction is uncertainty-aware restoration, where degradation cues are accompanied by confidence estimates to improve robustness against hallucination-induced errors. Despite these emerging strategies, developing principled mechanisms for handling degradation ambiguity, prompt variability, and language-model uncertainty remains an important challenge for future research.

5.2. Computational Efficiency

While VLM-/MLLM-based IR frameworks introduce enhanced semantic awareness and perceptual reasoning capabilities, computational efficiency remains a central challenge. The computational

 doi:10.20944/preprints202603.2366.v1

burden arises from multiple sources. First, pretrained VLMs and MLLMs typically contain a large number of parameters, making semantic feature extraction and cross-modal reasoning inherently expensive. Additional conditioning and reasoning mechanisms further increase memory consumption. Moreover, agentic paradigms often require multiple reasoning iterations, candidate evaluations, or tool invocation loops, leading to high latency and inference cost.

Recent studies have explored several strategies to mitigate these efficiency bottlenecks. For example, policy optimization [132] and dynamic routing mechanisms [126] have been applied to reduce unnecessary reasoning and model invocation steps. Hybrid designs [129] further improve efficiency by restricting expensive language-driven computation to critical stages. Despite these advances, computational efficiency remains a key challenge for practical deployment. Future research may focus on developing compact multimodal models tailored for restoration tasks as well as reducing redundancy in cross-modal representations through mechanisms such as knowledge distillation or token reduction. In addition, resolution-adaptive architectures present a promising direction, enabling language-driven reasoning to operate at coarse semantic scales while preserving high-resolution reconstruction within lightweight visual backbones.

5.3. Cross-Paradigm Trade-Offs

Different restoration paradigms rely on distinct design assumptions that introduce structural trade-offs. Language-driven supervision signals may conflict with low-level reconstruction constraints. Specifically, perceptual losses or evaluator-based rewards derived from VLMs/MLLMs typically operate at a global semantic level and may exhibit limited sensitivity to local structural artifacts. As a result, optimization may favor high-level consistency while under-constraining fine-grained visual fidelity. Moreover, paradigms also differ in balancing capability and system complexity. MLLM-centered frameworks enable flexible degradation interpretation and dynamic control but incur increased deployment overhead, whereas VLM-guided approaches preserve relatively structural simplicity with more limited decision-level flexibility.

5.4. Evaluation Reliability

In contrast to traditional evaluation metrics, language-driven IQA models do not always produce numerically stable or strictly consistent scores. Instead, they often operate in a semantic assessment space, where quality judgments are expressed through preferences, rankings, or linguistic interpretations [96,198]. Recently, several studies [45,99] have explored MLLMs as quantitative quality scorers. While such models demonstrate promising alignment with human perception in many scenarios, their evaluation behavior may exhibit variability and uncertainty. In particular, assessment results can be sensitive to prompt formulations, where small variations in phrasing or textual context may lead to inconsistent quality predictions. This prompt sensitivity introduces challenges for reproducibility and comparability across evaluation settings. Moreover, score calibration remains a nontrivial challenge, as language models are not explicitly optimized for metric-level numerical stability. This variability becomes particularly critical when language-based IQA outputs are directly used for model selection or benchmark comparisons.

The coexistence of conventional IQA metrics and language-driven evaluators introduces new considerations regarding evaluation reliability. Future research may explore standardized prompting protocols, confidence-aware evaluation mechanisms, and hybrid assessment frameworks that combine deterministic metrics with language-driven quality reasoning. Nevertheless, ensuring stable and comparable evaluation remains an open problem.

5.5. Dataset Design for VLM-/MLLM-Based IR

Numerous datasets have been proposed for various restoration tasks. However, several limitations still remain in existing datasets. Current restoration benchmarks rarely incorporate structured linguistic annotations describing degradation characteristics or perceptual attributes. While MLLMs and VLMs benefit from large-scale pretraining and strong semantic understanding, they may still exhibit failure

 doi:10.20944/preprints202603.2366.v1

modes in complex or ambiguous scenarios. Consequently, language-driven restoration frameworks often rely on synthetic prompts [110] or automatically generated descriptions [56,68], which can introduce semantic inconsistencies. This highlights the importance of reliable textual supervision for improving model stability and performance. Moreover, most restoration datasets are constructed under predefined degradation models, which may not adequately capture the complexity or compositional nature of real-world degradations. Such limitations may restrict the robustness and generalization capability of language-integrated restoration systems.

Future dataset construction may benefit from multi-level annotations beyond pixel-level ground truth, including degradation semantics, perceptual quality descriptions, and task-oriented linguistic guidance. Furthermore, developing standardized degradation annotation protocols represents a critical research direction for improving cross-modal consistency while reducing annotation ambiguity. Furthermore, incorporating diverse, composed, and open-world degradations may enhance model robustness under realistic conditions.

5.6. Leveraging Multimodal Data and High-Dimensional Representations

MLLMs and VLMs enable IR frameworks to incorporate textual priors and cross-modal semantics, opening new opportunities to exploit richer data modalities. However, existing studies mainly focus on RGB images, while other sensing modalities remain largely underexplored, including infrared imagery, event-based data, and depth measurements [227,228]. Integrating heterogeneous sensory inputs may improve the interpretation of degradation, structural reasoning, and contextual consistency. Therefore, extending language-driven restoration frameworks to these modalities presents a promising research direction.

Beyond static images, applying VLM-/MLLM-based restoration paradigms to high-dimensional data, such as videos, represents another important direction for future research. Video restoration inherently requires not only accurate frame-level reconstruction but also temporal consistency across frames. In this context, language-driven reasoning mechanisms may provide complementary benefits by scene-level coherence. For example, MLLMs may assist in interpreting dynamic degradation patterns [229] or evaluating video quality to support invoking tools.

5.7. Ethics and Trustworthiness

Ethical considerations and trustworthiness have become increasingly relevant in VLM-/MLLM-based IR frameworks. Certain approaches rely on online inference services, which require transmitting visual data for non-local processing [130]. Such designs may raise privacy and data security concerns, particularly in sensitive scenarios including medical imaging, unmanned aerial vehicle (UAV), and autonomous driving.

Beyond data privacy, the reliability of language-driven components introduces additional challenges. MLLMs are known to exhibit hallucinations, reasoning inconsistencies, and biases inherited from large-scale training data. When integrated into restoration pipelines, these factors may result in incorrect degradation interpretation, unstable guidance signals, or semantically inconsistent restoration outcomes. Unlike conventional restoration errors, failures caused by MLLMs or VLMs may be less predictable and more difficult to attribute.

Future research may investigate privacy-preserving multimodal restoration frameworks and robustness against hallucination-induced errors. In addition, the development of locally deployable multimodal models [72] represents a promising direction for mitigating privacy concerns while preserving semantic reasoning capabilities.

6. Conclusions

In this survey, we provided a structured review of language-driven IR frameworks and IQA methods based on MLLMs and VLMs. We introduced a unified taxonomy that characterizes how language models interact with restoration pipelines, encompassing paradigms such as representation-level, optimization-level, design-level, and hybrid coupling. Through this perspective, we analyzed

 doi:10.20944/preprints202603.2366.v1

representative methodologies and architectural designs, highlighting the evolving role of language information in IR. In addition, we reviewed emerging VLM-/MLLM-based IQA approaches and discussed their distinctions from conventional evaluation metrics. We also summarized widely used restoration datasets and compared state-of-the-art VLM-/MLLM-based methods across different restoration tasks.

The rapid progress of language-integrated restoration systems also raises fundamental challenges, including cross-paradigm trade-offs, generalization under complex degradations, computational efficiency, evaluation reliability, dataset construction, and trustworthiness. These challenges highlight that language-driven restoration is not merely an architectural extension of traditional frameworks, but rather a paradigm shift that raises new methodological and system-level questions. We further anticipate future research directions: exploring uncertainty-aware restoration mechanisms, more efficient semantic-guided architectures, and scalable annotation strategies for image–text paired datasets. Extending language-driven restoration paradigms to heterogeneous modalities and high-dimensional data also presents promising opportunities.

As language and vision models continue to develop, their integration is expected to play an increasingly important role in bridging low-level visual restoration with high-level semantic understanding. Overall, this survey aims to provide a comprehensive reference for understanding the evolving design principles of language-driven IR and to encourage further investigation into this rapidly developing research direction.

References

1. Jiang, B.; Li, J.; Lu, Y.; Cai, Q.; Song, H.; Lu, G. Efficient image denoising using deep learning: A brief survey. *Information Fusion* **2025**, p. 103013.
2. Tian, C.; Xu, Y.; Zuo, W. Image denoising using deep CNN with batch renormalization. *Neural Networks* **2020**, *121*, 461–473.
3. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing* **2017**, *26*, 3142–3155.
4. Chen, X.; Pan, J.; Dong, J.; Tang, J. Towards unified deep image deraining: A survey and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
5. Chen, X.; Li, H.; Li, M.; Pan, J. Learning a sparse transformer network for effective image deraining. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 5896–5905.
6. Xiao, J.; Fu, X.; Liu, A.; Wu, F.; Zha, Z.J. Image de-raining transformer. *IEEE transactions on pattern analysis and machine intelligence* **2022**, *45*, 12978–12995.
7. Gui, J.; Cong, X.; Cao, Y.; Ren, W.; Zhang, J.; Zhang, J.; Cao, J.; Tao, D. A comprehensive survey and taxonomy on single image dehazing based on deep learning. *ACM Computing Surveys* **2023**, *55*, 1–37.
8. Tsai, F.J.; Peng, Y.T.; Lin, Y.Y.; Lin, C.W. PHATNet: A Physics-guided Haze Transfer Network for Domain-adaptive Real-world Image Dehazing. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 5591–5600.
9. Yu, H.; Huang, J.; Zheng, K.; Zhao, F. High-quality image dehazing with diffusion model. *arXiv preprint arXiv:2308.11949* **2023**.
10. Quan, Y.; Tan, X.; Huang, Y.; Xu, Y.; Ji, H. Image desnowing via deep invertible separation. *IEEE Transactions on Circuits and Systems for Video Technology* **2023**, *33*, 3133–3144.
11. Guo, X.; Wang, X.; Fu, X.; Zha, Z.J. Deep unfolding network for image desnowing with snow shape prior. *IEEE Transactions on Circuits and Systems for Video Technology* **2025**.
12. Liu, Y.F.; Jaw, D.W.; Huang, S.C.; Hwang, J.N. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing* **2018**, *27*, 3064–3073.
13. Xiang, Y.; Zhou, H.; Li, C.; Sun, F.; Li, Z.; Xie, Y. Deep learning in motion deblurring: current status, benchmarks and future prospects. *The Visual Computer* **2025**, *41*, 3801–3827.
14. Abuolaim, A.; Brown, M.S. Defocus deblurring using dual-pixel data. In Proceedings of the European conference on computer vision. Springer, 2020, pp. 111–126.

 doi:10.20944/preprints202603.2366.v1

15. Nah, S.; Hyun Kim, T.; Mu Lee, K. Deep multi-scale convolutional neural network for dynamic scene deblurring. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3883–3891.
16. Li, C.; Guo, C.; Han, L.; Jiang, J.; Cheng, M.M.; Gu, J.; Loy, C.C. Low-light image and video enhancement using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *44*, 9396–9416.
17. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep Retinex Decomposition for Low-Light Enhancement. In Proceedings of the BMVC, 2018.
18. Yang, W.; Wang, W.; Huang, H.; Wang, S.; Liu, J. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing* **2021**, *30*, 2072–2086.
19. Wang, Z.; Chen, J.; Hoi, S.C. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 3365–3387.
20. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1905–1914.
21. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating more pixels in image super-resolution transformer. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 22367–22377.
22. Zhang, W.; Dong, L.; Pan, X.; Zou, P.; Qin, L.; Xu, W. A survey of restoration and enhancement for underwater images. *IEEE Access* **2019**, *7*, 182259–182279.
23. Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An underwater image enhancement benchmark dataset and beyond. *IEEE transactions on image processing* **2019**, *29*, 4376–4389.
24. Islam, M.J.; Xia, Y.; Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE robotics and automation letters* **2020**, *5*, 3227–3234.
25. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell* **2018**, *172*, 1122–1131.
26. McCollough, C.H.; Bartley, A.C.; Carter, R.E.; Chen, B.; Drees, T.A.; Edwards, P.; Holmes III, D.R.; Huang, A.E.; Khan, F.; Leng, S.; et al. Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. *Medical physics* **2017**, *44*, e339–e352.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
29. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In Proceedings of the First conference on language modeling, 2024.
30. Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; Zhang, Y. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 12504–12513.
31. Jiang, J.; Zuo, Z.; Wu, G.; Jiang, K.; Liu, X. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
32. Li, R.; Tan, R.T.; Cheong, L.F. All in one bad weather removal using architectural search. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3175–3185.
33. Potlapalli, V.; Zamir, S.W.; Khan, S.H.; Shahbaz Khan, F. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems* **2023**, *36*, 71275–71293.
34. Conde, M.V.; Geigle, G.; Timofte, R. Instructir: High-quality image restoration following human instructions. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 1–21.
35. Guan, C.; Yoshie, O. CLIP-driven rain perception: Adaptive deraining with pattern-aware network routing and mask-guided cross-attention. *arXiv preprint arXiv:2506.01366* **2025**.
36. Lin, Y.; Lin, Z.; Chen, H.; Pan, P.; Li, C.; Chen, S.; Wen, K.; Jin, Y.; Li, W.; Ding, X. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 22369–22380.
37. Zhu, K.; Gu, J.; You, Z.; Qiao, Y.; Dong, C. An intelligent agentic system for complex image restoration problems. *arXiv preprint arXiv:2410.17809* **2024**.

 doi:10.20944/preprints202603.2366.v1

38. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **2004**, *13*, 600–612.
39. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electronics letters* **2008**, *44*, 800–801.
40. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
41. Wang, J.; Chan, K.C.; Loy, C.C. Exploring clip for assessing the look and feel of images. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2023, Vol. 37, pp. 2555–2563.
42. Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In Proceedings of the Proceedings of the 2021 conference on empirical methods in natural language processing, 2021, pp. 7514–7528.
43. Agnolucci, L.; Galteri, L.; Bertini, M. Quality-aware image-text alignment for opinion-unaware image quality assessment. *arXiv preprint arXiv:2403.11176* **2024**.
44. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 34892–34916.
45. You, Z.; Cai, X.; Gu, J.; Xue, T.; Dong, C. Teaching large language models to regress accurate image quality scores using score distribution. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 14483–14494.
46. Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090* **2023**.
47. Zhang, Z.; Wu, H.; Jia, Z.; Lin, W.; Zhai, G. Teaching lmms for image quality scoring and interpreting. *arXiv preprint arXiv:2503.09197* **2025**.
48. Zhu, H.; Tian, Y.; Ding, K.; Chen, B.; Chen, B.; Wang, S.; Lin, W. Agenticiqa: An agentic framework for adaptive and interpretable image quality assessment. *arXiv preprint arXiv:2509.26006* **2025**.
49. Su, J.; Xu, B.; Yin, H. A survey of deep learning approaches to image restoration. *Neurocomputing* **2022**, *487*, 46–65.
50. Wang, L.; Zhou, W.; Wang, C.; Lam, K.M.; Su, Z.; Pan, J. Deep Learning-Driven Ultra-High-Definition Image Restoration: A Survey. *arXiv preprint arXiv:2505.16161* **2025**.
51. Zhang, J.; Huang, J.; Jin, S.; Lu, S. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence* **2024**, *46*, 5625–5644.
52. Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In Proceedings of the Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 6418–6428.
53. Xu, J.; Li, H.; Liang, Z.; Zhang, D.; Zhang, L. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603* **2018**.
54. Guo, Y.; Xiao, X.; Chang, Y.; Deng, S.; Yan, L. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 12097–12107.
55. Ancuti, C.O.; Ancuti, C.; Timofte, R. NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 444–445.
56. Lai, J.; Chen, S.; Lin, Y.; Ye, T.; Liu, Y.; Fei, S.; Xing, Z.; Wu, H.; Wang, W.; Zhu, L. SnowMaster: Comprehensive Real-world Image Desnowing via MLLM with Multi-Model Feedback Optimization. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 4302–4312.
57. Hai, J.; Xuan, Z.; Yang, R.; Hao, Y.; Zou, F.; Lin, F.; Han, S. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation* **2023**, *90*, 103712.
58. Zuo, Y.; Zheng, Q.; Wu, M.; Jiang, X.; Li, R.; Wang, J.; Zhang, Y.; Mai, G.; Wang, L.V.; Zou, J.; et al. 4kagent: agentic any image to 4k super-resolution. *arXiv preprint arXiv:2507.07105* **2025**.
59. Berman, D.; Levy, D.; Avidan, S.; Treibitz, T. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 2822–2837.
60. Knoll, F.; Zbontar, J.; Sriram, A.; Muckley, M.J.; Bruno, M.; Defazio, A.; Parente, M.; Geras, K.J.; Katsnelson, J.; Chandarana, H.; et al. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. *Radiology: Artificial Intelligence* **2020**, *2*, e190007.

 doi:10.20944/preprints202603.2366.v1

61. Kong, X.; Dong, C.; Zhang, L. Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy. *arXiv preprint arXiv:2401.03379* **2024**.
62. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17683–17693.
63. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5728–5739.
64. Liu, M.; Cui, Y.; Liu, X.; Strand, L.; Yin, H.; Knoll, A. Drfir: A dimensionality reduction framework for all-in-one image restoration in spatial and frequency domains. *Expert Systems with Applications* **2025**, p. 128959.
65. Zhang, X.; Zhang, H.; Wang, G.; Zhang, Q.; Zhang, L.; Du, B. UniUIR: Considering Underwater Image Restoration as An All-in-One Learner. *arXiv preprint arXiv:2501.12981* **2025**.
66. Zhang, X.; Zhang, H.; Wang, G.; Zhang, Q.; Zhang, L. ClearAIR: A Human-Visual-Perception-Inspired All-in-One Image Restoration. *arXiv preprint arXiv:2601.02763* **2026**.
67. Zeng, H.; Wang, X.; Chen, Y.; Su, J.; Liu, J. Vision-Language Gradient Descent-driven All-in-One Deep Unfolding Networks. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 7524–7533.
68. Jin, X.; Shi, Y.; Xia, B.; Yang, W. Llmra: Multi-modal large language model based restoration assistant. *arXiv preprint arXiv:2401.11401* **2024**.
69. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
70. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* **2024**.
71. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen technical report. *arXiv preprint arXiv:2309.16609* **2023**.
72. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* **2024**.
73. Luo, H.; Bao, J.; Wu, Y.; He, X.; Li, T. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 23033–23044.
74. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 4904–4916.
75. Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; Xu, H. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems* **2022**, *35*, 9125–9138.
76. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
77. Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106* **2025**.
78. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* **2024**.
79. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* **2025**.
80. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* **2023**.
81. Zhai, G.; Min, X. Perceptual image quality assessment: a survey. *Science China Information Sciences* **2020**, *63*, 211301.
82. Ding, K.; Ma, K.; Wang, S.; Simoncelli, E.P. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *44*, 2567–2581.

 doi:10.20944/preprints202603.2366.v1

83. Zheng, H.; Yang, H.; Fu, J.; Zha, Z.J.; Luo, J. Learning conditional knowledge distillation for degraded-reference image quality assessment. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10242–10251.
84. Lao, S.; Gong, Y.; Shi, S.; Yang, S.; Wu, T.; Wang, J.; Xia, W.; Yang, Y. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1140–1149.
85. Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; Lin, W. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing* **2024**, *33*, 2404–2418.
86. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* **2012**, *21*, 4695–4708.
87. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* **2012**, *20*, 209–212.
88. Venkatanath, N.; Praneeth, D.; Sumohana, S.C.; Swarup, S.M.; et al. Blind image quality evaluation using perception based features. In Proceedings of the 2015 twenty first national conference on communications (NCC). IEEE, 2015, pp. 1–6.
89. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. Musiq: Multi-scale image quality transformer. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 5148–5157.
90. Talebi, H.; Milanfar, P. NIMA: Neural image assessment. *IEEE transactions on image processing* **2018**, *27*, 3998–4011.
91. Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; Yang, Y. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1191–1200.
92. Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3667–3676.
93. Kang, L.; Ye, P.; Li, Y.; Doermann, D. Convolutional neural networks for no-reference image quality assessment. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1733–1740.
94. Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 25490–25500.
95. Zhu, H.; Wu, H.; Li, Y.; Zhang, Z.; Chen, B.; Zhu, L.; Fang, Y.; Zhai, G.; Lin, W.; Wang, S. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems* **2024**, *37*, 32611–32629.
96. You, Z.; Li, Z.; Gu, J.; Yin, Z.; Xue, T.; Dong, C. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 259–276.
97. Chen, Z.; Wang, J.; Wang, W.; Xu, S.; Xiong, H.; Zeng, Y.; Guo, J.; Wang, S.; Yuan, C.; Li, B.; et al. Seagull: No-reference image quality assessment for regions of interest via vision-language instruction tuning. *arXiv preprint arXiv:2411.10161* **2024**.
98. Chen, C.; Yang, S.; Wu, H.; Liao, L.; Zhang, Z.; Wang, A.; Sun, W.; Yan, Q.; Lin, W. Q-ground: Image quality grounding with large multi-modality models. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 486–495.
99. Tang, Z.; Yang, S.; Peng, B.; Wang, Z.; Dong, J. Revisiting MLLM Based Image Quality Assessment: Errors and Remedy. *arXiv preprint arXiv:2511.07812* **2025**.
100. Li, X.; Ren, Y.; Jin, X.; Lan, C.; Wang, X.; Zeng, W.; Wang, X.; Chen, Z. Diffusion models for image restoration and enhancement: a comprehensive survey. *International Journal of Computer Vision* **2025**, *133*, 8078–8108.
101. Cheng, J.; Liang, D.; Tan, S. Transfer clip for generalizable image denoising. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 25974–25984.
102. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.

 doi:10.20944/preprints202603.2366.v1

103. Yang, H.; Pan, L.; Yang, Y.; Hartley, R.; Liu, M. Ldp: Language-driven dual-pixel image defocus deblurring network. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24078–24087.
104. Chen, Z.; Chen, T.; Wang, C.; Gao, Q.; Niu, C.; Wang, G.; Shan, H. Low-dose CT denoising with language-engaged dual-space alignment. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2024, pp. 3088–3091.
105. Chen, Z.; Chen, T.; Wang, C.; Gao, Q.; Xie, H.; Niu, C.; Wang, G.; Shan, H. LangMamba: A Language-driven Mamba Framework for Low-dose CT Denoising with Vision-language Models. *IEEE Transactions on Radiation and Plasma Medical Sciences* **2025**.
106. Hu, B.; Liu, H.; Zheng, Z.; Liu, P. CLIP-SR: Collaborative Linguistic and Image Processing for Super-Resolution. *IEEE Transactions on Multimedia* **2025**.
107. Duan, H.; Min, X.; Wu, S.; Shen, W.; Zhai, G. Uniprocessor: a text-induced unified low-level image processor. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 180–199.
108. Li, J.; Wang, Y.; Yan, J.; Zhang, R.; Yang, B. MdaIF: Robust One-Stop Multi-Degradation-Aware Image Fusion with Language-Driven Semantics. *arXiv preprint arXiv:2511.12525* **2025**.
109. Mao, J.; Yang, Y.; Yin, X.; Shao, L.; Tang, H. AllRestorer: All-in-One Transformer for Image Restoration under Composite Degradations. *arXiv preprint arXiv:2411.10708* **2024**.
110. Wu, Z.; Chen, Y.; Yokoya, N.; He, W. MP-HSIR: A Multi-Prompt Framework for Universal Hyperspectral Image Restoration. *arXiv preprint arXiv:2503.09131* **2025**.
111. Lee, C.M.; Cheng, C.H.; Lin, Y.F.; Cheng, Y.C.; Liao, W.T.; Hsu, C.C.; Yang, F.E.; Wang, Y.C.F. Prompthsi: Universal hyperspectral image restoration framework for composite degradation. *arXiv e-prints* **2024**, pp. arXiv–2411.
112. Luo, Z.; Gustafsson, F.K.; Zhao, Z.; Sjölund, J.; Schön, T.B. Controlling vision-language models for multi-task image restoration. *arXiv preprint arXiv:2310.01018* **2023**.
113. Jiang, Y.; Zhang, Z.; Xue, T.; Gu, J. Autodir: Automatic all-in-one image restoration with latent diffusion. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 340–359.
114. Qi, C.; Tu, Z.; Ye, K.; Delbracio, M.; Milanfar, P.; Chen, Q.; Talebi, H. Spire: Semantic prompt-driven image restoration. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 446–464.
115. Ai, Y.; Huang, H.; Zhou, X.; Wang, J.; He, R. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 25432–25444.
116. Tu, Y.; Yan, Q.; Niu, A.; Tang, J. TPGDiff: Hierarchical Triple-Prior Guided Diffusion for Image Restoration. *arXiv preprint arXiv:2601.20306* **2026**.
117. Zhang, Z.; Lei, J.; Peng, B.; Zhu, J.; Xu, L.; Huang, Q. Advancing Real-World Stereoscopic Image Super-Resolution via Vision-Language Model. *IEEE Transactions on Image Processing* **2025**.
118. Yang, S.; Ding, M.; Wu, Y.; Li, Z.; Zhang, J. Implicit neural representation for cooperative low-light image enhancement. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 12918–12927.
119. Li, B.; Li, X.; Zhu, H.; Jin, Y.; Feng, R.; Zhang, Z.; Chen, Z. Sed: Semantic-aware discriminator for image super-resolution. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 25784–25795.
120. Song, W.; Liu, C.; Di Mauro, M.; Liotta, A. Unsupervised Underwater Image Enhancement Combining Imaging Restoration and Prompt Learning. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer, 2024, pp. 421–434.
121. Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; He, K. Scaling language-image pre-training via masking. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 23390–23400.
122. Zhang, X.; Ma, J.; Wang, G.; Zhang, Q.; Zhang, H.; Zhang, L. Perceive-ir: Learning to perceive degradation better for all-in-one image restoration. *IEEE Transactions on Image Processing* **2025**.
123. Sun, X.; Wang, L.; Wang, C.; Jin, Y.; Lam, K.m.; Su, Z.; Yang, Y.; Pan, J. Adapting Large VLMs with Iterative and Manual Instructions for Generative Low-light Enhancement. *arXiv preprint arXiv:2507.18064* **2025**.
124. Zhou, H.; Dong, W.; Liu, X.; Zhang, Y.; Zhai, G.; Chen, J. Low-light image enhancement via generative perceptual priors. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 10752–10760.

 doi:10.20944/preprints202603.2366.v1

125. Chen, H.; Li, W.; Gu, J.; Ren, J.; Chen, S.; Ye, T.; Pei, R.; Zhou, K.; Song, F.; Zhu, L. Restoreagent: Autonomous image restoration agent via multimodal large language models. *Advances in Neural Information Processing Systems* **2024**, *37*, 110643–110666.
126. Zhou, Y.; Cao, J.; Zhang, Z.; Wen, F.; Jiang, Y.; Jia, J.; Liu, X.; Min, X.; Zhai, G. Q-Agent: Quality-Driven Chain-of-Thought Image Restoration Agent through Robust Multimodal Large Language Model. *arXiv preprint arXiv:2504.07148* **2025**.
127. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **2022**, *35*, 24824–24837.
128. Jiang, X.; Li, G.; Chen, B.; Zhang, J. Multi-Agent Image Restoration. *arXiv preprint arXiv:2503.09403* **2025**.
129. Li, B.; Li, X.; Lu, Y.; Chen, Z. Hybrid agents for image restoration. *arXiv preprint arXiv:2503.10120* **2025**.
130. Wei, Y.; Zhang, Z.; Ren, J.; Xu, X.; Hong, R.; Yang, Y.; Yan, S.; Wang, M. Clarity chatgpt: An interactive and adaptive processing system for image restoration and enhancement. *arXiv preprint arXiv:2311.11695* **2023**.
131. Wang, T.; Xia, P.; Li, B.; Jiang, P.T.; Kong, Z.; Zhang, K.; Lu, T.; Luo, W. MOERL: When Mixture-of-Experts Meet Reinforcement Learning for Adverse Weather Image Restoration. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 13673–13683.
132. Lu, J.; Wu, Y.; Zhao, Z.; Wang, H.; Jimenez, F.; Majeedi, A.; Fu, Y. SimpleCall: A Lightweight Image Restoration Agent in Label-Free Environments with MLLM Perceptual Feedback. *arXiv preprint arXiv:2512.18599* **2025**.
133. Hugging Face. Introducing IDEFICS: An Open Reproduction of State-of-the-Art Visual Language Models. https://huggingface.co/blog/idefics, 2023.
134. Yu, Y.; Zeng, Z.; Hua, H.; Fu, J.; Luo, J. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785* **2024**.
135. Franzen, R. Kodak lossless true color image suite, 1999.
136. Zhang, L.; Wu, X.; Buades, A.; Li, X. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging* **2011**, *20*, 023016–023016.
137. Martin, D.; Fowlkes, C.; Tal, D.; Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Proceedings of the Proceedings eighth IEEE international conference on computer vision. ICCV 2001. IEEE, 2001, Vol. 2, pp. 416–423.
138. Huang, J.B.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5197–5206.
139. Agustsson, E.; Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 126–135.
140. Abdelhamed, A.; Lin, S.; Brown, M.S. A high-quality denoising dataset for smartphone cameras. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1692–1700.
141. Ma, K.; Duanmu, Z.; Wu, Q.; Wang, Z.; Yong, H.; Li, H.; Zhang, L. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing* **2016**, *26*, 1004–1016.
142. Arbelaez, P.; Maire, M.; Fowlkes, C.; Malik, J. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence* **2010**, *33*, 898–916.
143. Yang, W.; Tan, R.T.; Feng, J.; Liu, J.; Guo, Z.; Yan, S. Deep joint rain detection and removal from a single image. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1357–1366.
144. Zhang, H.; Sindagi, V.; Patel, V.M. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology* **2019**, *30*, 3943–3956.
145. Fu, X.; Huang, J.; Zeng, D.; Huang, Y.; Ding, X.; Paisley, J. Removing rain from single images via a deep detail network. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3855–3863.
146. Qian, R.; Tan, R.T.; Yang, W.; Su, J.; Liu, J. Attentive generative adversarial network for raindrop removal from a single image. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2482–2491.

 doi:10.20944/preprints202603.2366.v1

147. Li, R.; Cheong, L.F.; Tan, R.T. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1633–1642.
148. Quan, R.; Yu, X.; Liang, Y.; Yang, Y. Removing raindrops and rain streaks in one go. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9147–9156.
149. Huang, H.; Luo, M.; He, R. Memory uncertainty learning for real-world single image deraining. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 3446–3460.
150. Sakaridis, C.; Dai, D.; Van Gool, L. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision* **2018**, *126*, 973–992.
151. Sakaridis, C.; Wang, H.; Li, K.; Zurbrügg, R.; Jadon, A.; Abbeloos, W.; Reino, D.O.; Van Gool, L.; Dai, D. ACDC: The adverse conditions dataset with correspondences for robust semantic driving scene perception. *arXiv preprint arXiv:2104.13395* **2021**.
152. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing* **2018**, *28*, 492–505.
153. Ancuti, C.O.; Ancuti, C.; Sbert, M.; Timofte, R. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In Proceedings of the 2019 IEEE international conference on image processing (ICIP). IEEE, 2019, pp. 1014–1018.
154. Punnappurath, A.; Abuolaim, A.; Afifi, M.; Brown, M.S. Modeling defocus-disparity in dual-pixel sensors. In Proceedings of the 2020 IEEE International Conference on Computational Photography (ICCP). IEEE, 2020, pp. 1–12.
155. Pan, L.; Chowdhury, S.; Hartley, R.; Liu, M.; Zhang, H.; Li, H. Dual pixel exploration: Simultaneous depth estimation and image restoration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4340–4349.
156. Abuolaim, A.; Delbracio, M.; Kelly, D.; Brown, M.S.; Milanfar, P. Learning to reduce defocus blur by realistically modeling dual-pixel data. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2289–2298.
157. Lee, C.; Lee, C.; Kim, C.S. Contrast enhancement based on layered difference representation of 2D histograms. *IEEE transactions on image processing* **2013**, *22*, 5372–5384.
158. Wang, S.; Zheng, J.; Hu, H.M.; Li, B. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing* **2013**, *22*, 3538–3548.
159. Vonikakis, V.; Kouskouridas, R.; Gasteratos, A. On the evaluation of illumination compensation algorithms. *Multimedia Tools and Applications* **2018**, *77*, 9211–9231.
160. Ma, K.; Zeng, K.; Wang, Z. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing* **2015**, *24*, 3345–3356.
161. Cai, J.; Gu, S.; Zhang, L. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing* **2018**, *27*, 2049–2062.
162. Guo, X.; Li, Y.; Ling, H. LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing* **2016**, *26*, 982–993.
163. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE transactions on circuits and systems for video technology* **2020**, *30*, 4861–4875.
164. Bevilacqua, M.; Roumy, A.; Guillemot, C.; Alberi-Morel, M.L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding **2012**.
165. Zeyde, R.; Elad, M.; Protter, M. On single image scale-up using sparse-representations. In Proceedings of the International conference on curves and surfaces. Springer, 2010, pp. 711–730.
166. Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; Aizawa, K. Sketch-based manga retrieval using manga109 dataset. *Multimedia tools and applications* **2017**, *76*, 21811–21838.
167. Cheng, D.; Price, B.; Cohen, S.; Brown, M.S. Beyond white: Ground truth colors for color constancy correction. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 298–306.
168. Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; Zhang, L. Toward real-world single image super-resolution: A new benchmark and a new model. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3086–3095.

 doi:10.20944/preprints202603.2366.v1

169. Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; Lin, L. Component divide-and-conquer for real-world image super-resolution. In Proceedings of the European conference on computer vision. Springer, 2020, pp. 101–117.
170. Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; Zhang, L. Seesr: Towards semantics-aware real-world image super-resolution. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 25456–25467.
171. Zhang, K.; Liang, J.; Van Gool, L.; Timofte, R. Designing a practical degradation model for deep blind image super-resolution. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4791–4800.
172. Wang, Z.J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; Chau, D.H. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 893–911.
173. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 3965–3981.
174. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing* **2020**, *159*, 296–307.
175. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3974–3983.
176. Jia, F.; Tan, L.; Wang, G.; Jia, C.; Chen, Z. A super-resolution network using channel attention retention for pathology images. *PeerJ Computer Science* **2023**, *9*, e1196.
177. FUJIFILM Healthcare Europe.; SonoSkills. US-CASE: Ultrasound Cases Dataset, 2025.
178. Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; Ma, J. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion* **2022**, *83*, 79–92.
179. Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; Fan, X. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 8115–8124.
180. Guo, Y.; Gao, Y.; Lu, Y.; Zhu, H.; Liu, R.W.; He, S. Onerestore: A universal restoration framework for composite degradation. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 255–272.
181. Zhou, Y.; Ren, D.; Emerton, N.; Lim, S.; Large, T. Image restoration for under-display camera. In Proceedings of the Proceedings of the ieee/cvf conference on computer vision and pattern recognition, 2021, pp. 9179–9188.
182. Lin, C.H.; Hsu, C.C.; Young, S.S.; Hsieh, C.Y.; Tai, S.C. QRCODE: Quasi-residual convex deep network for fusing misaligned hyperspectral and multispectral images. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–15.
183. Arad, B.; Timofte, R.; Yahel, R.; Morag, N.; Bernat, A.; Cai, Y.; Lin, J.; Lin, Z.; Wang, H.; Zhang, Y.; et al. Ntire 2022 spectral recovery challenge and data set. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 863–881.
184. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* **2011**, *20*, 2378–2386.
185. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **2017**, *30*.
186. Du, Q.; Younan, N.H.; King, R.; Shah, V.P. On the performance evaluation of pan-sharpening techniques. *IEEE Geoscience and Remote Sensing Letters* **2007**, *4*, 518–522.
187. Blau, Y.; Mechrez, R.; Timofte, R.; Michaeli, T.; Zelnik-Manor, L. The 2018 PIRM challenge on perceptual image super-resolution. In Proceedings of the Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.
188. Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; Bovik, A. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3575–3585.
189. Network, A. Blind image quality assessment using a deep bilinear convolutional neural network. *Deep Bilinear Convolutional Neural* **2022**, *5*.

 doi:10.20944/preprints202603.2366.v1

190. Zhou, H.; Tang, L.; Yang, R.; Qin, G.; Zhang, Y.; Li, Y.; Li, X.; Hu, R.; Zhai, G. UniQA: Unified vision-language pre-training for image quality and aesthetic assessment. *arXiv preprint arXiv:2406.01069* **2024**.
191. Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; Ma, K. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14071–14081.
192. Li, X.; Huang, Z.; Zhang, Y.; Shen, Y.; Li, K.; Zheng, X.; Cao, L.; Ji, R. Few-Shot Image Quality Assessment via Adaptation of Vision-Language Models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 10442–10452.
193. Chen, Z.; Qin, H.; Wang, J.; Yuan, C.; Li, B.; Hu, W.; Wang, L. Promptiqa: Boosting the performance and generalization for no-reference image quality assessment via prompts. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 247–264.
194. Kwon, D.; Kim, D.; Ki, S.; Jo, Y.; Lee, H.E.; Kim, S.J. ATTIQA: Generalizable image quality feature extractor using attribute-aware pretraining. In Proceedings of the Proceedings of the Asian Conference on Computer Vision, 2024, pp. 4526–4543.
195. Dong, G.; Liao, X.; Li, M.; Guo, G.; Ren, C. Exploring semantic feature discrimination for perceptual image super-resolution and opinion-unaware no-reference image quality assessment. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 28176–28187.
196. Liu, K.; Zhang, Z.; Li, W.; Pei, R.; Song, F.; Liu, X.; Kong, L.; Zhang, Y. Dog-IQA: Standard-guided Zero-shot MLLM for Mix-grained Image Quality Assessment. *arXiv preprint arXiv:2410.02505* **2024**.
197. Li, W.; Zhang, X.; Zhao, S.; Zhang, Y.; Li, J.; Zhang, L.; Zhang, J. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679* **2025**.
198. You, Z.; Gu, J.; Li, Z.; Cai, X.; Zhu, K.; Dong, C.; Xue, T. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842* **2024**.
199. Chen, Z.; Hu, B.; Niu, C.; Chen, T.; Li, Y.; Shan, H.; Wang, G. IQAGPT: computed tomography image quality assessment with vision-language and ChatGPT models. *Visual Computing for Industry, Biomedicine, and Art* **2024**, *7*, 20.
200. Xie, W.; Dai, R.; Ding, R.; Liu, K.; Chu, X.; Hou, X.; Wen, J. Q-Hawkeye: Reliable Visual Policy Optimization for Image Quality Assessment. *arXiv preprint arXiv:2601.22920* **2026**.
201. Zhou, J.; Liu, C.; Jiang, Q.; Fu, X.; Hou, J.; Li, X. Semantic Contrast for Domain-Robust Underwater Image Quality Assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2026**.
202. Li, X.; Zhang, Z.; Xu, Z.; Xu, S.; Min, X.; Chen, Y.; Zhai, G. Decoupling Perception and Calibration: Label-Efficient Image Quality Assessment Framework. *arXiv preprint arXiv:2601.20689* **2026**.
203. Cai, Z.; Zhang, J.; Yuan, X.; Jiang, P.T.; Chen, W.; Tang, B.; Yao, L.; Wang, Q.; Chen, J.; Li, B. Q-ponder: A unified training pipeline for reasoning-based visual quality assessment. *arXiv preprint arXiv:2506.05384* **2025**.
204. Rifa, K.R.; Zhang, J.; Imran, A. CAP-IQA: Context-Aware Prompt-Guided CT Image Quality Assessment. *arXiv preprint arXiv:2601.01613* **2026**.
205. Wu, H.; Zhu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Li, C.; Wang, A.; Sun, W.; Yan, Q.; et al. Towards open-ended visual quality comparison. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 360–377.
206. Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181* **2023**.
207. Zhang, Z.; Wu, H.; Zhang, E.; Zhai, G.; Lin, W. Q-Bench $^{+}$: A Benchmark for Multi-Modal Foundation Models on Low-Level Vision From Single Images to Pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 10404–10418.
208. Zhao, S.; Zhang, X.; Li, W.; Li, J.; Zhang, L.; Xue, T.; Zhang, J. Reasoning as Representation: Rethinking Visual Reinforcement Learning in Image Quality Assessment. *arXiv preprint arXiv:2510.11369* **2025**.
209. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93.
210. Tinsley, H.E.; Weiss, D.J. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology* **1975**, *22*, 358.
211. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
212. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text summarization branches out, 2004, pp. 74–81.

 doi:10.20944/preprints202603.2366.v1

213. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
214. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
215. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4015–4026.
216. Sun, S.; Ren, W.; Wang, T.; Cao, X. Rethinking image restoration for object detection. *Advances in Neural Information Processing Systems* **2022**, *35*, 4461–4474.
217. Zhu, H.; Sui, X.; Chen, B.; Liu, X.; Chen, P.; Fang, Y.; Wang, S. 2AFC prompting of large multimodal models for image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology* **2024**.
218. Yi, X.; Xu, H.; Zhang, H.; Tang, L.; Ma, J. Diff-Retinex++: Retinex-driven reinforced diffusion model for low-light image enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
219. Li, B.; Liu, X.; Hu, P.; Wu, Z.; Lv, J.; Peng, X. All-in-one image restoration for unknown corruption. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17452–17462.
220. Zhang, J.; Huang, J.; Yao, M.; Yang, Z.; Yu, H.; Zhou, M.; Zhao, F. Ingredient-oriented multi-degradation learning for image restoration. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 5825–5835.
221. Cui, Y.; Zamir, S.W.; Khan, S.; Knoll, A.; Shah, M.; Khan, F.S. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In Proceedings of the 13th international conference on learning representations, ICLR 2025. International Conference on Learning Representations, ICLR, 2025, pp. 57335–57356.
222. Wu, G.; Jiang, J.; Jiang, K.; Liu, X.; Nie, L. DSwinIR: Rethinking Window-Based Attention for Image Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
223. Cui, Y.; Ren, W.; Shi, B.; Knoll, A. Visual-in-Visual: A Unified and Efficient Baseline for Image Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2026**.
224. Wang, C.; Fan, H.; Yang, H.; Karimi, S.; Yao, L.; Yang, Y. Adapting Text-to-Image Generation with Feature Difference Instruction for Generic Image Restoration. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 23539–23550.
225. Wang, T.; Zhang, K.; Shen, T.; Luo, W.; Stenger, B.; Lu, T. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2023, Vol. 37, pp. 2654–2662.
226. Zhang, T.; Liu, P.; Lu, Y.; Cai, M.; Zhang, Z.; Zhang, Z.; Zhou, Q. Cwnet: Causal wavelet network for low-light image enhancement. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 8789–8799.
227. Deng, X.; Dragotti, P.L. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 3333–3348.
228. Wang, Z.; Wu, Y.; Li, D.; Li, G.; Zhu, P.; Zhang, Z.; Jiang, R. LiDAR-assisted image restoration for extreme low-light conditions. *Knowledge-Based Systems* **2025**, *316*, 113382.
229. Janjua, M.K.; Ghasemabadi, A.; Zhang, K.; Salameh, M.; Gao, C.; Niu, D. Grounding Degradations in Natural Language for All-In-One Video Restoration. *arXiv preprint arXiv:2507.14851* **2025**.

 doi:10.20944/preprints202603.2366.v1

Short Biography of Authors

Mingyu Liu (Student Member, IEEE) is currently a PhD candidate in the Chair of Robotics, Artificial Intelligence and Real-time Systems at the Technical University of Munich (TUM), Germany. He received his dual master's degree in Electrical and Computer Engineering from TUM and Electronics and Communication Engineering from Tongji University, China, respectively. His research interests include computer vision in autonomous driving, deep learning, and artificial intelligence.

Haozhan Shu received his bachelor's degree from Harbin Institute of Technology and is currently pursuing his M.Sc. degree in Design and Engineering at the Technical University of Munich (TUM). His research interests include image restoration and autonomous driving.

Yuning Cui (Student Member, IEEE) received the B.Eng. degree from Central South University, China, in 2016, and the M.Eng. degree from National University of Defense Technology, China, in 2018. He is currently working towards the Ph.D. degree at the Chair of Robotics, Artificial Intelligence and Real-time Systems within the School of Computation, Information and Technology at the Technical University of Munich. His research interest lies in image restoration.

Xingcheng Zhou is currently a Ph.D. student in the Chair of Robotics, Artificial Intelligence and Real-time Systems at the Technical University of Munich (TUM), Germany. He completed his M.Sc. in Electrical and Computer Engineering at the Technical University of Munich in 2021. Before starting at TUM, he worked as an Industrial AI Researcher at Siemens. His current research interests include computer vision, autonomous driving, and vision language models.

Hu Cao is currently a professor at school of automation, Southeast University. He was a research associate at the Chair of Robotics, Artificial Intelligence, and Real-Time Systems of the Technical University of Munich (TUM), where he has also been working as a research assistant since October. 2019. He received his Ph.D. degree in computer engineering from the Technical University of Munich (TUM) in 2023. During his studies, he stayed abroad at the ETH Zürich and the University of Hong Kong (HKU), where he was involved in developing algorithms for dense prediction (classification, detection, and segmentation), autonomous driving, and robotic grasping. His current research interests include robotics, machine learning, computer vision, event-based vision, and embodied AI.

 doi:10.20944/preprints202603.2366.v1

Wenqi Ren (Senior Member, IEEE) received the Ph.D. degree from Tianjin University, China, in 2017. From 2015 to 2016, he worked with Prof. Ming-Hsuan Yang as a joint training student in the Electrical Engineering and Computer Science Department, University of California at Merced. He is currently a Professor with the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University. His research interests include image processing and high-level vision problems.

Boxin Shi (Senior Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, in 2007, the ME degree from Peking University, in 2010, and the PhD degree from the University of Tokyo, in 2013. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor with Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper, Runners-Up at CVPR 2024, ICCP 2015, and selected as Best Paper candidate at ICCV 2015. He is an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence/International Journal of Computer Vision* and an area chair of CVPR/ICCV/ECCV. His research interests include computational photography and computer vision.

Alois C. Knoll (Fellow, IEEE) received the M.Sc. degree in electrical / communications engineering from the University of Stuttgart, Stuttgart, Germany, in 1985, and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin (TU Berlin), Berlin, Germany, in 1988. He was with the Faculty of the Computer Science Department, TU Berlin, until 1993. He joined Bielefeld University, Bielefeld, Germany, as a Full Professor, where he has served as the Director for the Technical Informatics Research Group, until 2001. Since 2001, he has been a Professor with the Department of Informatics, Technical University of Munich (TUM), Munich. He was also on the Board of Directors of the Central Institute of Medical Technology, TUM (IMETUM). From 2004 to 2006, he was an Executive Director of the Institute of Computer Science, TUM. His research interests include cognitive, medical robotics, mult-iagent systems, data fusion, adaptive systems, multimedia information retrieval, modeldriven development of embedded systems with applications to automotive software and electric transportation, and simulation systems for robotics and traffic. He was a member of the EU's Highest Advisory Board on Information Technology, Information Society Technology Advisory Group (ISTAG), and its Future and Emerging Technologies (FET) subgroup from 2007 to 2009.