Review

Not peer-reviewed version

Toward Multimodal Understanding in Low-Level Vision: The Rise of Vision-Language Models

Chunming He [†], Dingming Zhang [†], Fengyang Xiao [*], Xinge Guo, Jingjia Feng, Yiqing Wang, Sina Farsiu [*]

Posted Date: 25 June 2026

doi: 10.20944/preprints202606.1854.v1

Keywords: vision-language models; low-level vision; image restoration; prompt learning; multimodal learning

 doi:10.20944/preprints202606.1854.v1

Review

Toward Multimodal Understanding in Low-Level Vision: The Rise of Vision-Language Models

Chunming He †, Dingming Zhang †, Fengyang Xiao *, Xinge Guo, Jingjia Feng, Yiqing Wang and Sina Farsiu *

Duke University, Durham, NC 27708, USA
* Correspondence: fengyang.xiao@duke.edu (F.X.); sina.farsiu@duke.edu (S.F.)
† These authors contributed equally to this work.

Abstract

Conventional low-level vision models often suffer from limited generalization and a lack of explicit semantic understanding, particularly when confronting complex or hybrid real-world degradations. The recent emergence of Vision-Language Models (VLMs) has fundamentally transformed this landscape by injecting visual-linguistic priors into pixel-level reconstruction. This paper presents the first comprehensive survey of VLM applications in low-level vision, proposing a unified taxonomy that organizes existing literature into two complementary paradigms. The first, *Direct VLM Adaptation*, focuses on modifying VLM architectures, including visual encoders and language branches, and employing prompt learning strategies to bridge the inherent gap between high-level semantics and the fine-grained precision required for pixel-level tasks. The second, *VLM as Auxiliary*, leverages pre-trained VLMs as external modules to assist restoration networks, serving distinct roles as Semantic Providers, Degradation Interpreters, Quality Evaluators, and Intelligent Controllers. Beyond natural images, we systematically review applications in specialized domains such as medical imaging and remote sensing, analyzing domain-specific challenges like data scarcity and distribution shifts. Through curated quantitative summaries and qualitative analyses, we identify critical open challenges, including semantic-pixel misalignment and computational efficiency. Finally, we outline promising future directions, emphasizing the transition toward unified, physics-informed, and generative foundation models for universal image restoration. A curated repository is available at: https://github.com/ChunmingHe/awesome-multimodal-large-language-models-in-low-level-vision.

Keywords: vision-language models; low-level vision; image restoration; prompt learning; multimodal learning

1. Introduction

Low-level vision encompasses fundamental image processing tasks that restore or enhance visual data degraded by real-world factors, including super-resolution [1,2], deblurring [3,4], dehazing [5,6], inpainting [7], fusion [8,9], and low-light enhancement [10,11]. They serve as critical preprocessing for downstream applications such as autonomous driving, medical diagnosis [12], remote sensing [13], and multimedia content creation [14]. Driven by deep learning, the field has progressed from early CNNs [15,16] to Transformers with global receptive fields [17] and, more recently, diffusion models with powerful generative priors [10], improving restoration fidelity and robustness.

Despite this progress, most methods remain task-specific and data-driven, with objectives dominated by pixel-wise reconstruction error and little explicit understanding of scene structure, object relations, or contextual intent. As a result, they generalize poorly to complex or hybrid degradations beyond the training distribution, yielding inconsistent perceptual quality on in-the-wild data [18,19]. This semantic gap, *i.e.*, insufficient global understanding within pixel-oriented paradigms, has become a central bottleneck in pushing low-level vision toward higher-level visual reasoning.

 doi:10.20944/preprints202606.1854.v1

Vision-Language Models (VLMs), such as CLIP [20], BLIP [21], and LLaVA [22], offer a transformative remedy. Trained on large-scale image–text data, they acquire rich visual–semantic knowledge and strong zero-shot generalization [20,23], motivating their use in low-level vision [24–26] (see Figure 1). They address the limitations along three axes: semantic priors that enforce consistency in restored outputs [27,28], broad visual–linguistic knowledge that improves generalization to out-of-distribution degradations [29,30], and reasoning that enables adaptive handling of multiple co-existing degradations [31,32] (for generative MLLMs).

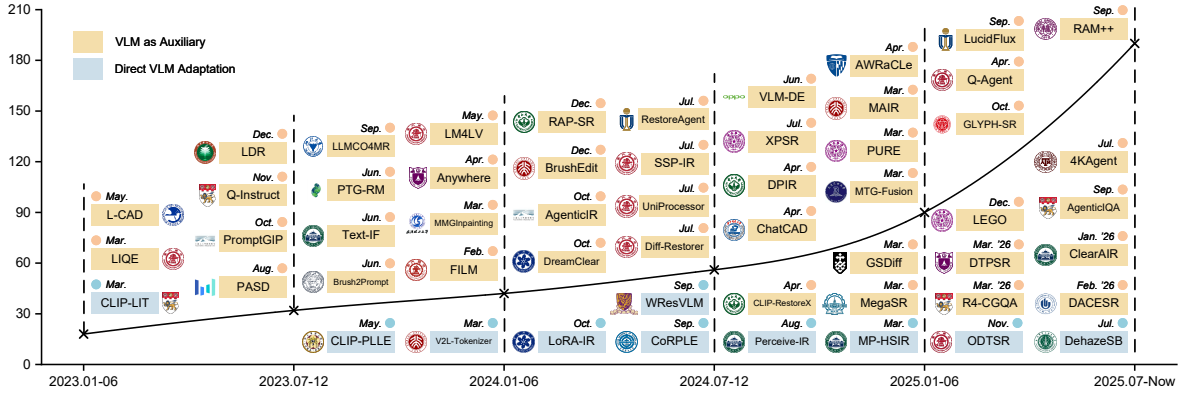


Figure 1. A timeline illustration depicting the evolution of VLM-guided low-level vision models, where Vision–Language Models are primarily used either as auxiliary modules to assist restoration pipelines or are adapted to directly process degraded content.

However, deploying VLMs here is non-trivial due to an objective mismatch: VLMs target high-level semantic understanding and image–text alignment, whereas low-level tasks demand pixel-level accuracy and spatial fidelity [33]. Their visual encoders operate at fixed low resolutions (typically 224×224), and, being trained on clean RGB images, they degrade under severe noise, blur, or haze [34,35]. Domain gaps are even larger for specialized modalities such as hyperspectral remote sensing [36–38] or medical CT/MRI [39–41]. These challenges have spurred research on VLM adaptation, prompt design, and restoration-oriented architectures.

From this review, we identify two complementary paradigms for integrating VLMs into low-level vision. *Direct VLM Adaptation* modifies VLM architectures and training to process and restore degraded images end-to-end [42–47], through visual-encoder adaptation, language-branch modification, and prompt learning. *VLM as Auxiliary* keeps the VLM largely frozen and uses it as an external module in four roles: semantic provider [48–50], degradation interpreter [51–53], quality evaluator [54–56], and intelligent controller [18,19]. The two are not mutually exclusive and can be combined.

Despite this rapid progress, a dedicated systematic survey is still lacking. Existing VLM surveys focus on high-level tasks [57–60] and overlook the distinct challenges of low-level applications; this work fills the gap with the first comprehensive survey of VLMs across low-level vision tasks. Two contemporaneous surveys warrant explicit positioning. Jiang *et al.* [61] review all-in-one restoration by architecture and learning paradigm, without centering the language–vision axis; Liu *et al.* [62] organize language-driven restoration by three coupling levels, restricted to restoration and quality assessment. In contrast, our taxonomy is integration-centered, splitting the literature by how VLM components enter the system (*Direct VLM Adaptation* versus *VLM as Auxiliary* with its provider, interpreter, evaluator, and controller roles) and spans image fusion, medical imaging, remote sensing, video and 3D processing, and zero-shot adaptation.

 doi:10.20944/preprints202606.1854.v1

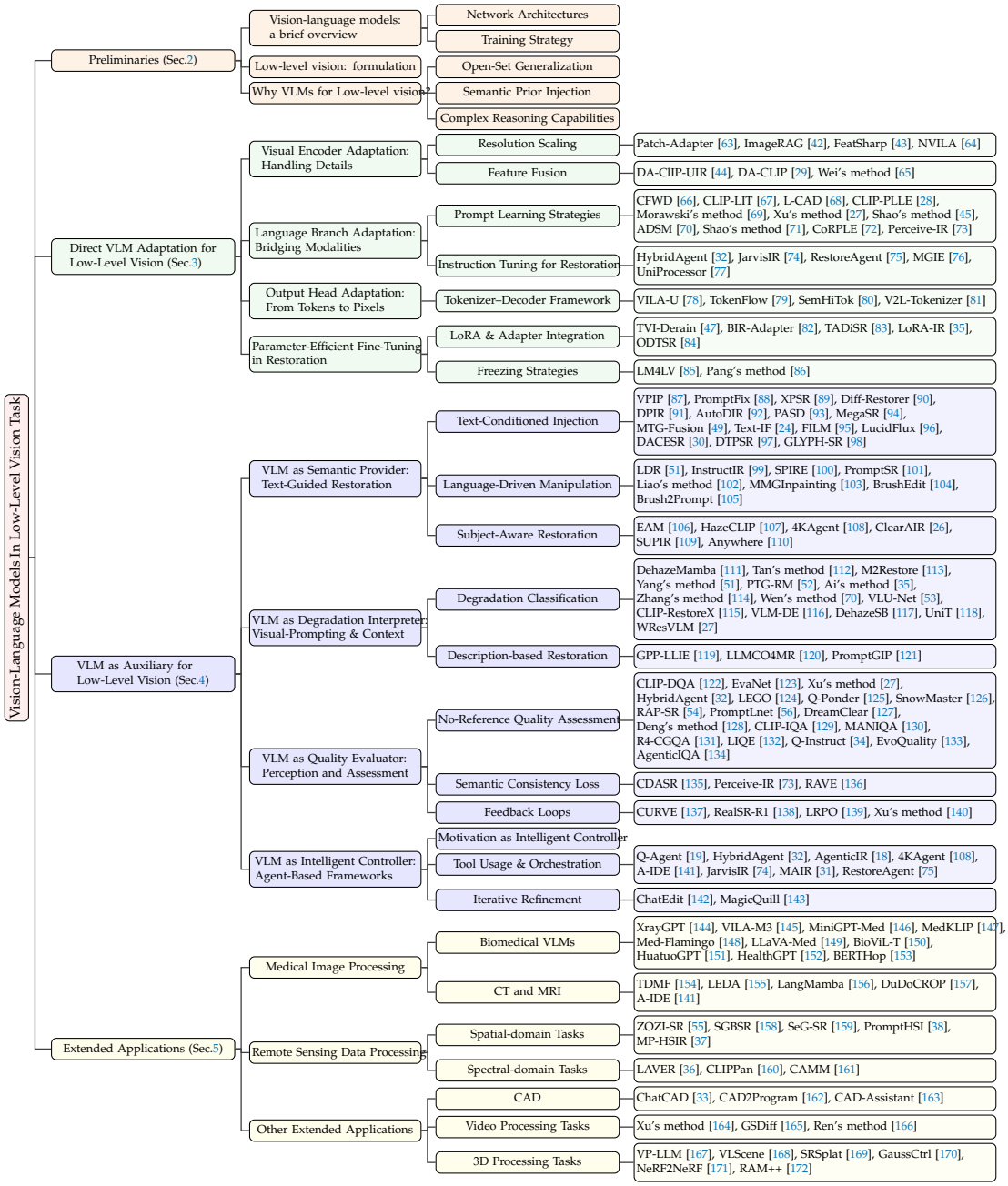


Figure 2. Overview and taxonomy of vision–language model paradigms for low-level vision. The taxonomy organizes methods by their integration roles; contrastive and autoregressive generative mechanisms can occur across its categories.

Studying VLMs in low-level vision is not purely application-oriented; it also informs VLM design. The demand for pixel fidelity and fine-grained spatial consistency exposes architectural and training limitations (resolution, degradation sensitivity, and semantic–pixel misalignment), while solutions such as multi-scale visual encoding, degradation-aware feature extraction, and continuous prompt optimization point toward more versatile VLMs that support semantic understanding and fine-grained reconstruction. This survey thus advances low-level vision with VLMs while informing the broader development of general-purpose vision–language systems.

Our main contributions are summarized as follows. (1) We propose a unified taxonomy that organizes VLM-based low-level vision methods into two paradigms, Direct VLM Adaptation and VLM as Auxiliary, the latter further divided into Provider, Interpreter, Evaluator, and Controller roles, offering a clear framework for categorizing existing and future work. (2) We systematically analyze architectural adaptations and prompt learning strategies that connect VLMs' high-level semantics

 doi:10.20944/preprints202606.1854.v1

with the pixel-level accuracy required in low-level tasks. (3) We extend the discussion beyond natural images to specialized domains such as medical imaging and remote sensing, highlighting domain-specific challenges and adaptations. (4) We compile experimental results on representative benchmarks, summarize open challenges, and outline future directions, with particular emphasis on emerging generative VLMs for image restoration. To facilitate ongoing research, we maintain an up-to-date repository at repository.

The remainder of this survey is organized as: Section 2 introduces the preliminaries. Sections 3 and 4 detail the two core paradigms: Direct VLM Adaptation and VLM as Auxiliary. Section 5 explores extended applications in specialized domains, followed by experiments in Section 6. Finally, Sections 7 and 8 discuss future directions and conclude the work.

Note. Our systematic search covers databases like DBLP and Google Scholar, emphasizing reputable venues such as IEEE T-PAMI, IJCV, and CVPR. We prioritize studies with *public code* and *high citations* as indicators of impact and practicality. Each selected work is carefully examined to assess its unique contribution and influence in the community. This allows our survey to provide a *comprehensive overview of the most representative and influential research*, and to highlight promising directions for future inquiry.

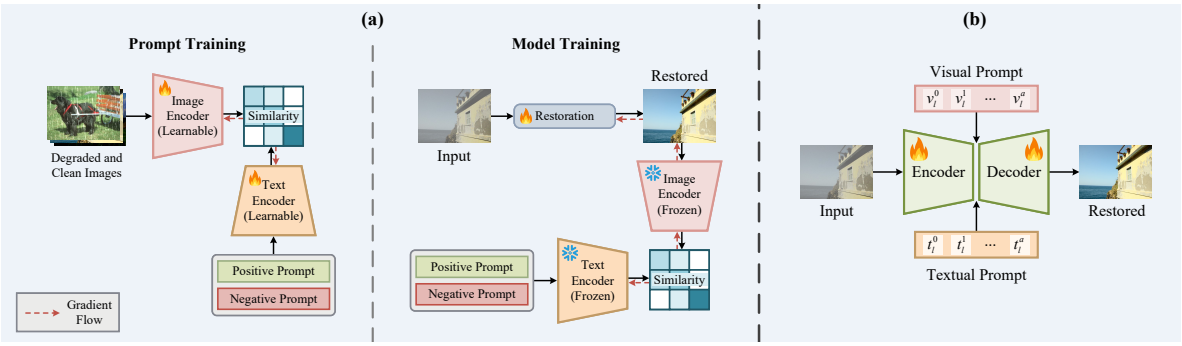


Figure 3. Main application methods of prompt learning. (a) learnable context vectors are adopted as prompts. This approach typically involves two stages: first, the prompt is optimized; second, the learned prompt is used for self-supervised training [27,45,66,69–73]; (b) prompts are simultaneously set in both visual and linguistic modalities [37,173].

2. Preliminaries

2.1. Vision-Language Models: A Brief Overview

VLMs are multimodal systems that model images and text in a shared embedding space, enabling image–text matching and cross-modal retrieval. A typical VLM comprises image and text encoders that map visual and textual inputs into aligned embeddings and, with large-scale contrastive pre-training, exhibit transferability and zero-shot generalization.

We use VLM as an umbrella term and distinguish two operating mechanisms relevant to this survey. Contrastive (dual-encoder) VLMs, represented by CLIP, map images and text into aligned embedding spaces for similarity estimation, retrieval, classification, and feature conditioning. Autoregressive generative VLMs, particularly multimodal large language models (MLLMs) such as LLaVA, connect visual representations to a language model and generate text for description, instruction following, reasoning, and tool use. Models such as BLIP and BLIP-2 support both contrastive representation learning and text generation; we describe them according to the capability used by each reviewed method. This mechanism axis is orthogonal to our integration-centered taxonomy.

2.1.1. Network Architectures

Given a dataset of paired samples $\mathcal{D} = \{x_n^I, x_n^T\}_{n=1}^N$, an image encoder f_θ and a text encoder f_ϕ map each image and its caption to embeddings $z_n^I = f_\theta(x_n^I)$ and $z_n^T = f_\phi(x_n^T)$.

Visual branch. Two families of encoders are widely used: CNN-based and Transformer-based. CNN backbones such as VGG [174], ResNet [175], and EfficientNet [176] are often adopted with minor

 doi:10.20944/preprints202606.1854.v1

changes for vision–language pre-training [20]. Vision Transformers (ViT) [177] instead tokenize an image into patches and process them via stacked Transformer blocks, capturing long-range dependencies in a global token space. Later works [20,178,179] introduce refinements such as pre-encoder normalization to improve stability and representation quality under large-scale VLM pre-training.

Language branch. Transformers and their variants [180–182] dominate text encoding. The standard Transformer consists of multi-head self-attention and multilayer perceptron (MLP) blocks in both encoder and decoder. Most VLMs, *e.g.*, CLIP [20], use Transformers similar to GPT-2 [182], sometimes initialized from GPT-like checkpoints, while others train the text encoder from scratch.

2.1.2. Training Strategy

Contrastive VLMs and generative MLLMs are trained differently. Contrastive VLMs such as CLIP are pre-trained on large image-text corpora with a contrastive objective that pulls matched image-text pairs together and pushes mismatched pairs apart in a shared embedding space, yielding transferable representations without generating text. Generative MLLMs instead follow three stages:

Pre-training. Pre-training aims to align heterogeneous modalities and acquire broad world knowledge using large-scale image–text pairs. A common paradigm involves autoregressively predicting textual descriptions from input images via standard cross-entropy loss. For optimization, many frameworks freeze pre-trained components (*e.g.*, the visual encoder and LLM) and train only a lightweight bridging module to preserve inherent knowledge [149,183]. Conversely, other approaches [184–186] opt to unfreeze specific components, such as the visual encoder, to enable stronger cross-modal adaptation at the cost of increased trainable parameters.

Instruction-tuning. This stage enhances zero-shot generalization by optimizing the model to generate a response $\mathcal{R}$ conditioned on an instruction $\mathcal{I}$ and multimodal input $\mathcal{M}$. Formally, given a triplet $(\mathcal{I}, \mathcal{M}, \mathcal{R})$, the MLLM is trained to minimize the standard autoregressive loss [187,188]:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p(\mathcal{R}_i \mid \mathcal{I}, \mathcal{M}, \mathcal{R}_{<i}; \theta), \tag{1}$$

where θ represents model parameters and $\mathcal{R}_{<i}$ denotes preceding tokens. This process enables the model to generalize to unseen tasks by following natural language instructions.

Preference Alignment. This stage aligns generative MLLMs with human preferences (*e.g.*, reducing hallucination). Reinforcement Learning from Human Feedback (RLHF) [189] utilizes a learned reward model to guide optimization [190], whereas Direct Preference Optimization (DPO) [191] simplifies the pipeline by directly learning from preference data without an explicit reward model. Recently, Group Relative Policy Optimization (GRPO) [192] has emerged to further enhance efficiency; it eliminates the critic (value function) by optimizing based on relative rewards within a sampled group of outputs, thereby significantly reducing memory overhead.

2.2. Low-Level Vision: Formulation

Low-level vision tasks focus on restoring or enhancing degraded visual signals, prioritizing pixel-level reconstruction and structural fidelity over semantic understanding. The primary objective is to recover the clean image x_0 from a degraded observation y. Representative tasks include image denoising [193], deblurring [4], deraining [194], dehazing [6], and super-resolution [1]. Mathematically, these tasks share a unified forward degradation model:

$$y = D(x_0) + n, \tag{2}$$

where x_0 and y denote the clean high-quality (HQ) and observed low-quality (LQ) images, respectively. $D(\cdot)$ represents the degradation operator, encompassing factors such as kernel blur, downsampling, atmospheric scattering, or illumination attenuation, while n denotes random perturbations like sensor noise. Reconstructing x_0 from y constitutes an inherently ill-posed inverse problem characterized

 doi:10.20944/preprints202606.1854.v1

by non-uniqueness and high sensitivity to noise. In real-world scenarios, $D(\cdot)$ is often unknown or difficult to model accurately, further exacerbating the challenge by limiting available constraints.

3. Direct VLM Adaptation for Low-Level Vision

3.1. Visual Encoder Adaptation: Handling Details

Recent work adapts VLM visual branches via patching (to bypass fixed-resolution limits) and controller modules (for degradation awareness).

Resolution Scaling. Input resolutions in low-level vision often exceed standard VLM limits (*e.g.*, 224×224). Resolution-scaling strategies use patching or tiling to process high-resolution images without modifying the backbone, partitioning the image into encoder-sized patches and then recomposing features or retrieving across patches to preserve context and detail. Zhang *et al.* [63] address structural inconsistencies in 4K inpainting by injecting CLIP-retrieved reference features into attention layers for consistency. ImageRAG [42] handles ultra-high-resolution remote sensing by formulating a "long-context selection" problem. FeatSharp [43] and NVILA [64] encode partitioned tiles and stitch them into a unified feature map, capturing both global context and local structure.

Table 1. Instruction tuning for restoration in language branch.

Dataset	Year	Related Method	Short Description	Type
Paired training samples [75]	2024	RestoreAgent [75]	A 23K-pair instruction dataset labeled with the optimal restoration pipeline	Real
CleanBench [74]	2025	JarvisIR [74]	150K synthetic pairs and 80K real degraded images accompanied by self-constructed instruction–response pairs	Real and Syn
SlowAgent [32]	2025	HybridAgent [32]	A total of 70K pairs composed of degraded images, predicted degradation types, and tool-calling instructions.	Syn
FeedbackAgent [32]	2025	HybridAgent [32]	66K pairs used to determine whether an image is clean and whether further restoration is required.	Syn
Low-level vision VQA dataset [77]	2024	UniProcessor [77]	Covering 30 degradation types and over 70K image patches, with diverse QA instructions generated via templates	Syn
InstructIR Prompt Set [99]	2024	InstructIR [99]	An instruction set with 10K+ distinct prompts covering 7 restoration/ enhancement tasks	Syn
PromptFix Dataset [88]	2024	PromptFix [88]	A 1,013,320-sample instruction-following dataset spanning 7 restoration/ editing tasks (e.g., dehazing, low-light, deblurring)	Real and Syn
PixTalk Dataset [195]	2025	PixTalk [195]	A 106K RAW↔sRGB paired dataset with language prompts for photorealistic image processing and editing	Real
Tri-IR [196]	2025	InstructRestore [196]	A 536,945-triplet dataset consisting of a high-quality image, a target-region description, and a region mask for region instruction restoration	Real and Syn
Pico-Banana-400K [197]	2025	Pico-Banana-400K [197]	A 400K instruction-based image editing dataset built on real photos, paired with edited results for instruction-following enhancement	Real and Syn

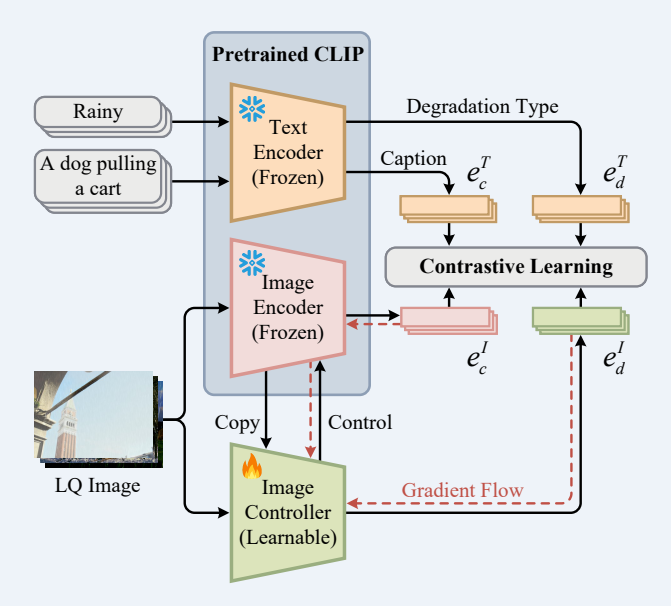


Figure 4. Illustration of combining frozen VLM semantic features with low-level features via trainable controlnet-like structures.

 doi:10.20944/preprints202606.1854.v1

Feature Fusion. VLMs pre-trained on clean images lack robustness to degradation, so their visual features often misalign with textual descriptions in restoration tasks. Recent works inject degradation awareness into the visual branch. Several approaches [29,44] add a parallel image controller (Fig. 4) that encodes degradation embeddings e_d^I from low-quality inputs and fuses them with semantic priors. The primary encoder stays frozen while the controller is fine-tuned, connected by zero-initialized layers for adaptation [198] and optimized with a joint contrastive loss $\mathcal{L}_{\text{con}}$:

$$\mathcal{L}_c(\omega) = \mathcal{L}_{\text{con}}(e_c^I, e_c^T; \omega) + \mathcal{L}_{\text{con}}(e_d^I, e_d^T; \omega), \tag{3}$$

where ω denotes the controller parameters, e_c^I / e_d^I correspond to content/degradation embeddings, and e_c^T / e_d^T are embeddings from the clean caption and degradation label.

3.2. Language Branch Adaptation: Bridging Modalities

Standard VLM language branches, built for high-level semantic matching, lack fine-grained degradation information and sensitivity to quality variations. Two adaptation strategies address this: soft prompt tuning, which injects learnable embeddings to steer the semantic space, and instruction tuning, which fine-tunes the model on restoration-specific tasks.

Prompt Learning Strategies. Soft prompt tuning injects continuous, learnable context vectors into the language branch, optimized by gradient descent to capture degradation-specific priors without modifying the core architecture (Fig. 3). A common paradigm is two-stage: first learn quality-distinguishing prompts, then use them to guide restoration [27,69]. Given a degraded image x encoded as $E_I(x)$ and prompt vectors t for different image characteristics, the matching probability for the i-th prompt is a softmax-normalized cosine similarity:

$$\hat{y}_i = \frac{\exp(\cos(E_I(x), E_T(t_i)))}{\sum_{j=1}^{N} \exp\big(\cos\big(E_I(x), E_T(t_j)\big)\big)}. \tag{4}$$

The prompts are trained with the cross-entropy loss. During restoration, contrastive learning aligns the restored output with the "clear" prompt embedding. Some methods [37,173] prompt both visual and textual modalities to maintain alignment. Yang *et al.* [173] address the feature shift from uni-modal prompting by modeling the layer-wise shift Ω_l.

Instruction Tuning for Restoration. Instruction tuning finetunes VLMs on restoration-centric image–text–image triplets, enabling the language encoder to perceive diverse degradation patterns and understand specific restoration goals; representative datasets are summarized in Table 1. Existing approaches fall into three categories: (1) *Agent Instruction Tuning*, where the VLM acts as a controller to output restoration task signals [32,74] (Sec. 4.4); (2) *Conditional Generation*, where the language model produces soft embeddings to guide a diffusion-based restoration model [76]; and (3) *Degradation Prediction*, where the VLM predicts degradation labels for a downstream reconstruction network [77] (Sec. 4.2).

 doi:10.20944/preprints202606.1854.v1

Table 2. Representative text generation methods.

Generation approach	Method	Explanation	Utilized VLMs
Direct Generation	RAP-SR [54]	Automatically generates image descriptions using Florence-2.	Florence-2 [199]
	RamIR [200]	Generates degradation (p_d^{emb}) and clean (P_c^{emb}) embeddings via BLIP.	BLIP [21]
	CyclicPrompt [102]	Encodes BLIP descriptions into p_t, combining learnable p_i and visual p_k.	BLIP [21]
	LDR [51]	Describes weather types and occluded regions via language instructions.	LISA [201]
	Diff-Dehazer [202]	Derives positive prompts by removing "haze" from BLIP-2 descriptions.	BLIP-2 [203]
	DA-CLIP [29]	Uses triplets of HQ images, descriptive texts, and degradation labels.	BLIP [21]
Progressive generation	PBI [48]	CogVLM describes masked objects; Mistral-7B rewrites them as addition instructions.	CogVLM [204], Mistral-7B [205]
	MTGFusion [49]	Encodes six types of GPT-4 text prompts via BLIP into vectors.	GPT-4 [206], BLIP [21]
	Chen's Method [50]	Fuses coarse-grained (BLIP-2) and fine-grained (MiniGPT-4) descriptions via templates.	BLIP-2 [203], MiniGPT-4 [207]
Multi level generation	XPSR [89]	Generates high-level semantic and low-level detail descriptions via LLaVA.	LLaVA [22]
	PASD [93]	Encodes object/location words (ResNet/YOLO) and scene contents (BLIP).	BLIP [21]
	MTGFusion [49]	Generates separate source, shared, and unique descriptions.	GPT-4 [206], BLIP [21]
	FILM [95]	Combines scene (BLIP-2), object (GRIT), and mask (SAM) prompts for ChatGPT generation.	BLIP-2 [203], ChatGPT [208]
Region level generation	DreamPaint [209]	Generates descriptive text based on Kosmos-2 object detection boxes.	LLaVA-1.6-Vicuna-7B [187]
	Anywhere [110]	Generates text descriptions for foreground regions extracted by RMBG-1.4.	Gemini-Pro-Vision [210]

3.3. Output Head Adaptation: From Tokens to Pixels

Tokenizer–Decoder Framework. Extending VLMs to visual generation requires bridging continuous pixels and discrete token-based LLMs via a codebook tokenizer–decoder. Yu *et al.* [46] address the

 doi:10.20944/preprints202606.1854.v1

limited capacity of traditional VQ-VAE [211] codebooks by reducing embedding dimensions. VILA-U [78] enhances capacity via residual vector quantization, jointly predicting indices from multiple residual stages. To reconcile high-level semantics with fine pixels, TokenFlow [79] shares indexing between semantic and pixel-level codebooks, while SemHiTok [80] uses a hierarchical coarse-semantic codebook with fine-grained pixel sub-codebooks.

Direct Pixel Regression Heads. Beyond discrete tokenization, an alternative line attaches lightweight regression heads that map VLM features directly to continuous pixel values, avoiding the quantization loss of codebook-based decoders. LM4LV [85] keeps the language model frozen and learns only linear adapter layers that regress pixel-level outputs, showing that a frozen LLM can support low-level reconstruction without discrete token prediction. Such heads trade the flexibility of token sequences for higher fidelity on detail-sensitive tasks and remain underexplored relative to tokenizer–decoder designs.

3.4. Parameter-Efficient Fine-Tuning (PEFT) in Restoration

Since direct VLM use is degradation-insensitive and full fine-tuning is costly, PEFT freezes most parameters and tunes lightweight modules to balance generalization and adaptation.

LoRA and Adapter Integration. Adapters [212] and LoRA [213] enable efficient fine-tuning via low-rank matrices or bottleneck layers. Inserting adapters into Transformer layers steers representations toward restoration [47]:

$$\begin{aligned} x_i' &= \text{Adapter}(\text{MSA}(\text{LN}(x_{i-1}))) + x_{i-1}, \\ x_i &= \text{MLP}\big(\text{LN}(x_i')\big) + s \cdot \text{Adapter}\big(\text{LN}(x_i')\big) + x_i', \end{aligned} \tag{5}$$

where s is a scaling factor weighting the adapter output. Adapter-based methods also extend pre-trained diffusion priors: BIR-Adapter [82] augments diffusion self-attention with a parameter-efficient restoring-attention mechanism for blind restoration without auxiliary extractors; TADiSR [83] adapts diffusion cross-attention layers; and LoRA-IR [35] maps CLIP features to degradation types via fine-tuned experts.

Freezing Strategies. Strategic freezing preserves pre-trained knowledge while allowing adaptation. LM4LV [85] freezes the LLM and tunes only linear adapters for pixel-level details, and Pang *et al.* [86] show that frozen LLM Transformer blocks can serve as effective visual encoders when paired with trainable linear projection layers.

4. VLM as Auxiliary for Low-Level Vision

As auxiliary modules, VLMs enhance restoration networks with minimal tuning in four roles mapping onto Equation (2): Semantic Providers shape the prior on x_0; Degradation Interpreters estimate $D(\cdot)$ and n; Quality Evaluators supply a no-reference fidelity surrogate; and Intelligent Controllers select the inversion sequence, dividing labor over prior, forward model, objective, and solver. These roles describe what a model does, not how it is implemented: contrastive VLMs supply embeddings, similarities, or probabilities, whereas generative MLLMs produce descriptions, judgments, reasoning, or tool decisions. Both span the provider, interpreter, and evaluator roles; the controller, generative by nature, is MLLM-based as a rule, with contrastive models entering only as scoring sub-components it reads.

4.1. VLM as Semantic Provider

VLM-based semantic generation falls into four categories, summarized in Table 2. First, *direct generation* constructs explicit image-text pairs to train restoration models, using VLMs such as BLIP [29,44], LLaVA [214], and mPLUG-Owl2 [54] to describe content, scene characteristics, and degradation types; text-only LLMs can further synthesize comprehensive descriptions from image-derived prompts, as in FILM [95]. Second, *progressive generation* [48–50] chains multiple VLMs into a coherent semantic pathway for higher-quality descriptions. Third, *multi-level generation* [89,92,93,215–217] produces prompts

 doi:10.20944/preprints202606.1854.v1

capturing both high-level semantics (scene categories, spatial layout) and fine-grained details. Finally, *region-level generation* [107,110,209] segments images to generate localized descriptions, capturing local features and spatial relationships.

Text-Conditioned Injection. Most works extract semantic or degradation-related embeddings from CLIP, while others use multimodal models for prompts. Diff-Restorer [90] disentangles semantic and degradation features from CLIP, whereas XPSR [89] uses a multimodal model for hierarchical prompts. Injection typically employs cross-attention: VPIP [87] integrates task-specific prompt maps at the U-Net bottom; PromptFix [88] injects user instructions and auxiliary prompts into cross-attention layers (Fig. 5); and XPSR [89] uses a Semantic Attention module. To prevent degradation contamination, Diff-Restorer [90] decomposes CLIP embeddings into content and degradation components, and DPIR [91] combines textual prompts with global-local visual prompts as control for diffusion-transformer restoration.

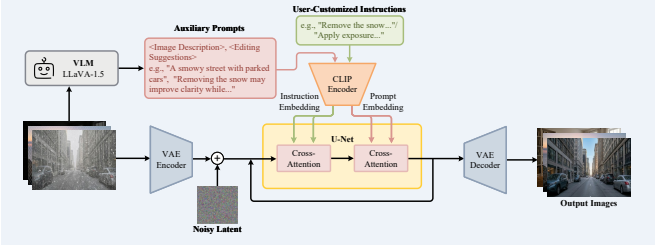


Figure 5. Framework of PromptFix [88], redrawn and reorganized for clarity and visual consistency.

Language-Driven Manipulation. VLMs enable restoration strength and style adjustment via natural-language or numerical prompts. For dynamic routing, LDR [51] uses a degradation map to guide a Mixture-of-Experts for kernel selection, while InstructIR [99] uses an Instruction Condition Block for soft gating. For strength control, SPIRE [100] embeds explicit degradation parameters (*e.g.*, sigma values) into prompts for continuous modulation, and PromptSR [101] allows discrete intensity selection (*e.g.*, heavy/light). Liao *et al.* [102] extract weather-independent priors to update degradation prompts.

Subject-Aware Restoration. VLMs provide semantic priors that resolve the ill-posed nature of restoration, enabling subject-aware control. EAM [106] detects focal regions (*e.g.*, humans) and generates descriptions, preventing background over-enhancement. HazeCLIP [107] segments sky/non-sky regions via SAM [218] for dehazing prompts. 4KAgent [108] invokes a face-repair pipeline for facial regions while preserving backgrounds (Fig. 6). Beyond focal-region enhancement, ClearAIR [26] addresses the all-in-one setting with an MLLM-based quality-assessment module for global degradation evaluation and a semantic guidance unit for region-level analysis.

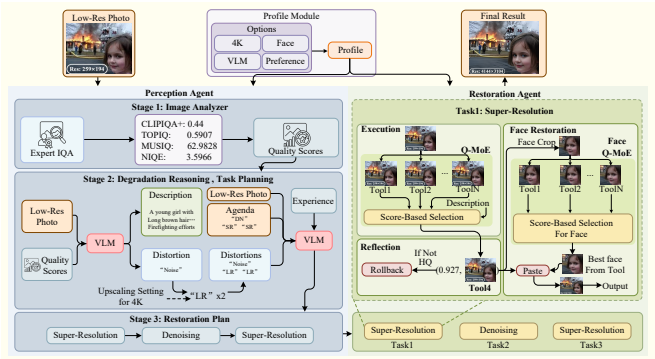


Figure 6. Framework of 4KAgent [108], redrawn and reorganized for clarity and visual consistency.

 doi:10.20944/preprints202606.1854.v1

4.2. VLM as Degradation Interpreter

VLMs function as degradation interpreters through two mechanisms: contrastive VLMs classify or score degradation attributes, while generative MLLMs describe degradation type, severity, and spatial location or produce structured signals.

Degradation Classification. Framing VLMs as degradation classifiers yields high-level descriptors that modulate restoration. A common paradigm extracts semantic features from degraded images [35, 53,113,114] and maps them to a probability distribution over degradation categories. To preserve degradation details in CLIP, Ai *et al.* [35] process both a downsampled image and sliding-window patches and optimized via cross-entropy. Zhang *et al.* [114] use Q-Align [219] to obtain token probability distributions for haze-level estimation. Alternatively, some methods map VLM-extracted tokens into visual embedding spaces [51,70,111,112]: Wen *et al.* [70] map degradation type and attribute prompts directly to high-dimensional vectors; Yang *et al.* [51] use CoT prompts to generate descriptions projected into a degradation prior p_{emb}; Tan *et al.* [112] compute layer-wise HQ–LQ CLIP differences for pixel-level and degradation-type priors; and PTG-RM [52] extracts implicit priors to learn a pixel-wise restoration intensity map.

Description-based Restoration. VLMs also evaluate attributes such as brightness and semantic correlations. In Figure 7, GPP-LLIE [119] uses the probability difference between positive and negative tokens to derive continuous scores for contrast, visibility, and sharpness: $z = -(P_{\text{pos}} - P_{\text{neg}})/\alpha$, $S = (1 + \exp(z))^{-1}$. LLMCO4MR [120] aligns ancient manuscripts using VLM confidence scores; VLU-Net [53] integrates VLM-derived degradation descriptions into a deep unfolding framework where vision-language gradient guidance drives all-in-one restoration; and PromptGIP [121] reformulates processing as QA-style visual prompts for dynamic task switching.

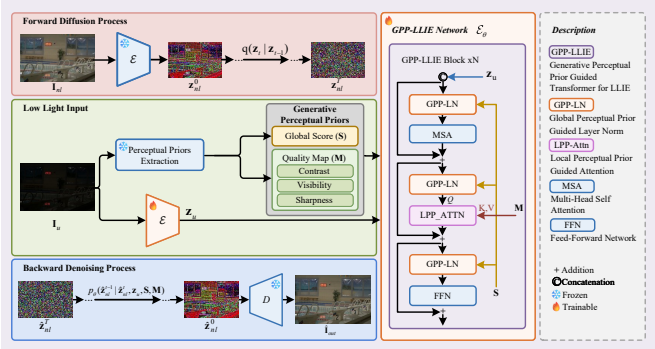


Figure 7. Framework of GPP-LLIE [119], redrawn and reorganized for clarity and visual consistency.

4.3. VLM as Quality Evaluator

VLMs increasingly serve as quality evaluators: contrastive VLMs score quality by comparing images against antonym prompts in a shared space, whereas generative MLLMs are prompted to produce attribute-level quality judgments.

No-Reference Quality Assessment. VLMs evaluate restoration quality when references are unavailable. CLIP-DQA [122] computes cosine similarity between the generated image and antonymous reference prompts (*e.g.*, "good"/"bad") in CLIP space. Beyond natural images, EvaNet [123] extends VLM-guided assessment to image fusion via an LLM-informed perceptual assessment. Others [27,32,124–126] prompt VLMs to estimate quality-level probabilities and convert them to scores; *e.g.*, Xu *et al.* [27] classify visibility into five levels and Ma *et al.* [124] use GPT-4o to score underwater attributes. Aesthetic assessment is also explored [54,56,127]: Wang *et al.* [54] use mPLUG-Owl2 for aesthetic scoring, and Yin *et al.* [56] train a reward model from BLIP and human preference data. Li *et al.* [131] propose R4-CGQA, a CG-specific quality dataset with a two-stream retrieval framework that supplies visually similar CG images.

Semantic Consistency Loss. While VGG-based perceptual loss [220] captures local textures, VLMs encode high-level semantics and aesthetic preferences. CLIP-IQA [129] shows zero-shot assessment,

 doi:10.20944/preprints202606.1854.v1

inspiring perceptual distances: CDASR [135] minimizes super-resolution distance in CLIP space, Perceive-IR [73] (Fig. 8) uses a CLIP-aware loss with fixed quality prompts ("excellent"/"poor"), and RAVE [136] introduces a CLIP-feature identity loss. These semantic losses are typically combined with pixel-level and VGG losses.

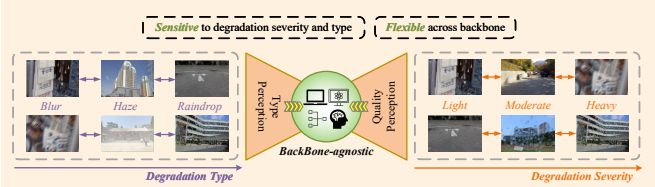


Figure 8. The mechanism of Perceive-IR [73], redrawn and reorganized for clarity and visual consistency.

Feedback Loops. VLM-derived quality scores increasingly serve as rewards in reinforcement learning (RL) loops. CURVE [137] computes rewards via CLIP embedding distances; RealSR-R1 [138] uses GRPO [192] with degradation, comprehension, and generation rewards; and LRPO [139] integrates human-preference, perceptual-quality (CLIP-IQA), and fidelity rewards. Xu *et al.* [140] use a VLM-trained IQA model as the reward with a difficulty-adaptive strategy.

4.4. VLM as Intelligent Controller

Recent studies position MLLMs as central controllers in agent frameworks for multi-degradation restoration. Because planning, tool selection, and reflection are generative by nature, the controller itself is an MLLM, and any contrastive model enters only as a quality-scoring sub-component.

Motivation. Natural images often suffer from concurrent degradations. "All-in-One" models [35, 221] address diverse types but trade generalization for accuracy, whereas combining specialized models requires precise orchestration, since improper sequencing can introduce artifacts [75]. MLLMs act as intelligent controllers to select and order these models.

Tool Usage and Orchestration. Agent-based methods [74,75,108,141] structure restoration into degradation perception, sequence optimization, model selection, and execution (Fig. 9). Given degradation types $\mathcal{D} = \{d_1, \ldots, d_n\}$ and a model library $\mathcal{M}$, the objective is an optimal sequence $\beta = (M_{d_1}^{r_1}, \ldots, M_{d_m}^{r_m})$ maximizing quality $E(I, \beta)$:

$$\beta^* = \arg \max_{\beta \in \mathcal{S}(\mathcal{D}, \mathcal{M})} E(I, \beta). \tag{6}$$

Rollback mechanisms mitigate error propagation. Improvements include standardized metrics [75], hierarchical multi-agent designs [31] distinguishing compression, imaging, and scene degradations, and CoT reasoning with greedy search for efficiency [19]. Beyond orchestration, controller reliability can be improved by training: JarvisIR [74] uses supervised fine-tuning on CleanBench followed by human-feedback alignment, which we treat as a canonical mitigation for the text-hallucination failure mode in Sec. 6.

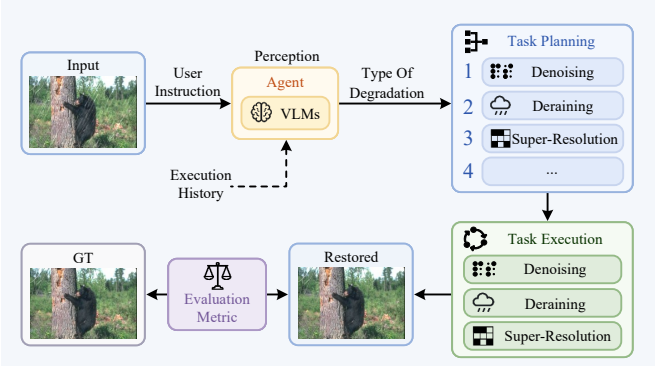


Figure 9. A typical pipeline for agent-based image restoration, where dashed lines indicate optional steps [18,32,75].

 doi:10.20944/preprints202606.1854.v1

Iterative Refinement. VLMs refine decisions through multi-turn human interaction. ChatEdit [142] tracks user requests via a dialogue module while an editing module applies modifications to original images to avoid error accumulation; MagicQuill [143] uses a brush-based paradigm that, after each stroke, infers user intent for real-time multi-turn editing.

5. Extended Applications

5.1. Medical Image Processing

Biomedical VLMs. Medical image acquisition relies on complex, physically constrained processes [224], making quality sensitive to hardware and clinical conditions such as radiation-dose limits, and demanding restoration that preserves clinically relevant semantics. While general-domain VLMs are limited by semantic discrepancies in biomedical data, domain-specific pretraining [40,41,225] and large-scale datasets such as PMC-15M [225] have made biomedical vision–language learning feasible (Table 3).

Table 3. Comparative Overview of Medical Vision-Language Models. "Open" means "Open-source".

Model	Year	Technical points	Function	Paper	Open
LLaVA-Med	2023	Self-supervised instruction tuning; Two-stage curriculum learning.	Medical VQA, Image captioning	[149]	Y
Med-Flamingo	2023	Few-shot learning; OpenFlamingo pre-training.	Rationale generation, Case solving	[148]	Y
BioViL-T	2023	CNN-Transformer hybrid; Temporal image-report pairing.	Disease progression modeling	[150]	Y
XrayGPT	2023	Frozen MedCLIP [222] and LLM; Alignment layer tuning.	Radiograph summarization & QA	[144]	Y
MedKLIP	2023	Entity-level alignment; Structured triplet supervision.	Zero-shot diagnosis, Grounding	[147]	Y
VILA-M3	2024	Expert-guided tuning; Integration of expert models.	Medical VQA, Classification	[145]	Y
MiniGPT-Med	2024	Frozen EVA [223] projection to LLaMA2; Task token guidance.	Radiology diagnosis, Report generation	[146]	Y
HuatuoGPT	2024	VLM-assisted data reformatting; Large-scale alignment.	Cross-modality reasoning	[151]	Y
HealthGPT	2025	Heterogeneous LoRA; Task-aware feature selection.	Modality conversion, Super-resolution	[152]	Y

Li *et al.* [149] proposed LLaVA-Med, pioneering biomedical instruction tuning via curriculum learning. Alkhaldi *et al.* [146] introduced MiniGPT-Med, a unified interface for report generation and disease detection. VILA-M3 [145] enhances reasoning with feedback from expert models such as VISTA3D [226]. Lin *et al.* [152] proposed HealthGPT, the first unified VLM for both understanding and generation across modalities, and Chen *et al.* [151] presented HuatuoGPT-Vision for Chinese medical tasks, supported by PubMedVision.

CT and MRI. MRI encodes hydrogen-proton concentration for soft-tissue visualization, whereas CT encodes tissue density. Chen *et al.* [156] proposed LangMamba for low-dose CT (LDCT) denoising (Fig. 10): a Language-guided AutoEncoder with a frozen medical VLM maps normal-dose CT into an anatomy-enriched semantic space, coupled with a semantic-enhanced denoiser and a dual-space alignment loss. For metal artifact reduction, DuDoCROP [157] uses LLaVA-1.5 [22] for structured descriptions and CLIP for embedding alignment. SCAN-PhysFed [227] uses MiniGPT-Med [146] to extract anatomical information from radiology reports for federated LDCT reconstruction. For fusion, Dong *et al.* [154] use BLIP-2 embeddings to guide a Restormer-based [17] process. Kim *et al.* [228] use language embeddings for controllable brain-MRI synthesis across modalities such as FLAIR and T1.

 doi:10.20944/preprints202606.1854.v1

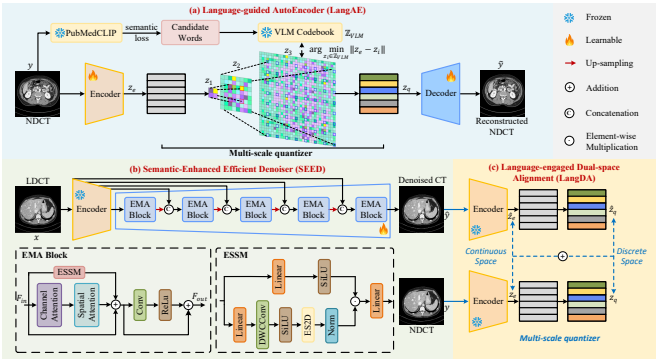


Figure 10. Framework of LangMamba [156], redrawn and reorganized for clarity and visual consistency.

5.2. Remote Sensing Data Processing

Heterogeneous satellite sensors yield increasingly large-scale, diverse data well-suited to the perceptual and reasoning capabilities of VLMs. We organize VLM-based methods into (i) *spatial-domain tasks*, improving spatial fidelity (*e.g.*, super-resolution, restoration), and (ii) *spectral-domain tasks*, recovering and enhancing spectral information.

Spatial-domain tasks. VLMs mainly provide semantic evaluation and guidance for spatial restoration and super-resolution. Wolters *et al.* [55] couple ESRGAN [229] with CLIP-based similarity as a perceptual alternative to PSNR. In Figure 11, Wu *et al.* [158] integrate CLIP features and learn a degradation vector through wavelet patch contrastive learning, mapped by an MLP to modulation parameters. Chen *et al.* [159] apply LoRA to adapt CLIP for module-wise semantic guidance maps. PromptHSI [38] converts text prompts into a task descriptor for frequency-domain modulation, whereas MP-HSIR [37] replaces CLIP visual features with learnable visual tokens fused with text prompts.

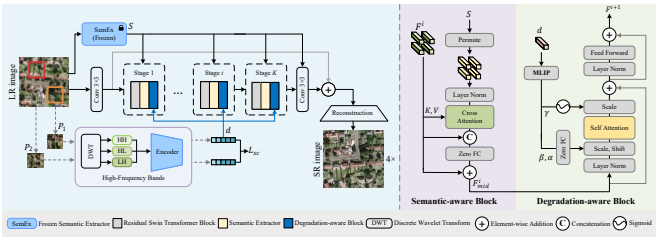


Figure 11. Framework of Wu's method [158], redrawn and reorganized for clarity and visual consistency.

Spectral-domain tasks. Spectral-domain restoration reconstructs rich hyperspectral signatures from limited spectral bands, a severely ill-posed mapping where VLM priors help disambiguate the few observed channels. Chen *et al.* [36] proposed LAVER, which modifies CLIP by adding interpolated positional encodings, enabling resolution-flexible semantic guidance for RGB-to-HSI reconstruction.

5.3. Spatiotemporal and Geometric Modality Processing

This section reviews emerging modalities where VLM-based low-level processing remains under-explored.

CAD. CAD drawings are often corrupted by software-version incompatibilities. Tang *et al.* [33] proposed ChatCAD, the first zero-shot VLM-based CAD restoration system, which simulates the industrial drawing-review workflow via a multi-agent framework that checks engineering-logic consistency and revises drawings when inconsistencies arise.

 doi:10.20944/preprints202606.1854.v1

Table 4. Common dataset for low-level vision tasks. "Size" reports the official train/validation/test split when available; otherwise, it reports the dataset size or the train/test split specified in the corresponding source.

Tasks	Datasets	Size	Source	Type	Description
SR	*DIV2K* [230]	800/100/100	NTIRE 2017	Syn	A high-resolution image benchmark containing diverse natural scenes with synthetically generated degradations.
	RealSR [231]	595	ICCV 2019	Real	A real-capture dataset obtained by varying the camera focal length under controlled settings.
	DRealSR [232]	2,507	ECCV 2020	Real	A real-world dataset acquired with focal-length adjustments across indoor and outdoor environments.
LLIE	*LOLv1* [233]	485/15	BMVC 2018	Real	A real-world paired low-/normal-light dataset collected under diverse illumination conditions.
	LIME [234]	10	TIP 2016	Real	A real-world low-light image set without paired ground truth, used for no-reference low-light enhancement evaluation.
	LOLv2 [235]	1,589/200	TIP 2021	Real&Syn	A combined dataset including real pairs with multi-stage imaging and synthetic pairs using luminance transformations.
Dehaze	*RESIDE* [236]	13,000/990	TIP 2019	Real&Syn	A dehazing benchmark with synthetic and real subsets, covering diverse indoor/outdoor conditions.
	NH-Haze [237]	55	CVPRW 2020	Real	A real-captured outdoor dataset using controlled haze machines, providing paired hazy/clear images.
	Haze-4K [238]	4,000	MM 2021	Syn	A synthetic benchmark including paired images with their transmission maps and atmospheric light annotations.
Inpainting	*CelebA* [239]	200,000	ICCV 2015	Real	A large-scale face benchmark with over 10,000 identities, used in semantic inpainting and human-centric restoration.
	CelebA-HQ [240]	30,000	ArXiv 2017	Real	A high-quality subset of CelebA produced through super-resolution, suitable for high-fidelity face restoration tasks.
	MSCOCO [241]	328,124	ECCV 2014	Real	A general-purpose benchmark offering diverse scenes across 91 object categories with segmentation annotations.
Derain	*Rain100H* [242]	1,800/100	CVPR 2017	Syn	A synthetic benchmark simulating heavy rain streaks with multiple directional patterns for supervised deraining.
	RainDrop [243]	861/239	CVPR 2018	Real	A real-world dataset by capturing paired images through dual glass surfaces, modeling adherent raindrop.
	GT-RAIN [244]	26,124/2,100	ECCV 2022	Real	A real paired deraining dataset with ground-truth clean/rainy pairs captured under controlled non-rain variations.
Deblur	*GoPro* [245]	2,103/1,111	CVPR 2017	Syn	A motion-blur benchmark with high-frame-rate video capture, approximating camera shake via temporal integration.
	RealBlur [246]	3,758/980	ECCV 2020	Real	A motion blur dataset acquired using paired short-/long-exposure imaging in both RAW and JPEG formats.
	HIDE [247]	8,422	ICCV 2019	Syn	A human-centered dynamic-scene dataset with both near-field and far-field motion.
CT denoising	*2016 NIH-AAPM-Mayo* [248]	5,936	Med. Phys. 2017	Real	A dataset with full-dose CT scans from 10 patients, widely used in dose-reduction and denoising studies.
	Mayo-2016 [249]	4,800/1,136	Med. Phys. 2021	Real	A dataset comprising abdominal CT images with normal-dose and simulated quarter-dose ones from 10 subjects.
	Mayo-2020 [249]	2,400/580	Med. Phys. 2021	Real	A dataset with abdominal CT images from 100 patients reconstructed at 25% of the routine radiation dose.

Video processing tasks. VLM-based video restoration is still nascent. Xu *et al.* [164] bring language prompts and VLMs to video face restoration with ControlNet [198] for prompt-free generalization, and build the first video-face degradation dataset on CelebV-HQ [250]. GSDiff [165] uses InstructBLIP [251] to describe the first frame's degradation, supporting generic and medical video reconstruction. Since CLIP-based pulling toward a "haze-free" prompt can induce adversarial behavior (imperceptible changes causing large score swings), Ren *et al.* [166] (Figure 12) propose a CLIP-regularized dehazing framework that measures the mean discrepancy between hazy and clean videos to guide training.

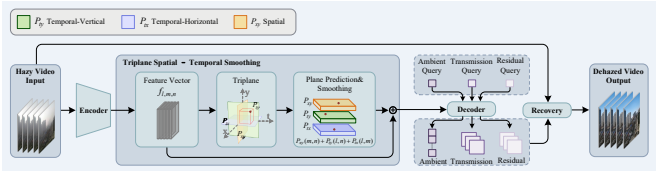


Figure 12. Framework of Ren *et al.*'s method [166], redrawn and reorganized for clarity and visual consistency. Video frames are represented by three complementary planes p_{xy}, p_{ty}, p_{tx}, and three queries extract ambient light, transmission, and residual to recover haze-free images from enhanced features.

Table 5. Results of VLM-guided SR Methods. RealSR uses a same-resolution SR, while DRealSR adopts 4× SR.

Methods	*RealSR* [231]				*DRealSR* [232]				**Params**
	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	[M]
LR Image	**27.80**	**0.7928**	0.3356	34.40	28.02	**0.8374**	0.4410	20.54	N/A
CoSeR [252]	21.24	0.6109	**0.2438**	70.29	19.95	0.5350	**0.2702**	70.18	720+1936
XPSR [89]	24.19	0.6870	0.3517	70.23	26.62	0.7220	0.3864	67.84	868+1066
PASD [93]	25.93	0.7105	0.2806	65.60	**29.09**	0.7937	0.2893	34.63	361+1314
MegaSR [94]	23.49	0.6903	0.3072	70.03	25.74	0.7367	0.3258	64.14	—
PURE [215]	22.83	0.6079	0.3821	**72.37**	24.73	0.6459	0.3452	**73.28**	7000

 doi:10.20944/preprints202606.1854.v1

Table 6. Results of VLM-guided Denoising Methods, where σ denotes the noise level. Values are reported in PSNR / SSIM.

Methods	*CBSD68* [253]			*Urban100* [254]			**Params**
	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	[M]
Noised Image	24.84/0.592	20.54/0.418	15.00/0.219	24.95/0.629	20.69/0.480	15.18/0.292	N/A
InstructIR-3D [99]	34.15/0.933	31.52/0.890	28.30/0.804	34.12/0.945	31.80/0.917	28.63/0.861	16+17
TextPromptIR [255]	34.17/**0.936**	31.52/0.893	28.26/0.805	34.76/0.951	32.54/0.929	29.47/0.882	124+110
CLIPDenoising [256]	33.97/0.930	31.02/0.878	26.69/0.731	33.15/0.930	30.72/0.893	26.27/0.769	11+9
VLU-Net [53]	34.13/0.935	31.48/0.892	28.23/0.804	**34.92**/**0.952**	**32.71**/**0.930**	**29.61**/**0.883**	35+88
DFPIR [257]	**34.32**/0.934	**31.71**/**0.897**	**28.49**/**0.814**	34.79/**0.952**	32.57/**0.930**	29.53/**0.883**	31+63

Table 7. Results of VLM-Guided Multi-Weather Image Restoration Methods.

Methods	*Snow100K-L [258]*		*Outdoor-Rain [259]*		*RainDrop [243]*		**Params**
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	[M]
Degraded Image	18.67	0.6468	12.89	0.5293	23.81	0.8363	N/A
MPerceiver [260]	31.11	0.9180	31.25	0.9246	**33.62**	0.9300	—
ADSM [70]	30.81	0.9058	31.28	0.9227	30.71	0.9188	—
VLCIR [45]	**32.28**	**0.9296**	32.62	0.9447	33.39	**0.9489**	—
CyclicPrompt [102]	32.16	0.9265	**32.81**	0.9371	32.57	0.9454	30+486
M2Restore [113]	31.20	0.9110	32.01	**0.9550**	31.73	0.9430	—

Table 8. Results of VLM-Guided Low-Light Image Enhancement Methods.

Methods	*LOLv1* [233]			*LOLv2-Real* [235]			*LIME* [234]	**Params**
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	[M]
Low-Light Image	7.77	0.191	0.417	9.72	0.196	0.394	4.351	N/A
NeRCo [261]	22.95	0.785	0.311	25.17	0.785	0.338	3.803	46+102
CFWD [66]	**29.19**	0.872	0.197	**29.86**	**0.891**	0.193	**3.568**	22+151
CLIP-LLA [262]	22.66	0.882	0.140	20.23	0.844	0.164	—	2+151
HVI-CIDNet+ [263]	28.85	**0.894**	**0.058**	24.31	0.873	0.107	3.760	69+246
GPP-LLIE [119]	27.51	0.872	0.081	29.23	—	**0.055**	4.24	76+55

Table 9. Results of VLM-Guided Image Deraining Methods.

Methods	*Rain100H* [242]		*Rain100L* [242]		*Test100* [264]		**Params**	**Time**
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	[M]	[s]
Rainy Image	12.13	0.344	25.52	0.814	21.11	0.640	N/A	N/A
DA-CLIP [29]	27.07	0.817	34.50	0.948	27.95	0.857	49+125	10.898
TVI-Derain [47]	30.51	0.893	37.72	0.973	31.17	0.912	20+278	—
TexDepth-Derain [265]	**33.97**	**0.949**	**41.97**	**0.991**	—	—	26+1469	—
PTG-RM-$c_{\text{Restormer}}$ [52]	31.77	0.913	39.27	0.985	**32.30**	**0.934**	27+86	0.142
AWRaCLe [266]	27.20	0.840	35.70	0.966	27.58	0.860	35+151	0.221

 doi:10.20944/preprints202606.1854.v1

Table 10. Results of VLM-guided Image Editing Methods.

Methods	*MagicBrush [267]*			*AnyEdit [268]*			**Params**	**Time**
	L1↓	DINO↑	CVS↑	L1↓	DINO↑	CVS↑	[B]	[s]
ML-MGIE [76]	0.083	84.54	**90.70**	0.211	60.31	78.85	0.86+7.43	5.11
InstructPix2Pix [269]	0.101	71.46	85.22	0.167	65.93	77.90	0.86+0.12	7.76
ICEdit [270]	0.088	84.37	90.17	0.163	66.66	80.90	12.00+4.82	9.95
UltraEdit-SD3 [271]	**0.042**	87.90	89.70	0.152	70.50	80.43	2.00+5.52	3.09
AnySD [268]	0.066	**88.10**	87.60	0.148	**72.45**	81.68	0.86+0.12	3.16
EditMGT [272]	0.109	76.96	86.22	**0.147**	68.80	**82.74**	1.01+3.70	12.75

3D processing tasks. VLMs are emerging as priors for 3D geometry and appearance recovery. DepthLM [273] uses vision-language features for metric depth prediction to reduce geometric ambiguity. For 3D completion, VP-LLM [167] formulates volume completion as token prediction, whereas VLScene [168] distills vision–language guidance into camera-based semantic completion for robustness against occlusion (Fig. 13). SRSplat [169] addresses super-resolution Gaussian splatting from sparse inputs using VLM-assisted priors, and SpatialStack [274] proposes a hierarchical fusion of vision, geometry, and language, in contrast with the late-stage fusion. For text-driven 3D editing, where naive per-view conditioning disrupts geometric consistency, Instruct-NeRF2NeRF [171] applies instruction-following 2D diffusion edits as view-wise supervision while enforcing cross-view coherence, and GaussCtrl [170] uses ControlNet-style conditioned diffusion to stabilize multi-view-consistent editing of Gaussian splats.

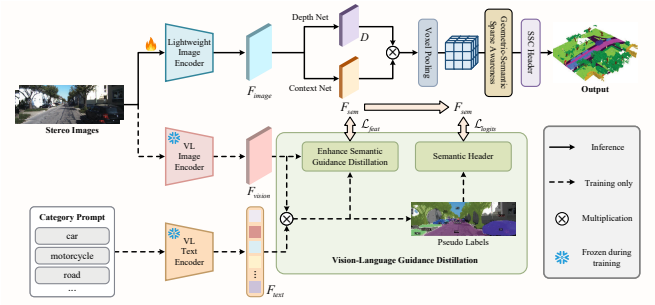
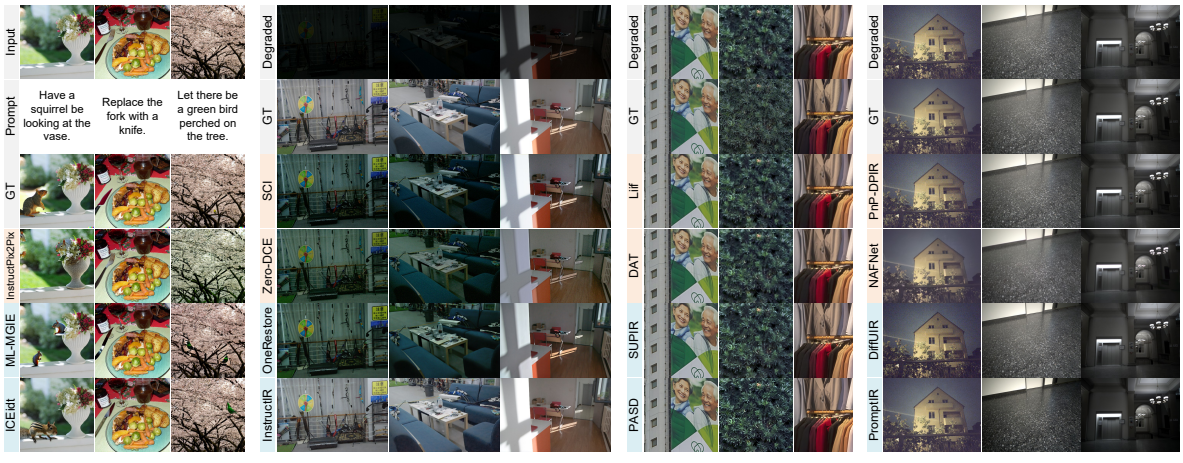


Figure 13. Framework of VLScene [168], redrawn and reorganized for clarity and visual consistency.



(a) Image Editing (b) Low-Light Image Enhancing (c) Super Resolution (d) Image Denoising

Figure 14. Image-level qualitative comparisons across four representative low-level vision tasks. Orange rows denote conventional task-specific models, while blue rows denote VLM-guided methods.

 doi:10.20944/preprints202606.1854.v1

6. Experiments

Protocol and caveats. The reported numbers are *curated summaries* from the original papers or public results, not a unified re-implementation: the methods differ in training data, parameters, degradation settings, and inference protocols. Hence, we use them for *trend-level* comparison rather than absolute ranking, and mark unreported values with "—".

6.1. Datasets

Pre-training datasets. Among open corpora, LAION-5B [275] is the largest (∼5.85B image–text pairs, of which 2.32B are English and 2.26B span over 100 languages); Conceptual Captions (CC3M/CC12M) [276] provide 3M and ∼12M English pairs, SBU Captions [277] 1M Flickr pairs, and MS COCO Captions [278] ∼120K images with 600K human-written descriptions. Proprietary corpora are typically larger: ALIGN [23] (∼1.8B pairs), Florence [279] (FLD-900M), Flamingo [280] (M3W, 185M), and PaLI [281] (WebLI, ∼10B images across 109 languages).

Low-level vision datasets. Table 4 summarizes task-specific datasets for important tasks. VLM-based approaches emphasize adaptation on real-world data, typically fine-tuning pre-trained models to address degradation.

6.2. Evaluation metrics

Distortion-based metrics. PSNR quantifies pixel-wise disparity via mean squared error, and SSIM assesses similarity across contrast, brightness, and structure.

Inception-based metrics. LPIPS [282] uses a pretrained AlexNet to emulate human perception, while FID [283] measures the fidelity and diversity of generated images via the Fréchet distance to reference images.

No-reference quality assessment. MUSIQ [284] handles resolution via multi-scale design, MANIQA [130] uses multi-dimensional attention for local-global cues, and CLIP-IQA [129] leverages CLIP to align with human perception.

Downstream application-based evaluations. Beyond visual quality, evaluation increasingly measures the utility of enhanced images for high-level tasks, *e.g.*, mean Average Precision (mAP) for detection and accuracy for classification.

Instruction-based editing metrics. For image editing, following MGIE [76], L1 measures pixel-level distance to the target, while DINO and CVS measure visual similarity to the target via DINO and CLIP features.

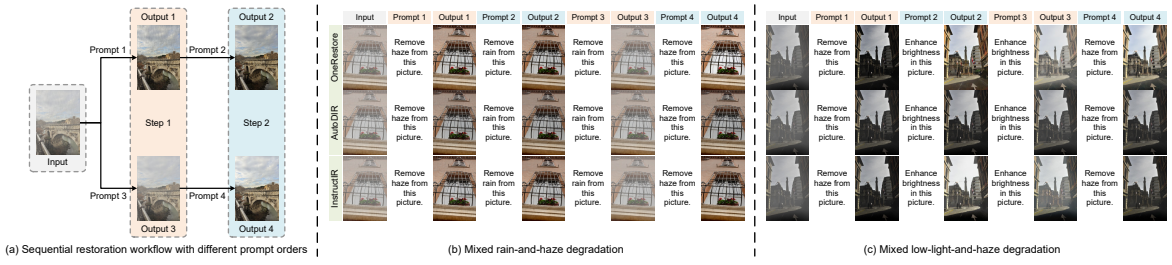


Figure 15. Framework-level analysis of VLM-guided restoration. (a) Text-driven controllability: given the same input, different prompt sequences yield distinct restoration outputs, a capability unique to VLM-guided frameworks. (b) Mixed rain-and-haze and (c) mixed low-light-and-haze degradations: three frameworks (OneRestore, AutoDIR, InstructIR) respond to four different prompts on the same input, demonstrating fine-grained text-conditioned control under compound degradations.

 doi:10.20944/preprints202606.1854.v1

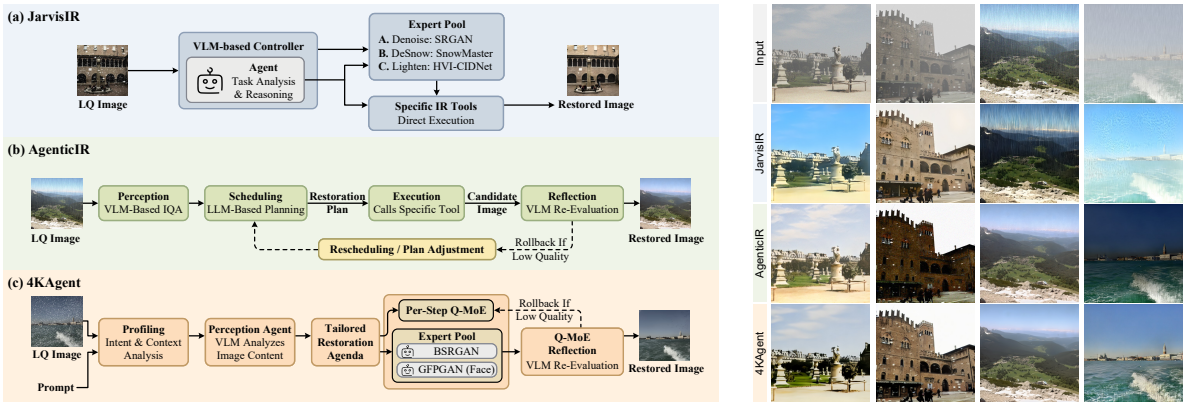


Figure 16. Paradigm-level visualization of agentic restoration workflows. (*Left*) Architectures of three representative VLM-driven agents: JarvisIR, AgenticIR, and 4KAgent. (*Right*) Corresponding qualitative results on the same low-quality inputs.

6.3. Experimental Results

Params are reported as backbone+auxiliary (in millions, M; billions, B in Table 10).

Results on super-resolution. On *RealSR* [231] and *DRealSR* [232] (Table 5), PASD [93] attains the highest PSNR via pixel-level cross-attention, while PURE [215] leads on MUSIQ [284] by combining structural and semantic cues. Qualitatively (Fig. 14(c)), VLM-guided methods (SUPIR, PASD) recover finer textures than conventional ones (Liif).

Results on denoising. On *CBSD68* [253] and *Urban100* [254] (Table 6), DFPIR [257] achieves the best overall results on *CBSD68*, while VLU-Net [53] leads on the structure-rich *Urban100*. TextPromptIR [255] also remains competitive across noise levels. Fig. 14(d) shows VLM-guided methods (DiffUIR, PromptIR) preserve more structural detail under heavy noise than conventional baselines (DPIR).

Results on multi-weather image restoration. On *Snow100K-L* [258], *Outdoor-Rain* [259], and *RainDrop* [243] (Table 7), VLCIR [45] performs well for snow removal, while CyclicPrompt [102] shows strong deraining by leveraging a rain-free prior for second-stage refinement.

Results on low-light enhancement. On *LOLv1* [233], *LIME* [234], and *LOLv2-Real* [235] (Table 8), CFWD [66] achieves the best PSNR and NIQE through multimodal guidance within a diffusion-based framework. In Fig. 14(b), VLM-guided methods (OneRestore, InstructIR) recover brightness and color fidelity in dark regions.

Results on deraining. On *Rain100H* [242], *Rain100L* [242], and *Test100* [264] (Table 9), TexDepthDerain [265] reports the highest PSNR/SSIM on Rain100H and Rain100L, while PTG-RM-$c_{\text{Restormer}}$ [52] achieves the best Test100 performance by leveraging CLIP-derived spatial priors. Inference time (Table 9, last column) varies by over an order of magnitude: prior-injection methods like PTG-RM stay near 0.1–0.2 s, whereas DA-CLIP, querying a CLIP model per image, exceeds 10 s, showing that VLM coupling dominates runtime more than the backbone.

 doi:10.20944/preprints202606.1854.v1

Table 11. Failure modes of VLM-guided low-level vision and representative works that *partially* mitigate them. None of these works fully resolves the corresponding failure; remaining gaps motivate Sec. 7.

Failure mode	Likely cause	Partial mitigation
Text hallucination	Over-association with semantic priors; weak text grounding	JarvisIR [74] human-feedback alignment stage
Noise introduction	Low-quality-caption noise leaking into the prior	VLMIR [285] caption-alignment cosine-similarity loss
Limited generative recovery	Insufficient quality-conditioned generative capacity	Quality-conditioned supervision: QualiTeacher [286], IQPIR [287]
Zero-shot failure	Poor transfer to unseen degradations	EvoQuality [133] (voting + GRPO); DA-CLIP [29]

Results on image editing. On *MagicBrush* [267] and *AnyEdit* [268] (Table 10), UltraEdit-SD3 [271] shows good edit fidelity on *MagicBrush*, AnySD [268] strong semantic alignment on *AnyEdit*, and ML-MGIE [76] stable performance across both. Fig. 14(a) shows VLM-guided editors (InstructPix2Pix, ML-MGIE) follow free-form instructions such as "*Have a squirrel be looking at the vase*". For editing (Table 10, last column), methods run at a few seconds per image, with diffusion-transformer designs (*e.g.*, EditMGT) the slowest, reflecting multi-step sampling rather than the VLM call.

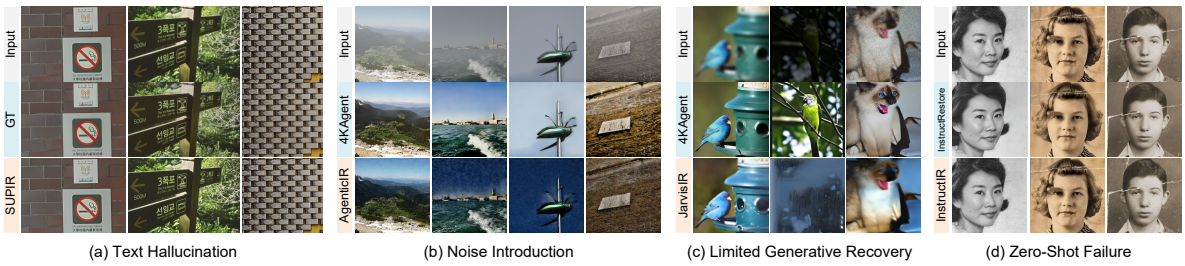


Figure 17. Representative failure cases of VLM-guided restoration: (a) text hallucination, (b) noise introduction, (c) limited generative recovery, and (d) zero-shot failure.

Framework-level analysis: text-driven controllability under mixed degradations. Beyond image-level metrics, VLM-guided frameworks offer *text-driven controllability*. In Fig. 15(a), the same input under different prompt sequences ("Prompt 1→2" vs. "Prompt 3→4") yields distinct restoration trajectories, which is inaccessible to task-specific models. Fig. 15(b,c) compares OneRestore, AutoDIR, and InstructIR under mixed rain-and-haze and low-light-and-haze degradations, showing that these frameworks adapt to each prompt's intent rather than merely handling compound degradations.

Paradigm-level analysis: agentic restoration workflows. At the paradigm level, recent works orchestrate multiple restoration tools via agents (Fig. 16): JarvisIR uses a controller–expert design dispatching inputs to specific IR tools; AgenticIR runs a closed loop of *Perception* → *Scheduling* → *Execution* → *Reflection* with rollback on insufficient quality; and 4KAgent adds a profiling stage and a per-step Q-MoE module. Progressively more sophisticated agentic designs yield better detail recovery and fewer artifacts.

Failure case analysis. Current VLM-guided methods still exhibit characteristic failure modes (Fig. 17, Table 11). *(a) Text hallucination:* the VLM perceives *objects* but not *textual content*, over-associating with semantic priors and distorting characters on signs. *(b) Noise introduction:* the agent commits to a plan from the *initial* degradation assessment and leaves residual noise introduced by intermediate tools uncorrected. *(c) Limited generative recovery:* under severe compound degradations, insufficient generative capacity yields blurry or incomplete reconstructions. *(d) Zero-shot failure:* unseen degradation types are not generalized. These failure modes motivate the directions in Sec. 7.

 doi:10.20944/preprints202606.1854.v1

7. Future Directions

While VLMs have advanced low-level vision, current methods remain dominated by the *VLM as Auxiliary* paradigm, with end-to-end *Direct VLM Adaptation* still nascent. We outline research vectors to bridge the granularity gap between semantic reasoning and pixel-level reconstruction.

7.1. Advancing VLM Architectures for Low-Level Vision

Bridging the semantic-pixel granularity gap. A core bottleneck is the dimensionality mismatch between VLM latent spaces and the pixel-space requirements of restoration [79,80,288]. Beyond simple cross-attention, learnable *hierarchical tokenization* with unified codebooks could jointly encode coarse semantics (*e.g.*, "foggy cityscape") and fine high-frequency details (*e.g.*, edge textures), letting VLMs reason about degradation globally while modulating local pixels via texture-aware decoders, a promising step toward a unified "All-in-One" restoration foundation model.

From guided restoration to direct generation. Most methods use VLMs to guide a separate diffusion or CNN backbone. A transformative direction empowers VLMs as *direct generative engines* for degradation removal, akin to DALL-E 2 [289] or GPT-4o [290], by extending discrete autoregressive prediction to continuous pixels via regression heads or integrated latent-diffusion decoders. Since naively adding high-level generative tasks can compromise detail-sensitive restoration [291], new objectives must synthesize restorations without hallucinating structural artifacts, reducing the error propagation of cascaded systems.

The perception-distortion tradeoff. Blau and Michaeli [292] show that distortion and perceptual realism cannot be jointly optimized beyond a point, which explains why the LR input leads on PSNR/SSIM in Table 5 while generative methods lead on LPIPS/MUSIQ. VLMs sharpen this tension: the semantic priors that enable realistic reconstructions also drive the hallucinations of Section 6. The goal is thus not to win both ends but to make the operating point controllable and to report both metric families explicitly.

7.2. Expanding VLM Capabilities and Applications

Restoration for downstream machine perception. Most research optimizes human perceptual quality (*e.g.*, LPIPS, FID), yet in autonomous driving and surveillance the consumer is a machine. In *machine-centric restoration*, VLMs act as differentiable proxies for downstream tasks [8,293]; aligning restoration with the perception model's semantic feature space [11] prioritizes features critical for recognition (*e.g.*, edge distinctness) over human aesthetics.

Physically-informed multimodal prompting. Beyond descriptive prompts (*e.g.*, "remove rain"), future systems should encode *explicit physical constraints*, prompting VLMs with physical parameters (scattering coefficients, ISO noise levels, sensor response) or non-visual modalities such as LiDAR depth, thermal, or audio [294] within an embodied framework [295,296], improving fidelity in ill-posed problems like underwater imaging or dehazing.

Quality-conditioned manifold learning. A complementary direction treats perceptual quality and degradation level as *controllable axes of the latent manifold*. Quality-conditioned approaches [286,287] steer generation toward a target quality, suggesting a restoration model conditioned on a continuous quality coordinate that traverses from degraded to clean states. Formalizing such a manifold, parameterized by VLM-derived quality estimates, could unify assessment and restoration and address the limited-generative-recovery failure of Sec. 6.

7.3. Optimizing VLM Performance and Deployment

Degradation-aware parameter-efficient tuning. While LoRA and Adapters have been widely adopted, they are often generic and agnostic to the spatial inductive biases required for image restoration [35,297]. Future research should investigate *degradation-aware efficient tuning*, where trainable modules are specifically designed to interact with the high-frequency components of the VLM's visual encoder. Furthermore, exploring neural architecture search [298] to automatically identify optimal

 doi:10.20944/preprints202606.1854.v1

insertion points for adapters within massive VLMs could maximize restoration performance while minimizing computational overhead.

Data-efficient learning from unpaired priors. The dependence on massive paired datasets [20, 23] is a barrier to scalability. A promising avenue is leveraging the pre-trained priors of VLMs to facilitate *unsupervised or weakly-supervised learning* [11]. By exploiting the rich semantic consistency inherent in CLIP or LLaVA, future models could learn to restore images by maximizing the semantic alignment between the output and a distinct "clean" text prompt. Techniques such as intelligent contrastive negative sampling [14] or distillation from larger models [299] can effectively distill restoration capabilities from the VLM's zero-shot knowledge without requiring pixel-perfect ground truth pairs.

7.4. Addressing Inherent Challenges and New Paradigms

Self-evolving and test-time adaptation. Training on paired synthetic data often leads to poor generalization on real-world degradations due to domain gaps [300]. VLMs offer a unique opportunity for *Test-Time Adaptation* as a new paradigm [301]. Future work should explore optimization objectives that leverage the pre-trained semantic consistency of VLMs as a supervision signal during inference. By dynamically minimizing the semantic discrepancy between the degraded input and the restored output in the VLM's feature space, models can adapt to specific noise distributions or degradation combinations on the fly, significantly enhancing robustness in open-world scenarios. A closely related route is **self-evolution**, in which the model generates its own training signal without human labels. EvoQuality [133] demonstrates this for no-reference quality assessment via pairwise voting and relative-reward updates; extending such label-free self-improvement from quality *assessment* to quality *restoration* remains open, since the supervisory signal for pixel reconstruction is harder to bootstrap than a ranking objective.

Trustworthiness and explainable restoration. As VLMs are deployed in safety-critical domains like medical imaging (CT/MRI) and forensic analysis, black-box restoration is insufficient [11]. Future frameworks must address the risk of hallucination by providing *explainable degradation reasoning* [302], where the model articulates its diagnosis of image quality issues and justifies its restoration strategy via CoT prompting. Additionally, developing human-in-the-loop interactive agents [24,303] that allow users to spatially refine restoration results through natural language feedback, correcting artifacts or adjusting enhancement strength in real-time, will be essential for establishing trust in clinical and professional workflows [27] under real-world conditions.

Convergence with Vision-Language-Action models. As Vision-Language-Action (VLA) models extend VLMs with action outputs for embodied agents, low-level vision becomes a robustness prerequisite rather than a downstream consumer: corrupted or degraded visual inputs propagate directly into action errors. Recent work restores corrupted observations before action prediction to harden VLA policies [304], and broader analyses identify input robustness and perception reliability as open challenges guiding VLA development [305]. A promising direction is to co-design restoration and action heads, so the restoration objective is shaped by downstream control performance and the VLA's action confidence signals which regions warrant restoration, tightening the loop between low-level vision and embodied decision-making.

8. Conclusions

This survey organizes VLM-based low-level vision into two paradigms: *Direct VLM Adaptation*, which refactors visual encoders and prompts to handle degradations end-to-end, and *VLM as Auxiliary*, where VLMs act as Semantic Providers, Degradation Interpreters, Quality Evaluators, or Intelligent Controllers. While VLMs bring strong zero-shot generalization and semantic consistency, particularly in domains such as medical imaging and remote sensing, aligning high-level linguistic representations with the fine-grained fidelity required for reconstruction remains the central bottleneck. As the field moves from cascaded guidance toward unified foundation models, balancing computational efficiency, physical grounding, and explainability will be key. We hope this survey and its taxonomy inspire future

 doi:10.20944/preprints202606.1854.v1

work on robust, open-world restoration; to support ongoing research, we maintain a continuously updated repository of VLM-based low-level vision resources.

Funding: This work was supported in part by the Foundation Fighting Blindness (BR-CL-0621-0812-DUKE) and Research to Prevent Blindness (Unrestricted Grant to Duke University).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, W.; Zhou, F.; Zhu, R.; Fukui, K.; Wang, G.; Xue, J.H. Deep learning for image super-resolution. *Neurocomputing* **2020**, *398*, 291–292.
2. Saharia, C.; Ho, J. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, pp. 713–726.
3. Biyouki, S.A.; Hwangbo, H. A comprehensive survey on deep neural image deblurring. *arXiv preprint arXiv:2310.04719* **2023**.
4. He, C.; Zhang, R.; Chen, Z.; Yang, B.; Fang, C.; Lin, Y.; Xiao, F.; Farsiu, S. UnfoldLDM: Deep Unfolding-based Blind Image Restoration with Latent Diffusion Priors. *arXiv preprint arXiv:2511.18152* **2025**.
5. Liu, Y.; Zhao, G.; Gong, B.; Li, Y.; Raj, R.; Goel, N.; Kesav, S.; Gottimukkala, S.; Wang, Z.; Ren, W.; et al. Improved techniques for learning to dehaze and beyond: A collective study. *arXiv preprint arXiv:1807.00202* **2018**.
6. Fang, C.; He, C.; Xiao, F.; Zhang, Y.; Tang, L.; Zhang, Y.; Li, K.; Li, X. Real-world Image Dehazing with Coherence-based Label Generator and Cooperative Unfolding Network. *NeurIPS* **2024**.
7. Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; Van Gool, L. DiffIR: Efficient diffusion model for image restoration. In Proceedings of the ICCV, 2023, pp. 13095–13105.
8. He, C.; Li, K.; Xu, G.; Zhang, Y.; Hu, R.; Guo, Z.; Li, X. Degradation-resistant unfolding network for heterogeneous image fusion. In Proceedings of the ICCV, 2023, pp. 12611–12621.
9. Xu, G.; He, C.; Wang, H.; Zhu, H.; Ding, W. DM-Fusion: Deep Model-Driven Network for Heterogeneous Image Fusion. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**.
10. He, C.; Fang, C.; Zhang, Y.; Ye, T.; Li, K.; Tang, L.; Guo, Z.; Li, X.; Farsiu, S. Reti-Diff: Illumination degradation image restoration with Retinex-based latent diffusion model. *ICLR* **2025**.
11. He, C.; Zhang, R.; Xiao, F.; Fang, C.; Tang, L.; Zhang, Y.; Farsiu, S. UnfoldIR: Rethinking Deep Unfolding Network in Illumination Degradation Image Restoration. *CVPR* **2026**.
12. Kaur, A.; et al. A complete review on image denoising techniques for medical images. *NPL* **2023**, *55*, 7807–7850.
13. Han, L.; Zhao, Y.; Lv, H.; Zhang, Y.; Liu, H.; Bi, G. Remote sensing image denoising based on deep and shallow feature fusion and attention mechanism. *Remote Sens.* **2022**, *14*, 1243.
14. He, C.; Li, K.; Zhang, Y.; Zhang, Y.; Guo, Z.; Li, X. Strategic Preys Make Acute Predators: Enhancing Camouflaged Object Detectors by Generating Camouflaged Objects. *ICLR* **2024**.
15. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307.
16. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155.
17. 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 CVPR, 2022, pp. 5728–5739.
18. 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**.
19. 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**.
20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the ICML, 2021, pp. 8748–8763.
21. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the ICML, 2022, pp. 12888–12900.

 doi:10.20944/preprints202606.1854.v1

22. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. In Proceedings of the CVPR, 2024, pp. 26296–26306.
23. 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 ICML. PMLR, 2021, pp. 4904–4916.
24. Yi, X.; Xu, H.; Zhang, H.; Tang, L.; Ma, J. Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In Proceedings of the CVPR, 2024, pp. 27026–27035.
25. Yu, Y.; Du, D.; Zhang, L.; Luo, T. Unbiased multi-modality guidance for image inpainting. In Proceedings of the ECCV, 2022, pp. 668–684.
26. Zhang, X.; Zhang, H.; Wang, G.; Zhang, Q.; Zhang, L. ClearAIR: A Human-Visual-Perception-Inspired All-in-One Image Restoration. In Proceedings of the AAAI, 2026.
27. Xu, J.; Wu, M.; Hu, X.; Fu, C.W.; Dou, Q.; Heng, P.A. Towards Real-World Adverse Weather Image Restoration: Enhancing Clearness and Semantics with Vision-Language Models. In Proceedings of the ECCV, 2024, pp. 147–164.
28. Morawski, I.; He, K.; Dangi, S.; Hsu, W.H. Leveraging Content and Context Cues for Low-Light Image Enhancement. *IEEE Trans. Multimedia* **2025**.
29. 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**.
30. Lei, X.; Zhang, W.; Luo, B.; Liang, H.; Cao, W.; Lin, Q. DACESR: Degradation-Aware Conditional Embedding for Real-World Image Super-Resolution. *arXiv preprint arXiv:2602.23890* **2026**.
31. Jiang, X.; Li, G.; Chen, B.; Zhang, J. Multi-Agent Image Restoration. *arXiv preprint arXiv:2503.09403* **2025**.
32. Li, B.; Li, X.; Lu, Y.; Chen, Z. Hybrid agents for image restoration. *arXiv preprint arXiv:2503.10120* **2025**.
33. Tang, J.; Xiao, H.; Li, X.; Wang, W.; Gong, Z. ChatCAD: An MLLM-Guided Framework for Zero-shot CAD Drawing Restoration. In Proceedings of the ICASSP, 2025, pp. 1–5.
34. 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 CVPR, 2024, pp. 25490–25500.
35. Ai, Y.; Huang, H.; He, R. LoRA-IR: Taming Low-Rank Experts for Efficient All-in-One Image Restoration. *arXiv preprint arXiv:2410.15385* **2024**.
36. Chen, B.; Chen, K.; Liu, L.; Shi, Z.; Zou, Z. Leveraging language-aligned visual knowledge for remote sensing image spectral super-resolution. In Proceedings of the WHISPERS, 2024, pp. 1–5.
37. 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**.
38. 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:2411.15922* **2024**.
39. Eslami, S.; de Melo, G.; Meinel, C. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906* **2021**.
40. Eslami, S.; Meinel, C.; De Melo, G. PubmedCLIP: How much does CLIP benefit visual question answering in the medical domain? In Proceedings of the EACL, 2023, pp. 1181–1193.
41. Boecking, B.; Usuyama, N.; Bannur, S.; Castro, D.C.; Schwaighofer, A.; Hyland, S.; Wetscherek, M.; Naumann, T.; Nori, A.; Alvarez-Valle, J.; et al. Making the most of text semantics to improve biomedical vision–language processing. In Proceedings of the ECCV, 2022, pp. 1–21.
42. Zhang, Z.; Shen, H.; Zhao, T.; Guan, Z.; Chen, B.; Wang, Y.; Jia, X.; Cai, Y.; Shang, Y.; Yin, J. ImageRAG: Enhancing Ultra High Resolution Remote Sensing Imagery Analysis with ImageRAG. *arXiv preprint arXiv:2411.07688* **2024**.
43. Ranzinger, M.; Heinrich, G.; Molchanov, P.; Kautz, J.; Catanzaro, B.; Tao, A. FeatSharp: Your Vision Model Features, Sharper. *arXiv preprint arXiv:2502.16025* **2025**.
44. Luo, Z.; Gustafsson, F.K.; Zhao, Z.; Sjölund, J.; Schön, T.B. Photo-realistic image restoration in the wild with controlled vision-language models. In Proceedings of the CVPR, 2024, pp. 6641–6651.
45. Shao, M.; Liu, W.; Li, Q.; Meng, L. Controlling vision-language model for enhancing image restoration. *IVC* **2025**, p. 105538.
46. Yu, L.; Lezama, J.; Gundavarapu, N.B.; Versari, L.; Sohn, K.; Minnen, D.; Cheng, Y.; Birodkar, V.; Gupta, A.; Gu, X.; et al. Language Model Beats Diffusion–Tokenizer is Key to Visual Generation. *arXiv preprint arXiv:2310.05737* **2023**.

 doi:10.20944/preprints202606.1854.v1

47. Yang, Q.; Li, P.; Jin, J.; Jin, G.; Song, T.; Fan, S.; Hou, H. Textual-visual interaction for enhanced single image deraining using adapter-tuned VLMs. *Visual Comput.* **2026**, *42*, 193.
48. Wasserman, N.; Rotstein, N.; Ganz, R.; Kimmel, R. Paint by inpaint: Learning to add image objects by removing them first. *arXiv preprint arXiv:2404.18212* **2024**.
49. Wang, Z.; Zhao, L.; Zhang, J.; Song, R.; Song, H.; Meng, J.; Wang, S. Multi-Text Guidance Is Important: Multi-Modality Image Fusion via Large Generative Vision-Language Model. *Int. J. Comput. Vis.* **2025**, pp. 1–23.
50. Chen, X.; Bai, C.; Wu, Z.; Wu, X.; Zou, Q.; Xia, Y.; Wang, S. Coarse-to-fine text injecting for realistic image super-resolution. *Neurocomputing* **2025**, p. 129591.
51. Yang, H.; Pan, L.; Yang, Y.; Liang, W. Language-driven all-in-one adverse weather removal. In Proceedings of the CVPR, 2024, pp. 24902–24912.
52. Xu, X.; Kong, S.; Hu, T.; Liu, Z.; Bao, H. Boosting image restoration via priors from pre-trained models. In Proceedings of the CVPR, 2024, pp. 2900–2909.
53. 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 CVPR, 2025, pp. 7524–7533.
54. Wang, J.; Fan, Q.; Chen, J.; Gu, H.; Huang, F.; Ren, W. RAP-SR: RestorAtion Prior Enhancement in Diffusion Models for Realistic Image Super-Resolution. In Proceedings of the AAAI, 2025, Vol. 39, pp. 7727–7735.
55. Wolters, P.; Bastani, F.; Kembhavi, A. Zooming out on zooming in: Advancing super-resolution for remote sensing. *arXiv preprint arXiv:2311.18082* **2023**.
56. Yin, J.; He, Y.; Zhang, M.; Zeng, P.; Wang, T.; Lu, S. PromptLnet: Region-adaptive aesthetic enhancement via prompt guidance in low-light enhancement net. *arXiv preprint arXiv:2503.08276* **2025**.
57. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X. A survey on multimodal large language models. *NSR* **2024**, *11*, nwae403.
58. Zhang, J.; Huang, J.; Jin, S.; Lu, S. Vision-language models for vision tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**.
59. Ghosh, A.; Acharya, A.; Saha, S.; Jain, V.; Chadha, A. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214* **2024**.
60. Danish, S.; Sadeghi-Niaraki, A.; Khan, S.U.; Dang, L.M.; Tightiz, L.; Moon, H. A comprehensive survey of vision-language models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Information Fusion* **2025**, p. 103623.
61. Jiang, J.; Zuo, Z.; Wu, G.; Jiang, K.; Liu, X. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**.
62. Liu, M.; Shu, H.; Cui, Y.; Zhou, X.; Cao, H.; Ren, W.; Shi, B.; Knoll, A.C. Language-Driven Image Restoration and Semantic-Aware Quality Assessment: A Survey **2026**.
63. Zhang, J.; Cheng, S.; Sun, Q.; Liu, J.; Luyang, W.; Feng, C.; Fang, C.; et al. Ultra High-Resolution Image Inpainting with Patch-Based Content Consistency Adapter. In Proceedings of the ICCV, 2025, pp. 16991–17000.
64. Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; et al. NVILA: Efficient frontier visual language models. In Proceedings of the CVPR, 2025, pp. 4122–4134.
65. Wei, Y.; Zhang, Y.; Li, K.; Wang, F.; Tang, S.; Zhang, Z. Leveraging vision-language prompts for real-world image restoration and enhancement. *Comput. Vis. Image Underst.* **2025**, *250*, 104222.
66. Xue, M.; He, J.; Wang, W.; Zhou, M. Low-light image enhancement via CLIP-Fourier guided wavelet diffusion. *arXiv preprint arXiv:2401.03788* **2024**.
67. Liang, Z.; Li, C.; Zhou, S.; Feng, R.; Loy, C.C. Iterative Prompt Learning for Unsupervised Backlit Image Enhancement. In Proceedings of the ICCV, 2023, pp. 8060–8069.
68. Chang, Z.; Weng, S.; Zhang, P.; Li, Y.; Li, S.; Shi, B. L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors. In Proceedings of the NeurIPS, 2023, Vol. 36, pp. 77174–77186.
69. Morawski, I.; He, K.; Dangi, S.; Hsu, W.H. Unsupervised Image Prior via Prompt Learning and CLIP Semantic Guidance for Low-Light Image Enhancement. In Proceedings of the CVPR, 2024, pp. 5971–5981.
70. Wen, Y.; Gao, T.; Li, Z.; Zhang, J.; Zhang, K.; Chen, T. All-in-one Weather-degraded Image Restoration via Adaptive Degradation-aware Self-prompting Model. *IEEE Trans. Multimedia* **2025**.
71. Shao, M.; Liu, Y.; Cheng, Y.; Wan, Y.; Wang, C. Adaptive Fuzzy Degradation Perception Based on CLIP Prior for All-in-one Image Restoration. *IEEE Trans. Fuzzy Syst.* **2024**.
72. Li, X.; Liu, J.; Chen, Z.; Zou, Y.; Ma, L.; Fan, X.; Liu, R. Contourlet residual for prompt learning enhanced infrared image super-resolution. In Proceedings of the ECCV. Springer, 2024, pp. 270–288.

 doi:10.20944/preprints202606.1854.v1

73. 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 Trans. Image Process.* **2025**.
74. 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 CVPR, 2025, pp. 22369–22380.
75. Chen, H.; Li, W.; Gu, J.; Ren, J.; Chen, S.; Ye, T.; Pei, R.; Zhou, K.; et al. RestoreAgent: Autonomous image restoration agent via multimodal large language models. *arXiv preprint arXiv:2407.18035* **2024**.
76. Fu, T.J.; Hu, W.; Du, X.; Wang, W.Y.; Yang, Y.; Gan, Z. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102* **2023**.
77. Duan, H.; Min, X.; Wu, S.; Shen, W.; Zhai, G. UniProcessor: a text-induced unified low-level image processor. In Proceedings of the ECCV. Springer, 2024, pp. 180–199.
78. Wu, Y.; Zhang, Z.; Chen, J.; Tang, H.; Li, D.; Fang, Y.; Zhu, L.; Xie, E.; et al. VILA-U: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429* **2024**.
79. Qu, L.; Zhang, H.; Liu, Y.; Wang, X.; Jiang, Y.; Gao, Y.; Ye, H.; Du, D.K.; et al. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In Proceedings of the CVPR, 2025, pp. 2545–2555.
80. Chen, Z.; Wang, C.; Chen, X.; Xu, H.; Huang, R.; Zhou, J.; Han, J.; Xu, H.; Liang, X. SemHiTok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. *arXiv preprint arXiv:2503.06764* **2025**.
81. Zhu, L.; Wei, F.; Lu, Y. Beyond Text: Frozen Large Language Models in Visual Signal Comprehension. In Proceedings of the CVPR, 2024, pp. 27047–27057.
82. Eteke, C.; Griessel, A.; Kellerer, W.; Steinbach, E. BIR-Adapter: A Low-Complexity Diffusion Model Adapter for Blind Image Restoration. *arXiv preprint arXiv:2509.06904* **2025**.
83. Hu, Q.; Fan, L.; Luo, Y.; Yu, Y.; Guo, X.; Fan, Q. Text-Aware Real-World Image Super-Resolution via Diffusion Model with Joint Segmentation Decoders. *arXiv preprint arXiv:2506.04641* **2025**.
84. Fang, Y.; Chen, Y.; Yin, S.; Hu, Q.; Yao, J.; Zhang, Y.; Zhang, X.; Wang, Y. One-Step Diffusion Transformer for Controllable Real-World Image Super-Resolution. In Proceedings of the CVPR, 2026.
85. Zheng, B.; Gu, J.; Li, S. LM4LV: A frozen large language model for low-level vision tasks. *arXiv preprint arXiv:2405.15734* **2024**.
86. Pang, Z.; Xie, Z.; Man, Y.; Wang, Y.X. Frozen transformers in language models are effective visual encoder layers. *arXiv preprint arXiv:2310.12973* **2023**.
87. Chen, X.; Liu, Y.; Pu, Y.; Zhang, W.; Zhou, J.; Qiao, Y.; Dong, C. Learning a low-level vision generalist via visual task prompt. In Proceedings of the ACM MM, 2024, pp. 2671–2680.
88. Zeng, Z.; Hua, H.; Fu, J.; Luo, J.; et al. PromptFix: You Prompt and We Fix the Photo. *NeurIPS* **2024**, *37*, 40000–40031.
89. Qu, Y.; Yuan, K.; Zhao, K.; Xie, Q.; Hao, J.; Sun, M.; Zhou, C. XPSR: Cross-modal priors for diffusion-based image super-resolution. In Proceedings of the ECCV. Springer, 2024, pp. 285–303.
90. Zhang, Y.; Zhang, H.; Chai, X.; Cheng, Z.; Xie, R.; Song, L.; Zhang, W. Diff-Restorer: Unleashing visual prompts for diffusion-based universal image restoration. *arXiv preprint arXiv:2407.03636* **2024**.
91. Kong, D.; Li, F.; Wang, Z.; Xu, J.; Pei, R.; Li, W.; Ren, W. Dual Prompting Image Restoration with Diffusion Transformers. In Proceedings of the CVPR, 2025.
92. Jiang, Y.; Zhang, Z.; Xue, T.; Gu, J. AutoDIR: Automatic all-in-one image restoration with latent diffusion. In Proceedings of the ECCV, 2024, pp. 340–359.
93. Yang, T.; Wu, R.; Ren, P.; Xie, X.; Zhang, L. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In Proceedings of the ECCV, 2024, pp. 74–91.
94. Li, X.; Wu, J.; Huang, X.; Chen, C.; Guan, W.; Hua, X.S.; Nie, L. MegaSR: Mining Customized Semantics and Expressive Guidance for Image Super-Resolution. *arXiv preprint arXiv:2503.08096* **2025**.
95. Zhao, Z.; Deng, L.; Bai, H.; Cui, Y.; Zhang, Z.; Zhang, Y.; Qin, H.; Chen, D.; Zhang, J.; Wang, P.; et al. Image fusion via vision-language model. *arXiv:2402.02235* **2024**.
96. Fei, S.; Ye, T.; Wang, L.; Zhu, L. LucidFlux: Caption-Free Universal Image Restoration via a Large-Scale Diffusion Transformer. In Proceedings of the ICLR, 2026.
97. Jiang, L.; Liu, X.; Tong, X.; Li, Z.; Liu, J.; Tang, J.; Wu, G. Disentangled Textual Priors for Diffusion-based Image Super-Resolution. In Proceedings of the CVPR, 2026.
98. Sung, M.; Ham, S.; Kim, K.; Yoon, Y.; Yun, S.; Kim, I.M.; Kang, J.M. GLYPH-SR: Can We Achieve Both High-Quality Image Super-Resolution and High-Fidelity Text Recovery via VLM-guided Latent Diffusion Model? *arXiv preprint arXiv:2510.26339* **2025**.

 doi:10.20944/preprints202606.1854.v1

99. Conde, M.V.; Geigle, G.; Timofte, R. InstructIR: High-quality image restoration following human instructions. In Proceedings of the ECCV, 2024, pp. 1–21.
100. Qi, C.; Tu, Z.; Ye, K.; Delbracio, M.; Milanfar, P.; Chen, Q.; Talebi, H. SPIRE: Semantic prompt-driven image restoration. In Proceedings of the ECCV, 2024, pp. 446–464.
101. Chen, Z.; Zhang, Y.; Gu, J.; Yuan, X.; Kong, L.; Chen, G.; Yang, X. Image super-resolution with text prompt diffusion. *arXiv preprint arXiv:2311.14282* **2023**.
102. Liao, R.; Li, F.; Wei, Y.; Shi, Z.; Zhang, L.; Bai, H.; Wang, M. Prompt to Restore, Restore to Prompt: Cyclic Prompting for Universal Adverse Weather Removal. *arXiv preprint arXiv:2503.09013* **2025**.
103. Zhang, C.; Yang, W.; Li, X.; Han, H. MMGInpainting: Multi-modality guided image inpainting based on diffusion models. *IEEE Trans. Multimedia* **2024**.
104. Li, Y.; Bian, Y.; Ju, X.; Zhang, Z.; Shan, Y.; Zou, Y.; Xu, Q. BrushEdit: All-in-one image inpainting and editing. *arXiv:2412.10316* **2024**.
105. Chiu, M.T.; Zhou, Y.; Zhang, L.; Lin, Z.; Barnes, C.; Amirghodsi, S.; Shechtman, E.; Shi, H. Brush2Prompt: Contextual Prompt Generator for Object Inpainting. In Proceedings of the CVPR, 2024, pp. 12636–12645.
106. Xie, H.; Du, K.; Yan, Q.; Lu, S.; Han, J.; Chen, H.; Hu, H.; Hu, J. EAM: Enhancing Anything with Diffusion Transformers for Blind Super-Resolution. *arXiv preprint arXiv:2505.05209* **2025**.
107. Wang, R.; Li, W.; Liu, X.; Li, C.; Zhang, Z.; Min, X.; Zhai, G. HazeCLIP: Towards language guided real-world image dehazing. In Proceedings of the ICASSP. IEEE, 2025, pp. 1–5.
108. 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**.
109. Yu, F.; Gu, J.; Li, Z.; Hu, J.; Kong, X.; Wang, X.; He, J.; et al. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In Proceedings of the CVPR, 2024, pp. 25669–25680.
110. Xie, T.; Ma, R.; Wang, Q.; Ye, X.; Liu, F.; Tai, Y.; Zhang, Z.; Wang, L.; Yi, Z. Anywhere: A Multi-Agent Framework for User-Guided, Reliable, and Diverse Foreground-Conditioned Image Generation. In Proceedings of the AAAI, 2025, Vol. 39, pp. 7410–7418.
111. Zhang, R.; Yang, Z.; Pan, L. DehazeMamba: Large multi-modal model guided single image dehazing via mamba. *Vis. Intell.* **2025**, *3*, 11.
112. Tan, Z.; Wu, Y.; Liu, Q.; Chu, Q.; Lu, L.; Ye, J.; Yu, N. Exploring the application of large-scale pre-trained models on adverse weather removal. *IEEE Trans. Image Process.* **2024**.
113. Wang, Y.; Li, Y.; Zheng, Z.; Zhang, X.P.; Wei, M. M2Restore: Mixture-of-Experts-based Mamba-CNN Fusion Framework for All-in-One Image Restoration. *arXiv preprint arXiv:2506.07814* **2025**.
114. Zhang, R.; Yang, H.; Yang, Y.; Fu, Y.; Pan, L. LMHaze: Intensity-aware Image Dehazing with a Large-scale Multi-intensity Real Haze Dataset. In Proceedings of the MM Asia, 2024, pp. 1–1.
115. Huang, X.; Zhang, Q.; Hu, J.F.; Zheng, W.S. CLIP-RestoreX: Restore Image Structure and Perception in Exposure Correction. In Proceedings of the AAAI, 2025, Vol. 39, pp. 3760–3768.
116. Cai, J.; Yang, K.; Ding, J.; Fu, L.; Ouyang, L.; Li, J.; Shen, J.; Meng, Z. Degradation-Aware Image Enhancement via Vision-Language Classification. *arXiv preprint arXiv:2506.05450* **2025**.
117. Lan, Y.; Cui, Z.; Luo, X.; Liu, C.; Wang, N.; Zhang, M.; Su, Y.; Liu, D. When Schrödinger Bridge Meets Real-World Image Dehazing with Unpaired Training. In Proceedings of the ICCV, 2025, pp. 8756–8765.
118. Kim, J.H.; Cho, P.H.; Kim, C.; Min, J.; Lee, J.; Park, J.; Choi, Y.; Kim, S. UniT: Unified Diffusion Transformer for High-fidelity Text-Aware Image Restoration. In Proceedings of the ICLR, 2026.
119. Zhou, H.; Dong, W.; Liu, X.; Zhang, Y.; Zhai, G.; Chen, J. Low-light image enhancement via generative perceptual priors. In Proceedings of the AAAI, 2025, Vol. 39, pp. 10752–10760.
120. Zhang, Y.; Li, H.; Zhang, S.; Wang, R.; He, B.; Dou, H.; Yan, J.; Zhang, Y.; Wu, F. LLMCO4MR: LLMs-Aided Neural Combinatorial Optimization for Ancient Manuscript Restoration from Fragments with Case Studies on Dunhuang. In Proceedings of the ECCV. Springer, 2024, pp. 253–269.
121. Liu, Y.; Chen, X.; Ma, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Unifying Image Processing as Visual Prompting Question Answering. In Proceedings of the ICML, 2024, pp. 30873–30891.
122. Zeng, Y.; Fu, J.; Amirpour, H.; Wang, H.; Yue, G.; Liu, H.; Chen, Y.; Zhou, W. CLIP-DQA: Blindly Evaluating Dehazed Images from Global and Local Perspectives Using CLIP. *arXiv preprint arXiv:2502.01707* **2025**.
123. Cheng, C.; Xu, T.; Wu, X.J.; Zhou, T.; Li, H.; Tang, Z.; Kittler, J. EvaNet: Towards More Efficient and Consistent Infrared and Visible Image Fusion Assessment. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**.
124. Ma, Y.; Xia, F.; Lin, L.; Guan, X. LEGO: LLM-enhanced genetic optimization for underwater robot image restoration. *Pattern Recognit.* **2025**, p. 111782.

 doi:10.20944/preprints202606.1854.v1

125. 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**.
126. Lai, J.; Chen, S.; Lin, Y.; Ye, T.; Liu, Y.; et al. SnowMaster: Comprehensive Real-world Image Desnowing via MLLM with Multi-Model Feedback Optimization. In Proceedings of the CVPR, 2025, pp. 4302–4312.
127. Ai, Y.; Zhou, X.; Huang, H.; Han, X.; Chen, Z.; You, Q.; Yang, H. DreamClear: High-Capacity Real-World Image Restoration with Privacy-Safe Dataset Curation. *NeurIPS* **2024**, *37*, 55443–55469.
128. Deng, J.; Wu, X.; Yang, Y.; Zhu, C.; Wang, S.; Wu, Z. Acquire and then Adapt: Squeezing out Text-to-Image Model for Image Restoration. *arXiv preprint arXiv:2504.15159* **2025**.
129. Wang, J.; Chan, K.C.; Loy, C.C. Exploring CLIP for assessing the look and feel of images. In Proceedings of the AAAI, 2023, Vol. 37, pp. 2555–2563.
130. 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 CVPR, 2022, pp. 1191–1200.
131. Li, Z.; Jin, J.; Cai, S.; Lin, W. R4-CGQA: Retrieval-based Vision Language Models for Computer Graphics Image Quality Assessment. *arXiv preprint arXiv:2603.10578* **2026**.
132. 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 CVPR, 2023, pp. 14071–14081.
133. Wen, W.; Zhi, T.; Fan, K.; Li, Y.; Peng, X.; Zhang, Y.; Liao, Y.; Li, J.; Zhang, L. Self-Evolving Vision-Language Models for Image Quality Assessment via Voting and Ranking. In Proceedings of the ICLR, 2026.
134. 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**.
135. Lu, Z.; Xia, Q.; Wang, W.; Wang, F. CLIP-aware domain-adaptive super-resolution. *arXiv preprint arXiv:2505.12391* **2025**.
136. Gaintseva, T.; Benning, M.; Slabaugh, G. RAVE: Residual vector embedding for CLIP-guided backlit image enhancement. In Proceedings of the ECCV, 2024, pp. 412–428.
137. Ogino, Y.; Toizumi, T.; Ito, A. CURVE: Clip-Utilized Reinforcement Learning for Visual Image Enhancement via Simple Image Processing. In Proceedings of the ICIP. IEEE, 2025, pp. 427–432.
138. Qiao, J.; Cai, M.; Li, W.; Liu, Y.; Huang, X.; He, G.; Xie, J.; Hu, J.; Chen, X.; Lin, S. RealSR-R1: Reinforcement Learning for Real-World Image Super-Resolution with Vision-Language Chain-of-Thought. *arXiv:2506.16796* **2025**.
139. Wu, B.; Liu, Y.; Zhang, C.; Zhao, Y.; Wang, W. LRPO: Enhancing Blind Face Restoration through Online Reinforcement Learning. *arXiv preprint arXiv:2509.23339* **2025**.
140. Xu, X.; Chu, R.; Wang, J.; Zhou, K.; Shu, W.; Yang, H.; Lim, S.N.; Chen, H.; Lin, L. Enhancing Diffusion-based Restoration Models via Difficulty-Adaptive Reinforcement Learning with IQA Reward. *arXiv preprint arXiv:2511.01645* **2025**.
141. Cho, U.; Kim, N. A-IDE: Agent-Integrated Denoising Experts. *arXiv preprint arXiv:2503.16780* **2025**.
142. Cui, X.; Li, Z.; Li, P.; Hu, Y.; Shi, H.; Cao, C.; He, Z. ChatEdit: Towards multi-turn interactive facial image editing via dialogue. In Proceedings of the EMNLP, 2023, pp. 14567–14583.
143. Liu, Z.; Yu, Y.; Ouyang, H.; Wang, Q.; Cheng, K.L.; Wang, W.; Liu, Z.; Chen, Q.; Shen, Y. MagicQuill: An intelligent interactive image editing system. In Proceedings of the CVPR, 2025, pp. 13072–13082.
144. Thawkar, O.; Shaker, A.; Mullappilly, S.S.; Cholakkal, H.; Anwer, R.M.; Khan, S.; Laaksonen, J.; Khan, F.S. XrayGPT: Chest radiographs summarization using medical vision-language models. *arXiv:2306.07971* **2023**.
145. Nath, V.; Li, W.; Yang, D.; Myronenko, A.; Zheng, M.; Lu, Y.; Liu, Z.; Yin, H.; Tang, Y.; Guo, P.; et al. VILA-M3: Enhancing vision-language models with medical expert knowledge. *arXiv preprint arXiv:2411.12915* **2024**.
146. Alkhaldi, A.; Alnajim, R.; Alabdullatef, L.; Alyahya, R.; Chen, J.; Zhu, D.; Alsinan, A.; Elhoseiny, M. MiniGPT-Med: Large language model as a general interface for radiology diagnosis. *arXiv preprint arXiv:2407.04106* **2024**.
147. Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; Xie, W. MedKLIP: Medical knowledge enhanced language-image pre-training for X-Ray diagnosis. In Proceedings of the ICCV, 2023, pp. 21372–21383.
148. Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E.P.; Rajpurkar, P. Med-Flamingo: a multimodal medical few-shot learner. In Proceedings of the ML4H, 2023, pp. 353–367.
149. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS* **2023**, *36*, 28541–28564.

 doi:10.20944/preprints202606.1854.v1

150. Bannur, S.; Hyland, S.; Liu, Q.; Perez-Garcia, F.; Ilse, M.; Castro, D.C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. Learning to exploit temporal structure for biomedical vision-language processing. In Proceedings of the CVPR, 2023, pp. 15016–15027.
151. Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G.H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; et al. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale. *arXiv preprint arXiv:2406.19280* **2024**.
152. Lin, T.; Zhang, W.; Li, S.; Yuan, Y.; Yu, B.; Li, H.; He, W.; Jiang, H.; Li, M.; Song, X.; et al. HealthGPT: A Medical Large Vision-Language Model for Unifying Comprehension and Generation via Heterogeneous Knowledge Adaptation. *arXiv preprint arXiv:2502.09838* **2025**.
153. Monajatipoor, M.; Rouhsedaghat, M.; Li, L.H.; Jay Kuo, C.C.; Chien, A.; Chang, K.W. BERTHop: An effective vision-and-language model for chest X-Ray disease diagnosis. In Proceedings of the MICCAI, 2022, pp. 725–734.
154. Dong, A.; Xu, J.; Wang, L.; Lv, G.; Zhao, G.; Cheng, J. TDMF: Text-Guided Denoising and Interactive Medical Image Fusion. In Proceedings of the ICASSP. IEEE, 2025, pp. 1–5.
155. 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 BIBM, 2024, pp. 3088–3091.
156. 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 Trans. Radiat. Plasma Med. Sci.* **2025**.
157. Zhang, X.; Cai, A.; Wang, S.; Wang, L.; Zheng, Z.; Li, L.; Yan, B. Dual-Domain CLIP-Assisted Residual Optimization Perception Model for Metal Artifact Reduction. *arXiv preprint arXiv:2408.14342* **2024**.
158. Wu, B.; Hao, S.; Wang, W. Semantic-aware Guidance for Blind Super-resolution of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2025**.
159. Chen, B.; Chen, K.; Yang, M.; Zou, Z.; Shi, Z. SeG-SR: Integrating Semantic Knowledge into Remote Sensing Image Super-Resolution via Vision-Language Model. *arXiv preprint arXiv:2505.23010* **2025**.
160. Jian, L.; Liu, J.; Wu, S.; Chen, L. CLIPPan: Adapting CLIP as A Supervisor for Unsupervised Pansharpening. *arXiv preprint arXiv:2511.10896* **2025**.
161. Zhang, M.; Li, L.; Gao, F.; Zhang, Q.; Guo, J. Multimodal Prior Learning with Double Constraint Alignment for Snapshot Spectral Compressive Imaging. In Proceedings of the IJCAI, 2025, pp. 2359–2367.
162. Wang, X.; Zheng, J.; Hu, Y.; Zhu, H.; Yu, Q.; Zhou, Z. From 2D CAD Drawings to 3D Parametric Models: A Vision-Language Approach. *arXiv preprint arXiv:2412.11892* **2024**.
163. Mallis, D.; Karadeniz, A.S.; Cavada, S.; Rukhovich, D.; Foteinopoulou, N.; Cherenkova, K.; Kacem, A.; Aouada, D. CAD-Assistant: Tool-Augmented VLLMs as Generic CAD Task Solvers. *arXiv preprint arXiv:2412.13810* **2025**.
164. Xu, Y.; Song, Z.; Lu, J. Universal Video Face Restoration Method Based on Vision-Language Model. In Proceedings of the ACML, 2025.
165. Liu, J.; Zhang, J.; Yang, S.; Xiang, J.; Wang, X.; Zhao, J.; Yang, Z.; Zhao, J. Towards General-Purpose Video Reconstruction through Synergy of Grid-Splicing Diffusion and Large Language Models. *IEEE Trans. Circuits Syst. Video Technol.* **2025**.
166. Ren, J.; Chen, H.; Ye, T.; Wu, H.; Zhu, L. Triplane-smoothed video dehazing with CLIP-enhanced generalization. *Int. J. Comput. Vis.* **2025**, *133*, 475–488.
167. Liu, J.; Liu, Y.; Zhang, Y.; Meng, Z.; Tai, Y.W.; Tang, C.K. VP-LLM: Text-Driven 3D Volume Completion with Large Language Models through Patchification. *arXiv preprint arXiv:2406.05543* **2024**.
168. Wang, M.; Pi, H.; Li, R.; Qin, Y.; Tang, Z.; Li, K. VLScene: Vision-Language Guidance Distillation for Camera-Based 3D Semantic Scene Completion. *arXiv preprint arXiv:2503.06219* **2025**.
169. Hu, X.; Shi, C.; Yang, C.; Chen, M.; Ding, J.; Wei, T.; Wei, C.; Yu, Z.; Tan, M. SRSplat: Feed-Forward Super-Resolution Gaussian Splatting from Sparse Multi-View Images. *arXiv preprint arXiv:2511.12040* **2025**.
170. Wu, J.; Bian, J.W.; Li, X.; Wang, G.; Reid, I.; Torr, P.; Prisacariu, V.A. GaussCtrl: Multi-View Consistent Text-Driven 3D Gaussian Splatting Editing. *arXiv preprint arXiv:2403.08733* **2024**.
171. Haque, A.; Tancik, M.; Efros, A.A.; Holynski, A.; Kanazawa, A. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. *arXiv preprint arXiv:2303.12789* **2023**.
172. Zhang, Z.; Qin, C.; Guo, C.; Zhang, Y.; Xue, C.; Cheng, M.M.; Li, C. RAM++: Robust Representation Learning via Adaptive Mask for All-in-One Image Restoration. *arXiv preprint arXiv:2509.12039* **2025**.
173. Yang, Y.; Zhang, C.; Yang, Z.; Gao, Y.; Qin, Y.; Li, K.; Sun, X.; Yang, J.; Gu, Y. RESTORE: Towards Feature Shift for Vision-Language Prompt Learning. *arXiv preprint arXiv:2403.06136* **2024**.

 doi:10.20944/preprints202606.1854.v1

174. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
175. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the CVPR, 2016, pp. 770–778.
176. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the ICML, 2019, pp. 6105–6114.
177. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
178. Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; Xu, C. FILIP: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783* **2021**.
179. Mu, N.; Kirillov, A.; Wagner, D.; Xie, S. SLIP: Self-supervision meets language-image pre-training. In Proceedings of the ECCV, 2022, pp. 529–544.
180. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *NeurIPS* **2017**, *30*.
181. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT, 2019, pp. 4171–4186.
182. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
183. Pi, R.; Gao, J.; Diao, S.; Pan, R.; Dong, H.; Zhang, J.; Yao, L.; Han, J.; Xu, H.; Kong, L.; et al. DetGPT: Detect what you need via reasoning. *arXiv:2305.14167* **2023**.
184. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* **2023**, *1*, 3.
185. Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* **2023**.
186. Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; et al. VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. *NeurIPS* **2023**, *36*, 61501–61513.
187. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *NeurIPS* **2023**, *36*, 34892–34916.
188. Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; Xie, W. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415* **2023**.
189. Ziegler, D.M.; Stiennon, N.; Wu, J.; Brown, T.B.; Radford, A.; Amodei, D.; Christiano, P.; Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* **2019**.
190. Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.; Wang, Y.X.; Yang, Y.; et al. Aligning large multimodal models with factually augmented RLHF. In Proceedings of the ACL, 2024, pp. 13088–13110.
191. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS* **2023**, *36*, 53728–53741.
192. Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* **2024**.
193. Fan, L.; Zhang, F.; Fan, H.; Zhang, C. Brief review of image denoising techniques. *Vis. Comput. Ind. Biomed. Art.* **2019**, p. 7.
194. Yang, W.; Tan, R.T.; Wang, S.; Fang, Y.; Liu, J. Single image deraining: From model-based to data-driven and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4059–4077.
195. Conde, M.V.; Lu, Z.; Timofte, R. PixTalk: Controlling Photorealistic Image Processing and Editing with Language. In Proceedings of the ICCV, 2025, pp. 19269–19279.
196. Liu, S.; Ma, J.; Sun, L.; Kong, X.; Zhang, L. InstructRestore: Region-Customized Image Restoration with Human Instructions. *arXiv preprint arXiv:2503.24357* **2025**.
197. Qian, Y.; Bocek-Rivele, E.; Song, L.; Tong, J.; Yang, Y.; Lu, J.; Hu, W.; Gan, Z. Pico-Banana-400K: A Large-Scale Dataset for Text-Guided Image Editing. *arXiv preprint arXiv:2510.19808* **2025**.
198. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the ICCV, 2023, pp. 3836–3847.
199. Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; Yuan, L. Florence-2: Advancing a unified representation for a variety of vision tasks. In Proceedings of the CVPR, 2024, pp. 4818–4829.
200. Tang, A.; Wu, Y.; Zhang, Y. RamIR: Reasoning and action prompting with Mamba for all-in-one image restoration. *Appl. Intell.* **2025**, *55*, 258.
201. Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; Jia, J. LISA: Reasoning segmentation via large language model. In Proceedings of the CVPR, 2024, pp. 9579–9589.

 doi:10.20944/preprints202606.1854.v1

202. Lan, Y.; Cui, Z.; Liu, C.; Peng, J.; Wang, N.; Luo, X.; Liu, D. Exploiting Diffusion Prior for Real-World Image Dehazing with Unpaired Training. In Proceedings of the AAAI, 2025, Vol. 39, pp. 4455–4463.
203. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the ICML, 2023, pp. 19730–19742.
204. Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; XiXuan, S.; et al. CogVLM: Visual expert for pretrained language models. *NeurIPS* **2024**, *37*, 121475–121499.
205. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv preprint arXiv:2310.06825* **2023**.
206. 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:2303.08774* **2023**.
207. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* **2023**.
208. 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. *NeurIPS* **2020**, *33*, 1877–1901.
209. Fanelli, N.; Vessio, G.; Castellano, G. I Dream My Painting: Connecting MLLMs and Diffusion Models via Prompt Generation for Text-Guided Multi-Mask Inpainting. In Proceedings of the WACV, 2025, pp. 6073–6082.
210. 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**.
211. Van Den Oord, A.; Vinyals, O.; et al. Neural discrete representation learning. *NeurIPS* **2017**, *30*.
212. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the ICML, 2019, pp. 2790–2799.
213. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. LoRA: Low-rank adaptation of large language models. *ICLR* **2022**.
214. Fan, G.; Zhou, S.; Hua, Z.; Li, J.; Zhou, J. LLaVA-based semantic feature modulation diffusion model for underwater image enhancement. *Inform. Fus.* **2025**, p. 103566.
215. Wei, H.; Liu, S.; Yuan, C.; Zhang, L. Perceive, Understand and Restore: Real-World Image Super-Resolution with Autoregressive Multimodal Generative Models. *arXiv preprint arXiv:2503.11073* **2025**.
216. Wang, C.; An, W.; Jiang, K.; Liu, X.; Jiang, J. LLV-FSR: Exploiting Large Language-Vision Prior for Face Super-resolution. *arXiv preprint arXiv:2411.09293* **2024**.
217. Sun, H.; Li, W.; Liu, J.; Zhou, K.; Chen, Y.; Guo, Y.; Li, Y.; Pei, R.; Peng, L.; Yang, Y. Beyond Pixels: Text Enhances Generalization in Real-World Image Restoration. *arXiv preprint arXiv:2412.00878* **2024**.
218. 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 ICCV, 2023, pp. 4015–4026.
219. 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**.
220. Johnson, J.; Alahi, A.; et al. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the ECCV, 2016, pp. 694–711.
221. He, X.; Li, L.; Wang, Y.; Zheng, H.; Cao, K.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; Zhou, M. Training-Free Large Model Priors for Multiple-in-One Image Restoration. *arXiv preprint arXiv:2407.13181* **2024**.
222. Wang, Z.; Wu, Z.; Agarwal, D.; Sun, J. MedCLIP: Contrastive learning from unpaired medical images and text. In Proceedings of the EMNLP, 2022, Vol. 2022, p. 3876.
223. Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; Cao, Y. Eva-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389* **2023**.
224. Kumari, S.; Singh, P. Data efficient deep learning for medical image analysis: A survey. *arXiv preprint arXiv:2310.06557* **2023**.
225. Zhang, S.; Xu, Y.; Usuyama, N.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915* **2023**.
226. He, Y.; Guo, P.; Tang, Y.; Myronenko, A.; Nath, V.; Xu, Z.; Yang, D.; Zhao, C.; Simon, B.; Belue, M.; et al. Vista3D: Versatile imaging segmentation and annotation model for 3D computed tomography. *arXiv preprint arXiv:2406.05285* **2024**.
227. Yang, Z.; Chen, Y.; Wang, Z.; Shan, H.; Chen, Y.; Zhang, Y. Patient-level anatomy meets scanning-level physics: Personalized federated Low-Dose CT denoising empowered by large language model. *arXiv preprint arXiv:2503.00908* **2025**.

 doi:10.20944/preprints202606.1854.v1

228. Kim, K.; Na, Y.; Ye, S.J.; Lee, J.; Ahn, S.S.; Park, J.E.; Kim, H. Controllable text-to-image synthesis for multi-modality MR images. In Proceedings of the WACV, 2024, pp. 7936–7945.
229. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. ESRGAN: Enhanced super-resolution generative adversarial networks. In Proceedings of the ECCVW, 2018, pp. 0–0.
230. Agustsson, E.; Timofte, R. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In Proceedings of the CVPRW, July 2017.
231. 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 ICCV, 2019, pp. 3086–3095.
232. 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 ECCV, 2020, pp. 101–117.
233. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep Retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560* **2018**.
234. Guo, X.; Li, Y.; Ling, H. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.* **2016**, *26*, 982–993.
235. Yang, W.; Wang, W.; Huang, H.; Wang, S.; Liu, J. Sparse gradient regularized deep Retinex network for robust low-light image enhancement. *IEEE Trans. Image Process.* **2021**, *30*, 2072–2086.
236. Li, B.; Ren, W.; Fu, D. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.* **2018**, pp. 492–505.
237. Ancuti, C.; Ancuti, C. An image dehazing benchmark with non-homogeneous hazy images. In Proceedings of the CVPRW, 2020, pp. 444–445.
238. Liu, Y.; Zhu, L. From synthetic to real: Image dehazing collaborating with real data. In Proceedings of the ACM MM, 2021, pp. 50–58.
239. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the ICCV, 2015, pp. 3730–3738.
240. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* **2017**.
241. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the ECCV, 2014, pp. 740–755.
242. 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 CVPR, 2017, pp. 1357–1366.
243. 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 CVPR, 2018, pp. 2482–2491.
244. Ba, Y.; Zhang, H.; Yang, E.; Suzuki, A.; Pfahnl, A.; Chandrappa, C.C.; de Melo, C.M.; You, S.; Soatto, S.; Wong, A.; et al. Not Just Streaks: Towards Ground Truth for Single Image Deraining. In Proceedings of the ECCV, 2022, pp. 723–740.
245. Nah, S.; Hyun Kim, T.; Mu Lee, K. Deep multi-scale convolutional neural network for dynamic scene deblurring. In Proceedings of the CVPR, 2017, pp. 3883–3891.
246. Rim, J.; Lee, H.; Won, J.; Cho, S. Real-world blur dataset for learning and benchmarking deblurring algorithms. In Proceedings of the ECCV, 2020, pp. 184–201.
247. Shen, Z.; Wang, W.; Lu, X.; Shen, J.; Ling, H.; Xu, T.; Shao, L. Human-aware motion deblurring. In Proceedings of the ICCV, 2019, pp. 5572–5581.
248. 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. *Med. Phys.* **2017**, *44*, e339–e352.
249. Moen, T.R.; Chen, B.; Holmes III, D.R.; Duan, X.; Yu, Z.; Yu, L.; Leng, S.; Fletcher, J.G.; McCollough, C.H. Low-Dose CT image and projection dataset. *Med. Phys.* **2021**, *48*, 902–911.
250. Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; Loy, C.C. CelebV-HQ: A large-scale video facial attributes dataset. In Proceedings of the ECCV, 2022, pp. 650–667.
251. Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.N.; Hoi, S. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *NeurIPS* **2023**, *36*, 49250–49267.
252. Sun, H.; Li, W.; Liu, J.; Chen, H.; Pei, R.; Zou, X.; Yan, Y.; Yang, Y. CoSeR: Bridging image and language for cognitive super-resolution. In Proceedings of the CVPR, 2024, pp. 25868–25878.
253. 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 ICCV. IEEE, 2001, Vol. 2, pp. 416–423.

 doi:10.20944/preprints202606.1854.v1

254. Huang, J.B.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. In Proceedings of the CVPR, 2015, pp. 5197–5206.
255. Yan, Q.; Jiang, A.; Chen, K.; Peng, L.; Yi, Q.; Zhang, C. Textual prompt guided image restoration. *arXiv:2312.06162* **2023**.
256. Cheng, J.; Liang, D.; Tan, S. Transfer CLIP for generalizable image denoising. In Proceedings of the CVPR, 2024, pp. 25974–25984.
257. Tian, X.; Liao, X.; Liu, X.; Li, M.; Ren, C. Degradation-Aware Feature Perturbation for All-in-One Image Restoration. In Proceedings of the CVPR, 2025, pp. 28165–28175.
258. Liu, Y.F.; Jaw, D.W.; Huang, S.C.; Hwang, J.N. DesnowNet: Context-aware deep network for snow removal. *IEEE Trans. Image Process.* **2018**, *27*, 3064–3073.
259. Li, R.; Cheong, L.F.; Tan, R.T. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In Proceedings of the CVPR, 2019, pp. 1633–1642.
260. 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 CVPR, 2024, pp. 25432–25444.
261. Yang, S.; Ding, M.; Wu, Y.; Li, Z.; Zhang, J. Implicit neural representation for cooperative low-light image enhancement. In Proceedings of the ICCV, 2023, pp. 12918–12927.
262. Song, S. Noise-Resilient Low-Light Image Enhancement with CLIP Guidance and Pixel-Reordering Subsampling. *Electronics* **2025**, *14*, 4839.
263. Yan, Q.; Shi, K.; Feng, Y.; Hu, T.; Wu, P.; Pang, G.; Zhang, Y. HVI-CIDNet+: Beyond Extreme Darkness for Low-Light Image Enhancement. *arXiv preprint arXiv:2507.06814* **2025**.
264. Zhang, H.; Sindagi, V.; Patel, V.M. Image de-raining using a conditional generative adversarial network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3943–3956.
265. Wei, X.; Ye, X.; Mei, X.; Wang, J.; Ma, H. A single image deraining algorithm guided by text generation based on depth information conditions. *Appl. Soft Comput.* **2025**, p. 113506.
266. Rajagopalan, S.; Patel, V.M. AWRaCLe: All-weather image restoration using visual in-context learning. In Proceedings of the AAAI, 2025, Vol. 39, pp. 6675–6683.
267. Zhang, K.; Mo, L.; Chen, W.; Sun, H.; Su, Y. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. *arXiv preprint arXiv:2306.10012* **2024**.
268. Yu, Q.; Chow, W.; Yue, Z.; Pan, K.; Wu, Y.; Wan, X.; Li, J.; Tang, S.; Zhang, H.; Zhuang, Y. AnyEdit: Mastering Unified High-Quality Image Editing for Any Idea. *arXiv preprint arXiv:2411.15738* **2025**.
269. Brooks, T.; Holynski, A.; Efros, A.A. InstructPix2Pix: Learning to follow image editing instructions. In Proceedings of the CVPR, 2023, pp. 18392–18402.
270. Zhang, Z.; Xie, J.; Lu, Y.; Yang, Z.; Yang, Y. In-Context Edit: Enabling Instructional Image Editing with In-Context Generation in Large Scale Diffusion Transformer. *arXiv preprint arXiv:2504.20690* **2025**.
271. Zhao, H.; Ma, X.; Chen, L.; Si, S.; Wu, R.; An, K.; Yu, P.; Zhang, M.; Li, Q.; Chang, B. UltraEdit: Instruction-based Fine-Grained Image Editing at Scale. *arXiv preprint arXiv:2407.05282* **2024**.
272. Chow, W.; Li, L.; Kong, L.; Li, Z.; Xu, Q.; Song, H.; Ye, T.; Wang, X.; Bai, J.; Xu, S.; et al. EditMGT: Unleashing Potentials of Masked Generative Transformers in Image Editing. *arXiv preprint arXiv:2512.11715* **2026**.
273. Cai, Z.; Yeh, C.F.; Xu, H.; Liu, Z.; Meyer, G.; Lei, X.; Zhao, C.; Li, S.W.; Chandra, V.; Shi, Y. DepthLM: Metric Depth From Vision Language Models. *arXiv preprint arXiv:2509.25413* **2025**.
274. Zhang, J.; Zhou, S.; Liu, B.; Kadambi, A.; Fan, Z. SpatialStack: Layered Geometry-Language Fusion for 3D VLM Spatial Reasoning. *arXiv preprint arXiv:2603.27437* **2026**.
275. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS* **2022**, *35*, 25278–25294.
276. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the ACL, 2018, pp. 2556–2565.
277. Ordonez, V.; Kulkarni, G.; Berg, T. Im2Text: Describing images using 1 million captioned photographs. *NeurIPS* **2011**, *24*.
278. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* **2015**.
279. Yuan, L.; Chen, D.; Chen, Y.L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* **2021**.

 doi:10.20944/preprints202606.1854.v1

280. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a visual language model for few-shot learning. *NeurIPS* **2022**, *35*, 23716–23736.
281. Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794* **2022**.
282. 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 CVPR, 2018, pp. 586–595.
283. 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. *NeurIPS* **2017**, *30*.
284. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. MUSIQ: Multi-scale image quality transformer. In Proceedings of the ICCV, 2021, pp. 5148–5157.
285. Yang, C.; Dong, R.; Lam, K.M. Vision-Language Model Guided Image Restoration. *arXiv preprint arXiv:2512.17292* **2025**.
286. Xiao, F.; Feng, J.; Hu, P.; Zhang, D.; Xu, L.; Qin, G.; Li, L.; He, C.; Farsiu, S. Qualiteacher: Quality-conditioned pseudo-labeling for real-world image restoration. *arXiv preprint arXiv:2603.08030* **2026**.
287. Xiao, F.; Hu, P.; Xu, L.; Guo, X.; Qin, G.; Shen, Y.; Fang, C.; Zhang, R.; He, C.; Farsiu, S. Beyond Ground-Truth: Leveraging Image Quality Priors for Real-World Image Restoration. *CVPR* **2026**.
288. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* **2023**.
289. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125* **2022**, *1*, 3.
290. Hurst, A.; Lerer, A.; Goucher, A.P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276* **2024**.
291. Pu, Y.; Zhuo, L.; Zhu, K.; Xie, L.; Zhang, W.; Chen, X.; Gao, P.; Qiao, Y.; Dong, C.; Liu, Y. Lumina-OmniLV: A unified multimodal framework for general low-level vision. *arXiv preprint arXiv:2504.04903* **2025**.
292. Blau, Y.; Michaeli, T. The perception-distortion tradeoff. In Proceedings of the CVPR, 2018, pp. 6228–6237.
293. He, C.; Zhang, R.; Zhang, D.; Xiao, F.; Fan, D.P.; Farsiu, S. Nested Unfolding Network for Real-World Concealed Object Segmentation. *arXiv preprint arXiv:2511.18164* **2025**.
294. Zhu, H.; Luo, M.D.; Wang, R.; Zheng, A.H.; He, R. Deep audio-visual learning: A survey. *Int. J. Autom. Comput.* **2021**, *18*, 351–376.
295. Duan, J.; Yu, S.; Tan, H.L.; Zhu, H.; Tan, C. A survey of Embodied AI: From simulators to research tasks. *IEEE Trans. Emerging Top. Comput. Intell.* **2022**, *6*, 230–244.
296. Wang, T.; Mao, X.; Zhu, C.; Xu, R.; Lyu, R.; et al. EmbodiedScan: A holistic multi-modal 3D perception suite towards Embodied AI. In Proceedings of the CVPR, 2024, pp. 19757–19767.
297. Zhang, C.; Gong, D.; He, J.; Zhu, Y.; Sun, J.; Zhang, Y. UIR-LoRA: Achieving Universal Image Restoration through Multiple Low-Rank Adaptation. *arXiv preprint arXiv:2409.20197* **2024**.
298. Ren, P.; Xiao, Y.; Chang, X.; Huang, P.Y.; Li, Z.; Chen, X.; Wang, X. A comprehensive survey of neural architecture search: Challenges and solutions. *CSUR* **2021**, *54*, 1–34.
299. Wang, P.; Luo, X.; Xie, Y.; Qu, Y. Data-free Distillation with Degradation-prompt Diffusion for Multi-weather Image Restoration. *arXiv preprint arXiv:2409.03455* **2024**.
300. He, C.; Shen, Y.; Fang, C.; Xiao, F.; Tang, L. Diffusion Models in Low-Level Vision: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**.
301. Chen, D.; Wang, D.; Darrell, T.; Ebrahimi, S. Contrastive test-time adaptation. In Proceedings of the CVPR, 2022, pp. 295–305.
302. Tang, J.; Chen, J.; Wei, W.; Xu, X.; Liu, R.; Wu, X.; Xie, Q.; Wu, J.; Zhang, L.; Chen, Q. Robust-R1: Degradation-Aware Reasoning for Robust Visual Understanding. *arXiv preprint arXiv:2512.17532* **2025**.
303. 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**.
304. Orjuela, D.Y.G.; Scappatura, L.; Di Gennaro, V.; Izzo, R.A.; Bardaro, G.; Matteucci, M. Improving Robustness of Vision-Language-Action Models by Restoring Corrupted Visual Inputs. *arXiv preprint arXiv:2602.01158* **2026**.
305. Poria, S.; Majumder, N.; Hung, C.Y.; Bagherzadeh, A.A.; Li, C.; Kwok, K.; Wang, Z.; Tan, C.; Wu, J.; Hsu, D. 10 open challenges steering the future of vision-language-action models. In Proceedings of the AAAI, 2026, pp. 39771–39779.

 doi:10.20944/preprints202606.1854.v1