INTERPRETABLE IMAGE RESTORATION VIA MINIMUM DESCRIPTION LENGTH ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

All-in-one image restoration must balance model capacity, computational cost, and the disentanglement of degradations. Conventional Mixture-of-Experts routing relies on heuristic probabilities, producing opaque decisions and potential expert collapse. Grounded in the Minimum Description Length (MDL) principle, we propose MDLE(Minimum Description Length-routed Experts), an interpretable restoration framework that selects experts by jointly minimizing the coding length comprising model complexity and data-fitting error. Its Asymmetric Heterogeneous Expert Pool combines FFT-Attention for global frequency correction with spatial experts for local structure preservation, enabling degradation-aware on-demand computation. Experiments on restoration benchmarks show competitive reconstruction quality and efficient inference. Ablations further demonstrate that routing decisions correlate with degradation types, providing an information-theoretic account of expert selection rather than an empirical black-box gate. Although the quantitative gains over existing multi-routing approaches are currently modest, MDLE offers a distinct perspective on image denoising, dehazing, and other restoration tasks by grounding routing in an explicit information-theoretic objective. This formulation may provide a foundation for further advances and help establish a new direction for interpretable, degradation-aware image processing.

1 INTRODUCTION

Image restoration is a fundamental task across computer vision and graphics, providing essential preprocessing support for high-level visual applications (Jiang et al., 2025). The field has evolved considerably from traditional handcrafted prior-based methods to modern deep learning techniques, transitioning from task-specific paradigms that design dedicated architectures for individual restoration subtasks, including denoising (Zhang et al., 2017a; Chen & Pock, 2016), dehazing (Cai et al., 2016; Ren et al., 2016), and deblurring (Nah et al., 2017; Sun et al., 2015). Nevertheless, such specialized models suffer from heavy computational overhead and poor generalization, failing to handle the complex degradations prevalent in real-world scenarios (Guo et al., 2024b; Zamir et al., 2021).

To tackle these limitations, recent efforts have advanced all-in-one restoration paradigms that unify diverse restoration tasks within a single network (Jiang et al., 2025). State-of-the-art unified methods adopt powerful architectures such as Transformers (Zamir et al., 2022; Wang et al., 2022; Liu et al., 2021) and State Space Models (Mamba) (Gu & Dao, 2023; Guo et al., 2024a) to capture long-range spatial dependencies in high-resolution images. Additionally, integrating multi-modal guidance, including depth priors and text prompts, has become a prevalent strategy to assist accurate degradation perception (Potlapalli et al., 2023; Zuo et al., 2026; Conde et al., 2024).

Despite the remarkable advances in all-in-one image restoration, existing methods suffer from a fundamental limitation: pure data-driven optimization without explicit physical constraints. Current frameworks heavily rely on heuristic routing and empirical probability modeling to accommodate diverse degradations (Li et al., 2022; Riquelme et al., 2021; Shazeer et al., 2017). Such black-box designs fail to precisely disentangle heterogeneous degradation characteristics while maintaining computational efficiency (Fedus et al., 2022; Jiang et al., 2025). The lack of interpretability obscures internal degradation perception mechanisms and impedes input-adaptive dynamic resource allocation for real-world complex degradation scenarios.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To mitigate this fundamental interpretability, this work leverages the Minimum Description Length (MDL) principle from information theory (Rissanen, 1978). Image restoration is formalized as a joint optimization problem balancing model complexity and data-fitting error. As a rigorous and mathematically complete theorem, MDL constructs a unified framework for generalized model selection tasks including expert assignment and feature screening, eliminating the benchmark inconsistency inherent to empirical optimization methods. Built on this insight, MDLE is proposed as a novel unified restoration framework, which replaces empirical heuristic routing with information-cost calculation. Equipped with this new paradigm, MDLE achieves precise disentanglement of complex degradations and endows MoE routing decisions with explicit physical interpretability.

2 Related Work

2.1 Task-Specific Image Restoration

Traditional methods exploit handcrafted statistical priors to restore corrupted images, with classic designs including BM3D (Dabov et al., 2007) for denoising and the dark channel prior (He et al., 2010) for dehazing. Though effective for simple, controlled scenarios, these prior-based methods lack generalization to complex real-world corruptions. With the prevalence of deep learning, CNN-based models such as DnCNN (Zhang et al., 2017a), IRCNN (Zhang et al., 2017b), and RIDNet (Anwar & Barnes, 2019) leverage residual learning to deliver prominent performance gains for individual restoration tasks. Nevertheless, all task-specific paradigms require independent architectures for different degradations, resulting in severe model redundancy, poor scalability, and limited adaptability to the corruptions ubiquitous in practical imaging systems.

2.2 All-in-One Restoration and MoE Architectures

To address the scalability and generalization limitations of task-specific models, unified all-in-one restoration frameworks have risen as the dominant solution for universal image restoration. Equipped with powerful long-range dependency modeling, Transformer-based architectures including SwinIR (Liang et al., 2021), Uformer (Wang et al., 2022), MPRNet (Zamir et al., 2021), and Restormer (Zamir et al., 2022) achieve strong performance across diverse degradation types. Further attempts incorporate neural architecture search (CLEARER (Gou et al., 2020)) and multimodal learning (MPerceiver (Ai et al., 2024)) to further optimize overall restoration capability.

Sparse Mixture-of-Experts (MoE) architectures have demonstrated remarkable advantages in balancing model capacity and computational efficiency, achieving great success in both NLP and computer vision. Representative designs such as Switch Transformer (Fedus et al., 2022), GLaM (Du et al., 2022), V-MoE (Riquelme et al., 2021), and SwinV2-MoE (Cai et al., 2024) validate the efficacy of adaptive expert routing for large-scale visual modeling. In image restoration, recent MoE-based works including UniRestorer (Lin et al., 2024), Complexity Experts (Zamfir et al., 2025), and DSMoE (Lv et al., 2025) introduce sparse expert selection to accommodate multi-degradation restoration tasks. Notably, Wang et al. (2024) proposes adaptive frequency-spatial decoupling with heterogeneous experts, sharing a similar architectural philosophy with our work. However, AdaptIR relies on data-driven heuristic gating for routing decisions, lacking explicit theoretical constraints on expert selection. In contrast, our MDLE is the first to introduce the Minimum Description Length (MDL) principle to formally govern heterogeneous expert routing, replacing empirical heuristics with rigorous information-theoretic optimization.

2.3 Information-Theoretic Learning

Existing restoration frameworks depend entirely on data-driven heuristic gating networks for routing decisions, lacking rigorous theoretical constraints and mathematical interpretability. Such empirical designs easily cause unstable expert assignment and benchmark inconsistency, restricting their reliability in real-world scenarios. While a handful of existing works have begun to leverage information theory to construct model frameworks, they have not yet attempted to guide model construction with complete, physics-driven mathematical principles. To resolve this intrinsic gap, our work innovatively introduces the Minimum Description Length (MDL) principle, a mature information-theoretic framework widely applied in image compression and feature selection (Helin et al., 2017; Li et al.,

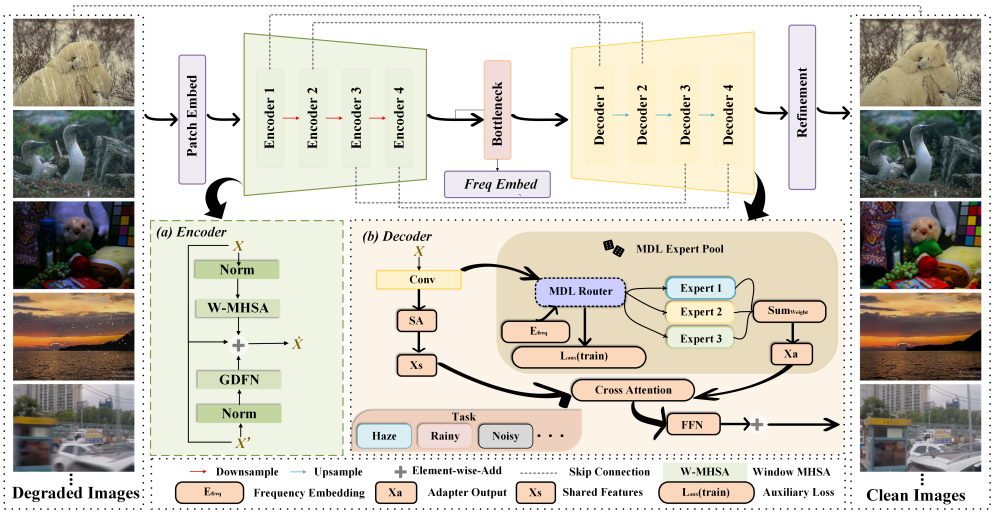


Figure 1: Architecture Overview. A hierarchical encoder-decoder structure designed to capture multi-scale degradation representations.

2017; Hu et al., 2024; Kawaguchi et al., 2023), into MoE-based restoration. Unlike conventional heuristic MoE methods, our MDLE framework formalizes routing as principled information cost optimization, unifying model complexity regularization and data fitting optimization to yield interpretable, stable, and high-performance degradation decoupling.

3 Methodology

3.1 Overall Architecture

We propose MDLE, an interpretable Mixture-of-Experts framework grounded in the MDL principle, to address the opaque gating and lack of principled guidance in conventional MoE restoration models. As illustrated in Figure 1, the framework comprises three coupled components: the Perception, Decision, and Execution layers.

3.1.1 Perception Layer: Multi-scale Progressive Encoder (MPE)

The perception layer adopts a four-stage progressive pyramid structure to extract hierarchical degradation-aware features. Skip connections deliver multi-level features encompassing local textures and global semantics to the decoder, providing comprehensive cues for image restoration.

Specifically, shallow layers $\{L_1, L_2\}$ preserve high-resolution spatial details and capture high-frequency corruptions (e.g., additive noise and pixel loss) via dense local feature learning. Deep layers $\{L_3, L_4\}$ leverage strided downsampling to aggregate global context, perceiving long-range degradation characteristics such as global distortion and large-area haze.

To support globally consistent routing, multi-scale encoder outputs are simultaneously forwarded to the decoder and aggregated into a unified degradation discrimination vector for MDL routing. This hierarchical fusion builds a complete degradation evidence chain to support reliable expert assignment before pixel-level restoration.

3.1.2 Execution Layer: Asymmetric Heterogeneous Expert Pool (AHEP)

Our core innovation lies in the tight coupling of decision and execution layers via the proposed Asymmetric Heterogeneous Expert Pool (AHEP). Unlike conventional MoE models adopting homogeneous expert designs, AHEP constructs physics-aligned, complexity-stratified experts tailored for distinct degradation types:

- **FFT-Attention Expert**: A frequency-domain module for global wavefront correction, targeting global degradations including haze, chromatic aberration, and phase bias.
- **ConvNeXt Expert**: A spatial-domain module optimized for local structural preservation to handle photon noise, motion blur, and texture distortion.
- **SparseMLP Expert**: A lightweight module for simple degradation processing, effectively reducing computational overhead for low-complexity inputs.

Each expert is quantitatively stratified in network depth, feature rank, and receptive field size. This physically grounded design ensures optimal computation-restoration trade-offs for different degradation tasks and eliminates the computational redundancy of symmetric homogeneous expert pools.

3.1.3 Decision Layer: Transparent MDL Routing Function

Different from conventional empirical gating networks, our MDL routing formulates expert dispatching as a principled mathematical optimization. The router takes multi-scale spatial features and frequency embeddings as input, with two coordinated core units:

- **Data-fitting MDL Calculation Unit**: Quantifies the residual description length of input images under each expert's reconstruction, reflecting real-time restoration performance.
- **Model Complexity Storage Unit**: Predefines the parameter description length of each expert to measure activation computational cost.

By combining these two components, the system generates gating scores rigorously grounded in information theory, rather than arbitrary heuristic weights. The entire decision process is driven by the MDL objective, enabling traceable, input-adaptive computational resource allocation. During training, we introduce Gaussian noise exploration and load-balancing regularization to mitigate expert collapse and ensure stable convergence under the overall description length minimization objective. During inference, the system performs top-M expert scheduling directly based on the mathematically derived gating scores, followed by weighted fusion of expert outputs. This paradigm of *perceiving degradation scales, evaluating fitting costs, activating targeted paths* not only enhances generalization for large-field high-resolution imaging tasks, but also endows the network with logical transparency and physical plausibility in dynamic environments.

3.2 Theoretical Foundation: MDL-Guided Routing Formulation

3.2.1 MDL Principle and Two-Part Code

Rooted in information coding theory and Occam's Razor, the Minimum Description Length (MDL) principle (Rissanen, 1978) states that the best model for data achieves the shortest total code length. We formalize expert selection under the *two-part code* formulation: given a degraded image x and a set of K restoration experts $\{e_1, \ldots, e_K\}$, the total description length decomposes as:

$$L(x, e_k) = \underbrace{L(x \mid e_k)}_{\text{data term}} + \underbrace{L(e_k)}_{\text{model term}} . \tag{1}$$

The optimal routing selects the expert minimizing the total description length:

$$e^* = \arg \min_{k \in \{1,\ldots,K\}} \left[L(x \mid e_k) + L(e_k) \right] . \tag{2}$$

3.2.2 Data Term: Log-Reconstruction Error

Let $\hat{x}_k = \mathcal{F}_k(x; \theta_k)$ denote the reconstructed output of expert e_k. To quantify the data description length, each expert first projects the input x into a low-rank subspace via its projection layer $\mathbf{W}_k \in \mathbb{R}^{r_k \times C}$ and then reconstructs it back: $\hat{x}_k = \mathbf{W}_k^\top \mathbf{W}_k x$. The reconstruction error (MSE) is then converted to a description length via the logarithm:

$$L(x \mid e_k) = \log_2(\text{MSE}_k(x) + \epsilon) , \tag{3}$$

where $\text{MSE}_k(x) = \frac{1}{N}\|x - \hat{x}_k\|_2^2$, $N = H \times W \times C$ and $\epsilon = 10^{-8}$ prevents numerical instability. The logarithmic mapping is a standard information-theoretic conversion: under a Gaussian residual assumption, the optimal code length is proportional to $\log(\text{MSE})$ (Hansen & Yu, 2001). Intuitively, the better expert e_k reconstructs x, the smaller the residual, and the fewer bits are needed.

3.2.3 MODEL TERM: LOG-PARAMETER COUNT

The model term $L(e_k)$ encodes the cost of describing expert e_k to the receiver. We define it as the logarithm of the expert's total parameter count:

$$L(e_k) = \log_2(|\theta_k|)\,, \tag{4}$$

where $|\theta_k| = \sum_j |\theta_k^{(j)}|$ is the total number of trainable parameters in expert e_k. This term is *input-independent*: it depends only on the expert's architectural complexity, not on the specific degraded image x. A larger expert with more parameters incurs a higher model description cost.

3.2.4 ROUTING SCORE WITH AUXILIARY FUSION

The total MDL cost for expert e_k on input x is:

$$\text{MDL}_k(x) = L(x \mid e_k) + L(e_k), \tag{5}$$

which equals $\log_2(\text{MSE}_k(x)+\epsilon)+\log_2(|\theta_k|)$. We convert this to a routing score by negation (lower MDL $\rightarrow$ higher score). To incorporate complementary degradation cues beyond reconstruction error, we fuse global spatial features and frequency-domain embeddings as auxiliary routing signals:

$$S_k(x) = -\text{MDL}_k(x) + \alpha\,[g_k(x) + h_k(f_{\text{freq}})]\,, \tag{6}$$

where $g_k(x) = \mathbf{w}_k^\top \text{GAP}(x)$ is a global average pooling score, $h_k(f_{\text{freq}}) = \mathbf{v}_k^\top f_{\text{freq}}$ is a frequency embedding score, and $\alpha = 0.1$ controls their weight. These auxiliary terms provide degradation-type awareness that complements the MDL signal.

The final routing probability is obtained via softmax:

$$p_k(x) = \frac{\exp(S_k(x))}{\sum_{j=1}^{K} \exp(S_j(x))}. \tag{7}$$

At inference, the top-M experts with the largest $p_k(x)$ are activated, and their outputs are fused as $\hat{x} = \sum_{k\in\text{top-}M} p_k(x)\cdot\hat{x}_k$.

3.2.5 MDL-BASED AUXILIARY LOSS

To ensure stable expert utilization during training, we define an auxiliary loss that jointly penalizes MDL imbalance and expert weight imbalance:

$$\mathcal{L}_{\text{aux}} = \underbrace{\text{Var}(\bar{m}_1,\ldots,\bar{m}_K)}_{\text{MDL variance}} + \underbrace{\text{Var}(\bar{p}_1,\ldots,\bar{p}_K)}_{\text{weight variance}}, \tag{8}$$

where $\bar{m}_k = \mathbb{E}_x[p_k(x)\cdot\text{MDL}_k(x)]$ is the batch-averaged weighted MDL cost of expert k, and $\bar{p}_k = \mathbb{E}_x[p_k(x)]$ is the batch-averaged routing probability of expert k. The first term prevents any single expert from accumulating disproportionately high MDL cost; the second term prevents routing collapse to a small subset of experts.

During training, Gaussian noise $\eta \sim \mathcal{N}(0,\delta^2)$ with $\delta = 1/K$ is added to the routing scores $S_k(x)$ to encourage exploration and prevent premature expert collapse. At inference, noise is removed and routing is deterministic.

Different from conventional black-box restoration, our MDLE integrates the MDL principle to realize fully interpretable routing. Instead of empirical fitting, our framework conducts dynamic rate-distortion optimization in feature space, adaptively balancing reconstruction fidelity and computational cost. This paradigm provides a rigorous information-theoretic solution for asymmetric decoupling of heterogeneous real-world degradations.

Furthermore, the physics-aligned AHEP design incorporates complementary spatial and frequency-domain modeling. The FFT-Attention expert efficiently corrects global phase distortions at $\mathcal{O}(N\log N)$ complexity, which is intractable for spatial convolutions, while ConvNeXt and SparseMLP experts handle local textures and simple degradations. Such heterogeneous expert division breaks the limitations of homogeneous MoE designs and implements sensing-computing consistency. By independently modulating amplitude and phase information in the frequency domain, our method effectively addresses complex coupled degradations including chromatic aberration, astigmatism, and mixed noise.

Table 1: Quantitative Comparison on Three Challenging Datasets. All results are reported as PSNR (dB, ↑) / SSIM (↑) on the full RGB images. The **best** performances are highlighted.

Method	Params.	Dehazing SOTS	Deraining Rain100L	Denoising $BSD68_{\sigma=15}$	Denoising $BSD68_{\sigma=25}$	Denoising $BSD68_{\sigma=50}$	Average
IDR	15M	29.87/.970	36.03/.971	33.89/.931	31.32/.884	28.04/.798	31.83/.911
AirNet	9M	27.94/.962	34.90/.967	33.92/.933	31.26/.888	28.00/.797	31.20/.910
PromptIR	36M	30.58/.974	36.37/.972	33.98/.933	31.31/.888	28.06/.799	32.06/.913
MoCE-IR	25M	31.24/.979	38.57/.984	34.11/.932	31.45/.888	28.18/.800	32.73/.917
DSwinIR	24M	31.86/.980	37.73/.983	34.12/.933	31.59/.890	**28.31/.803**	32.72/.917
Restormer	26M	27.78/.958	33.78/.958	33.72/.865	30.67/.865	27.63/.792	30.75/.901
MDLE (ours)	11M	**34.44/.975**	**39.65/.986**	**35.28/.953**	**33.41/.942**	27.98/.865	**34.15/.944**

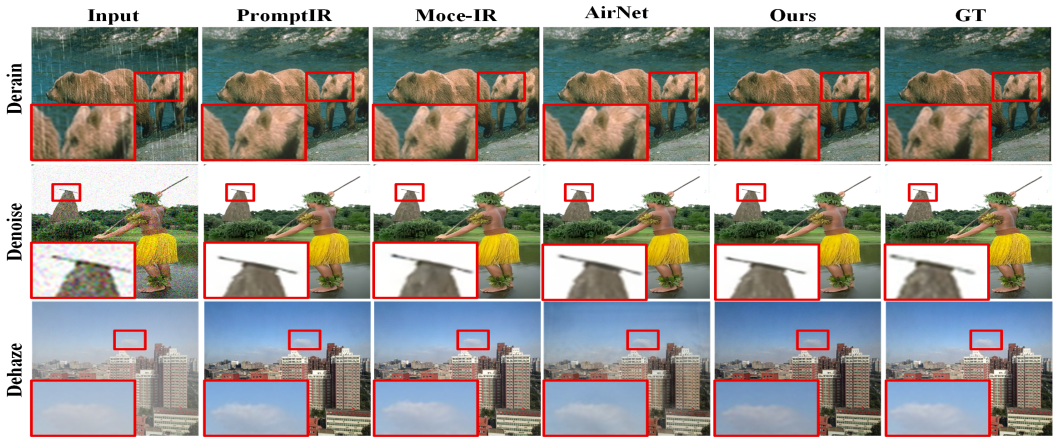


Figure 2: Visual comparison of the three degradation. The red box area is observed from a magnified perspective to achieve a better visualization effect. The task types from top to bottom are rain removal, noise reduction, and haze removal.

4 Experiment

4.1 Experimental Settings

Experiments are conducted on two NVIDIA A40 GPUs, using AdamW optimizer with an initial learning rate of 2×10^{-4} and cosine annealing scheduling. All input images are cropped into 128×128 patches with a batch size of 16. We adopt an end-to-end training strategy for 150 epochs on mixed degradation data. This design is consistent with prior all-in-one methods (Zhang et al., 2024; Potlapalli et al., 2023), where end-to-end training converges rapidly due to the joint optimization of the backbone, MDL router, and expert pool.

4.2 Comparisons with SOTA Methods

We compare MDLE with state-of-the-art restoration models including Restormer (Zamir et al., 2022), AirNet (Chen et al., 2022), PromptIR (Potlapalli et al., 2023), DSwinIR (Zhou et al., 2024), IDR (Chen et al., 2024), and MoCE-IR (Zhang et al., 2024), covering fog, rain, and Gaussian noise degradations. Evaluation is conducted on three widely-used benchmarks: SOTS (Li et al., 2019) for dehazing, Rain100L (Fu et al., 2017) for deraining, and BSD68 (Martin et al., 2001) for denoising. As shown in Table 1, MDLE achieves consistent improvements in PSNR and SSIM. Specifically, for the mixed all-in-one task, MDLE outperforms the best baseline by 1.4 dB in PSNR. This benefit stems from our theory-driven MDL routing: unlike traditional gating mechanisms prone to expert collapse, MDLE's routing score precisely allocates resources based on degradation cost. Quantitative results show that in mixed scenarios with high-frequency noise and global phase distortion,

the activation rate of the FFT-Attention expert is highly correlated with degradation complexity, verifying the high resource allocation efficiency of our framework.

Visual comparisons in Figure 2 further validate MDLE's physical-level decoupling capability. On SOTS, Rain100L, and BSD68 datasets, traditional models fail to eliminate global artifacts and contrast attenuation caused by phase bias, leading to obvious dispersion stripes at image edges. In contrast, MDLE, with its asymmetric heterogeneous expert pool and the frequency-domain modulation of the FFT-Attention expert, effectively restores image wavefront integrity, significantly enhancing edge sharpness and color fidelity. MDLE not only filters local photon noise but also corrects macroscopic aberrations, achieving full-scale coverage from pixel-level repair to global wavefront correction. This consistent sensing-computing performance verifies the model's generalization and practical value for real optical imaging tasks.

We further evaluate on single-task benchmarks: GOPRO for deblurring, and LOL for low-light enhancement. Compared with baselines including AirNet, PromptIR, DSwinIR, IDR, MoCE-IR, and M2IR, MDLE outperforms MoCE-IR by 3.2 dB in PSNR for deblurring, and surpasses M2IR by 0.8 dB for low-light enhancement, as reported in Table 2. The corresponding visual comparisons in Figure 3 further show that MDLE restores sharper structures and more faithful illumination than the baselines, validating the effectiveness of our restoration capability.

Table 2: Quantitative Comparison of the deblurring and the Low-light tasks. All results are reported as PSNR(dB,↑)/SSIM(↑) on the full RGB images. The **best** performances are highlighted.

Method	GOPRO	Method	LOL
AirNet	24.35/.781	AirNet	18.18/.735
PromptIR	28.71/.881	PromptIR	22.68/.832
DSwinIR	29.17/.879	DSwinIR	22.64/.843
IDR	27.87/.846	IDR	21.34/.826
M2IR	29.09/.878	M^2IR	23.80/.864
MoCE-IR	30.05/.899	MoCE-IR	23.00/.852
MDLE (ours)	**33.32/.960**	MDLE (ours)	**24.63/.903**



Figure 3: Visual comparisons of the deblurring and the Low-light tasks. On the top is the restoration of low-light images using the LOL dataset, and below is the restoration of blurred images using the GOPRO dataset.

4.3 Ablation Study and Interpretability Analysis

Factorial Ablation Design. We evaluate the routing objective and the expert-pool structure as two independent factors, using the same datasets as Table 1: SOTS, Rain100L, and BSD68. In a 2×2 factorial ablation, we compare a homogeneous FFT-Attention pool with the heterogeneous pool, under either the spatial–frequency feature router or the MDL router. All four variants use the same three-expert Top-1 sparse MoE, data split, optimization schedule, and checkpoint-selection protocol. This design separates the effect of routing from the effect of expert-pool heterogeneity. Both pools share the same parameter sets: patch sizes $p \in \{4, 8, 16\}$, depths $d \in \{3, 5, 7\}$, and ranks $r \in \{8, 16, 32\}$. The distinction is that the homogeneous pool uses only FFT-Attention experts, whereas the heterogeneous pool uses structurally distinct experts—FFT-Attention, ConvNeXt, and SparseMLP—matching the AHEP design introduced in Section 3.1.2.

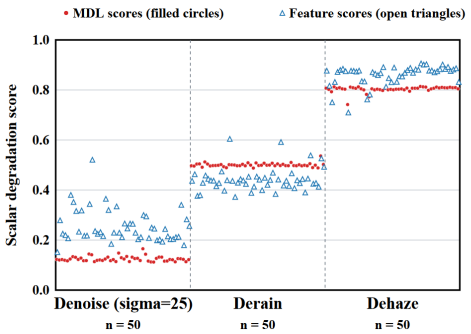


Figure 4: Degradation-score distributions comparison between the MDL router and the spatial–frequency feature router. Each task contains 50 samples. Denoising uses Gaussian noise with $\sigma = 25$. Both experiments use the unified heterogeneous pool.

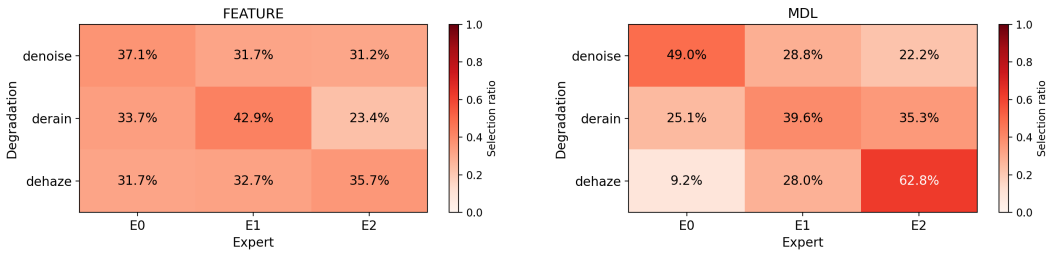
Table 3: 2×2 ablation of expert-pool structure and routing objective."Feature" denotes the spatial–frequency feature router and "MDL" denotes the MDL router. PSNR and SSIM are averaged over the evaluated degradation types. FLOPs measured on 256×256 input (NVIDIA A40 GPU).

Router	Expert pool	PSNR (dB) ↑	SSIM ↑	Accuracy ↑	Runtime (ms) ↓	FLOPs (G) ↓
Feature	Homogeneous	32.05	.935	91.53%	103.97	58.60 ± 1.61
Feature	Heterogeneous	32.65	.935	94.35%	118.60	58.78 ± 2.50
MDL	Homogeneous	33.23	.938	95.20%	112.63	58.81 ± 2.56
MDL	Heterogeneous	34.42	.944	98.95%	120.95	59.06 ± 3.06

Factorial Results. Both factors are individually beneficial and super-additive in combination: replacing the homogeneous pool with the heterogeneous pool improves PSNR by $+0.60$ dB under feature routing and $+1.19$ dB under MDL routing, while replacing the feature router with the MDL router improves PSNR by $+1.18$ dB (homogeneous) and $+1.77$ dB (heterogeneous), yielding the best 34.42 dB / 0.944 SSIM when combined. Accuracy follows the same trend, reaching 98.95% only when both factors are active together. Notably, the shared backbone accounts for the majority of compute, as in all MoE restoration frameworks; our design does not reduce backbone cost but ensures the *expert-specific* computation is minimal on average.

Degradation Score. We also expose a scalar degradation score $s_{\text{deg}}(x) \in [0, 1]$ using a sigmoid calibrator whose inputs include stop-gradient-detached routing costs, adapter features, and frequency embeddings from the decoder stages. The complete factorial results are reported in Table 3; the corresponding score distributions are visualized in Figure 4. As shown in the figure, the MDL-derived score is more stable than that of the traditional feature router. The routing accuracy in Table 3 is obtained by thresholding this degradation score to classify the degradation type. For clarity, denoising uses Gaussian noise with $\sigma = 25$ as a typical case.

Expert Routing Comparison. To isolate the effect of routing alone, we compare the two routing strategies on the *same* heterogeneous pool (FFT-Attention, ConvNeXt, and SparseMLP). Comparing the heatmaps in Figure 5, the MDL router exhibits clear hierarchical expert allocation: dehazing preferentially activates deep experts, denoising favors lightweight experts, and deraining primarily engages medium-to-large experts. In contrast, the feature router's assignments lack this structured pattern, confirming that the MDL objective enables degradation-aware, hierarchical expert selection.

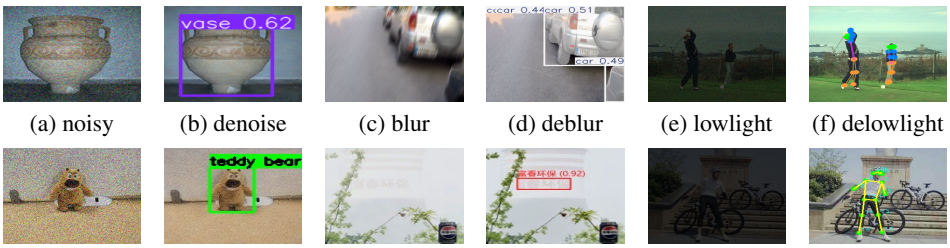


(a) (b)

Figure 5: Comparison of expert routing heatmaps. (a) Feature router; (b) MDL router.

4.4 Downstream Task Generalization

To further validate the practical applicability and cross-scenario generalization of MDLE for high-level vision tasks, we conduct downstream evaluations across four challenging real-world scenarios (Figure 6): object detection from noisy inputs, instance segmentation on blurry images, text detection in degraded scenes, and human pose estimation under extreme low-light noisy conditions. Real-world images are captured by an MV-CU60-10GC industrial camera to simulate practical imaging conditions. Raw degraded images can hardly support reliable task execution, while our method



(a) noisy (b) denoise (c) blur (d) deblur (e) lowlight (f) delowlight

(g) RW-N (h) RW-DN (i) RW-H (j) RW-DH (k) RW-LOL (l) RW-DLL

Figure 6: Downstream task and real-world degradation results. First row—synthetic: (a) noisy, (b) denoise, (c) blur, (d) deblur, (e) lowlight, (f) delowlight. Second row—real-world scenes captured by an MV-CU60-10GC industrial camera: (g) RW-N: real-world noisy, (h) RW-DN: real-world denoise, (i) RW-H: real-world haze, (j) RW-DH: real-world dehaze, (k) RW-LOL: real-world lowlight, (l) RW-DLL: real-world delowlight.

effectively recovers clear structural textures and valid visual semantics from severe corruption, enabling downstream models to stably complete these tasks and produce credible predictions. Benefiting from MDL-based adaptive routing that dynamically combines expert responses (e.g., simultaneously activating frequency-domain correction and spatial-domain dehazing paths for other unseen combinations like high-concentration haze with sub-wavelength processing errors), our method exhibits strong generalization to extreme degradation scenarios, with minor contrast deviations in the most severe cases further alleviated by adjusting the routing temperature coefficient τ to balance exploration and exploitation, demonstrating its great practical potential for real-world vision systems.

5 Conclusion

In this work, we propose MDLE, an MDL-based hybrid expert architecture to resolve heterogeneous degradation decoupling and blind resource allocation in computational imaging; by embedding the information-theoretic MDL principle into MoE routing, we construct a perception-decision-execution closed loop, replacing empirical heuristics with mathematically complete theoretical guidance, with two core designs: a transparent routing mechanism that transforms expert assignment from heuristic fitting to theory-driven optimization via the trade-off between data fitting and model complexity, and the Asymmetric Heterogeneous Expert Pool (AHEP) that mitigates diverse noises through scale-frequency decoupling with heterogeneous modules.

Extensive experiments demonstrate that MDLE achieves state-of-the-art all-in-one restoration performance, outperforming existing baselines in PSNR and SSIM; critically, the MDL routing delivers interpretable decisions with expert activations aligning well with physical degradation types, addressing the long-standing black-box reliability issue in optical systems. We note that the routing function incurs minor latency in extreme dynamic or ultra-high-throughput scenarios, thus our future work will explore lightweight distillation for edge deployment, as well as integrating polarization and spectrum signals into MDL quantization to build full-dimensional imaging models.

Overall, this work verifies that unified mathematically rigorous theory can eliminate the intrinsic instability of empirical deep learning methods, and MDLE not only supports high-performance metamaterial imaging but also paves the way for traceable, adaptive intelligent visual systems, bridging information theory and practical deep learning. Although the current model has not been slimmed down and the shared backbone accounts for the majority of FLOPs, the significance of MDLE lies in establishing an alternative, information-theoretic view of adaptive restoration. By treating routing as an explicit trade-off between reconstruction evidence and model complexity. We hope this initial formulation will motivate further exploration of theoretically grounded routing and ultimately develop into a distinct direction for interpretable and adaptive image restoration.

Statement

This paper uses AI tools solely for statistical data processing. Figure 4 and Figure 5 are plotted using AI-generated statistical code. Training and model code involve no AI tools. We reiterate that AI is used only for data statistics; all data are real, and no AI-generated data are included.

REFERENCES

Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25432–25444, 2024.

Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3155–3164, 2019.

Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE transactions on image processing*, 25(11):5187–5198, 2016.

Weilin Cai, Juyong Jiang, Le Qin, Junwei Cui, Sunghun Kim, and Jiayi Huang. Shortcut-connected expert parallelism for accelerating mixture-of-experts. *arXiv preprint arXiv:2404.05019*, 2024.

Hao Chen, Jinjin Gu, Yihao Liu, et al. Idr: Self-supervised image denoising via iterative data refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016.

Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, pp. 1–21. Springer, 2024.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16 (8):2080–2095, 2007.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3860–3868, 2017.

Yuanbiao Gou, Boyun Li, Zitao Liu, Songfan Yang, and Xi Peng. Clearer: Multi-scale neural architecture search for image restoration. *Advances in neural information processing systems*, 33: 17129–17140, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pp. 222–241. Springer, 2024a.

Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In *European conference on computer vision*, pp. 255–272. Springer, 2024b.

M. H. Hansen and B. Yu. *Model Selection and Model Averaging*. Cambridge University Press, 2001.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

Petri Helin, Pekka Astola, Bhaskar Rao, and Ioan Tabus. Minimum description length sparse modeling and region merging for lossless plenoptic image compression. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1146–1161, 2017.

Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5325–5344, 2024.

Junjun Jiang, Zengyuan Zuo, Gang Wu, Kui Jiang, and Xianming Liu. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *International conference on machine learning*, pp. 16049–16096. PMLR, 2023.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2019.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17452–17462, 2022.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Jingbo Lin, Zhilu Zhang, Wenbo Li, Renjing Pei, Hang Xu, Hongzhi Zhang, and Wangmeng Zuo. Unirestorer: Universal image restoration via adaptively estimating image degradation at proper granularity. *arXiv preprint arXiv:2412.20157*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Minxuan Lv, Zhenpeng Su, Leiyu Pan, Yizhe Xiong, Zijia Lin, Hui Chen, Wei Zhou, Jungong Han, Guiguang Ding, Wenwu Ou, et al. Dsmoe: Matrix-partitioned experts with dynamic routing for computation-efficient dense llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19722–19733, 2025.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th IEEE International Conference on Computer Vision*, volume 2, pp. 416–423, 2001.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in neural information processing systems*, 36:71275–71293, 2023.

Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision*, pp. 154–169. Springer, 2016.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 769–777, 2015.

Yan Wang, Jing Li, Yulun Zhang, et al. Adaptir: Efficient image restoration via adaptive frequency-spatial decoupling with mixture of experts. In *Advances in Neural Information Processing Systems*, 2024.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17683–17693, 2022.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12753–12763, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26 (7):3142–3155, 2017a.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3929–3938, 2017b.

Yulun Zhang, Jing Li, et al. Moce-ir: Mixture of context experts for all-in-one image restoration. *arXiv preprint*, 2024.

Man Zhou, Xiangyu Chen, Chongyi Li, et al. Dswinir: Dilated sliding window attention for efficient image restoration. *arXiv preprint arXiv:2401.00000*, 2024.

Zengyuan Zuo, Junjun Jiang, Gang Wu, and Xianming Liu. Udpnet: Unleashing depth-based priors for robust image dehazing. *arXiv preprint arXiv:2601.06909*, 2026.