Zero-Reference Joint Low-Light Enhancement and Deblurring via Visual Autoregressive Modeling with VLM-Derived Modulation

Wei Dong, Han Zhou*, Junwei Lin, Jun Chen
McMaster University
{dongw22, zhouh115, lin523, chenjun}@mcmaster.ca

Abstract

Real-world dark images commonly exhibit not only low visibility and contrast but also complex noise and blur, posing significant restoration challenges. Existing methods often rely on paired data or fail to model dynamic illumination and blur characteristics, leading to poor generalization. To tackle this, we propose a generative framework based on visual autoregressive (VAR) modeling, guided by perceptual priors from the vision-language model (VLM). Specifically, to supply informative conditioning cues for VAR models, we deploy an adaptive curve estimation scheme to modulate the diverse illumination based on VLM-derived visibility scores. In addition, we integrate dynamic and spatial-frequency-aware Rotary Positional Encodings (SF-RoPE) into VAR to enhance its ability to model structures degraded by blur. Furthermore, we propose a recursive phase-domain modulation strategy that mitigates blur-induced artifacts in the phase domain via bounded iterative refinement guided by VLM-assessed blur scores. Our framework is fully unsupervised and achieves state-of-the-art performance on benchmark datasets.

Code — https://github.com/LowLevelAI/VAR-LIDE

Introduction

The degradation of images captured in real-world dark environments can be formulated as: $\mathbf{x}_{LQ} = \gamma f(\mathbf{x}_{HQ}, \mathbf{k}) + \mathbf{n}$, with $\mathbf{x}_{HQ}$ denoting the high-quality (HQ) image and $\mathbf{x}_{LQ}$ its low-quality (LQ) counterpart. Here, $\mathbf{n}$ denotes sensor noise, f is the convolution with the blur kernel $\mathbf{k}$ and γ models dynamic range compression and saturation from exposure. Long exposure, a common strategy to improve photon capture, frequently leads to motion-induced blur and elevated noise levels. These combined artifacts degrade image quality and pose significant challenges for both human perception and high-level vision systems (Xu, Dong, and Zhou 2022).

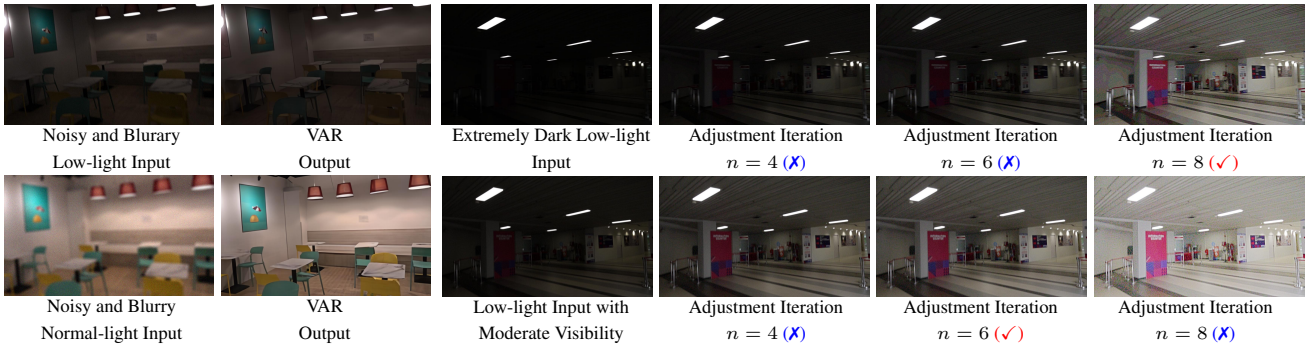
Similar to other image restoration tasks (Zhou et al. 2023; Dong et al. 2024b,a, 2025b), deep learning has led to notable progress in both low-light image enhancement (LLIE) (Cai et al. 2023; Jiang et al. 2021; Li, Guo, and Loy 2022; Ma et al. 2022) and deblurring (Pham et al. 2024; Guo et al. 2024; Dong, Roth, and Schiele 2022), most methods treat them as separate tasks. LLIE models primarily boost brightness and reduce noise but often leave motion blur unaddressed. Conversely, deblurring algorithms typically assume sufficient illumination and perform poorly under low visibility. Although sequential pipelines may appear viable, they often disrupt blur cues during enhancement or fail to recover motion details when visibility is too low.

Although recent efforts explore joint LLIE and deblurring, most rely on supervised pipelines that require paired training data (Feijoo et al. 2025; Zhou, Li, and Change Loy 2022), which is difficult to obtain in real-world scenarios. Though diffusion-based unsupervised methods (Lv et al. 2024) show promising results, their reliance on lengthy sampling steps (*e.g.*, 1,000 iterations) severely limits efficiency, making them unsuitable for practical applications.

Recently, visual autoregressive (VAR) (Tian et al. 2024) models offer a compelling alternative to diffusion methods by progressively generating high-resolution images via scale-wise token prediction, achieving superior structural fidelity and significantly faster inference without relying on costly iterative denoising. Technically, VAR models effectively preserve bidirectional spatial correlations while aligning with the unidirectional nature of autoregressive modeling, making them particularly suitable for image restoration tasks (Qu et al. 2025; Wang et al. 2025), where LQ inputs serve as conditioning signals to guide the generation process. As illustrated in Fig. 1(a), our preliminary experiments indicate that pre-trained VAR models (Qu et al. 2025) possess inherent capabilities for noise suppression and partial blur reduction, making them a promising backbone for reference-free joint LLIE and deblurring.

On the other hand, we observe that this VAR backbone demonstrates insufficient capacity for illumination enhancement and blur-compensated recovery, motivating the integration of specialized components to address visibility and detail degradation more effectively. An intuitive way to enable VAR with illumination correction capability is to incorporate a lightweight enhancement module prior to the generative process, allowing the model to operate on visibility-improved inputs while leveraging its inherent noise suppression properties. In our preliminary exploration, we adopt Zero-DCE (Guo et al. 2020), a lightweight yet highly effective method for real-world LLIE, where deep neural networks are employed to predict the parameters of a differ-

*Corresponding author and project leader



(a) Observations on blurry and noisy inputs with pre-trained VAR models.

(b) Zero-DCE presents unfavorable enhancement outcomes on low-light inputs with varying illumination conditions using a fixed adjustment iteration.

Figure 1: Motivations of our proposed method. (a) We observe that the pre-trained VAR model (Qu et al. 2025) exhibits a certain degree of noise suppression and blur reduction, suggesting its potential for joint LLIE and deblurring. However, it struggles to substantially enhance visibility and recover fine structural details, underscoring the need for task-specific modules to better cope with real-world degradations. (b) Although the iterative illumination adjustment is generally effective, Zero-DCE (Guo et al. 2020) fails to provide satisfactory results under diverse illumination conditions, often causing under- or over-exposure artifacts due to its reliance on a fixed number of adjustment iterations.

entiable curve-based model that iteratively adjusts image illumination. Once training converges, we observe that most pixels in the predicted curve parameter maps exhibit positive values, implying that increased iteration steps correspond to stronger illumination enhancement. However, this introduces two critical issues: (i) for moderately bright images, the default setting of 8 iterations tends to cause overexposure; (ii) reducing the iteration number alleviates overexposure but leads to insufficient enhancement on extremely dark images, as illustrated in Fig. 1(b). These observations indicate that a fixed iteration setting cannot robustly handle the diverse luminance conditions. This motivates further exploration into mechanisms that enable adaptive brightness modulation, ensuring consistent and perceptually compelling enhancement across diverse lighting scenarios.

Building upon these observations and insights, we introduce **VAR-LIDE**, a fully unsupervised generative framework for joint **LLI**E and **DE**blurring, which leverages the strengths of **VAR** models and perceptual guidance from the vision-language model (VLM). To effectively condition the VAR backbone, we develop a VLM-informed conditioning module that predicts adaptive enhancement curves based on VLM-assessed visibility, enabling robust performance under varying lighting conditions. Furthermore, we augment the pre-trained VAR backbone with dynamic and frequency-aware Rotary Positional Encodings to better model spatial structures degraded by motion blur. To mitigate motion-induced repeated edge artifacts in the Fourier phase domain, we introduce a recursive modulation mechanism that progressively refines the phase via a bounded parametric update, guided by blur-related VLM assessments.

We summarize our contributions as following:

◇ We introduce **VAR-LIDE**, a fully unsupervised VAR-based framework that jointly addresses low-light image enhancement and deblurring, leveraging perceptual priors.

◇ Based on perceptual priors derived from VLM, we develop a VLM-informed conditioning module to support informative conditioning for VAR backbone, and design a recursive phase refinement mechanism to suppress blur-induced edge artifacts in the Fourier domain.

◇ We enhance the VAR backbone with content-aware spatial-frequency rotary positional encodings to better capture structural information under blur degradation.

◇ Our **VAR-LIDE** relaxes the reliance on paired supervision and achieves compelling performance on challenging real-world low-light benchmarks.

Related Works

LLIE and Deblurring as Separate Tasks LLIE and deblurring are traditionally handled separately. Early LLIE methods (Pizer et al. 1987; Jobson, Rahman, and Woodell 1997; Wang et al. 2013) used hand-crafted priors, while recent deep models (Guo et al. 2020; Dong et al. 2025a; Zhou et al. 2024; Dong et al. 2024c; Zhou, Dong, and Chen 2025) learn brightness correction but neither are able to remove real-world blur, limiting their practical value. In parallel, traditional deblurring methods utilize predefined kernels for deconvolution (Kundur and Hatzinakos 1996; Anger, Facciolo, and Delbracio 2018), while deep learning models (Dong, Roth, and Schiele 2022; Li et al. 2023; Pham et al. 2024) aim for better generalization. However, these methods assume well-lit inputs, which LLIE results may not meet, leading to artifacts and degraded performance.

Joint LLIE and Deblurring Joint LLIE and deblurring has attracted increasing attention. Supervised methods (Zhou, Li, and Change Loy 2022; Feijoo et al. 2025) depend on costly paired data, whereas unsupervised approaches (Li et al. 2024; Lv et al. 2024) use reconstruction or contrastive objectives to avoid this limitation. The architectures have evolved from CNNs (Zhou, Li, and Change Loy 2022) to transformers (Cai et al. 2023), Mamba (Liu et al. 2025), normalizing flow (Li et al. 2024), and diffusion models (Lv et al. 2024). However, achieving efficient and gener-

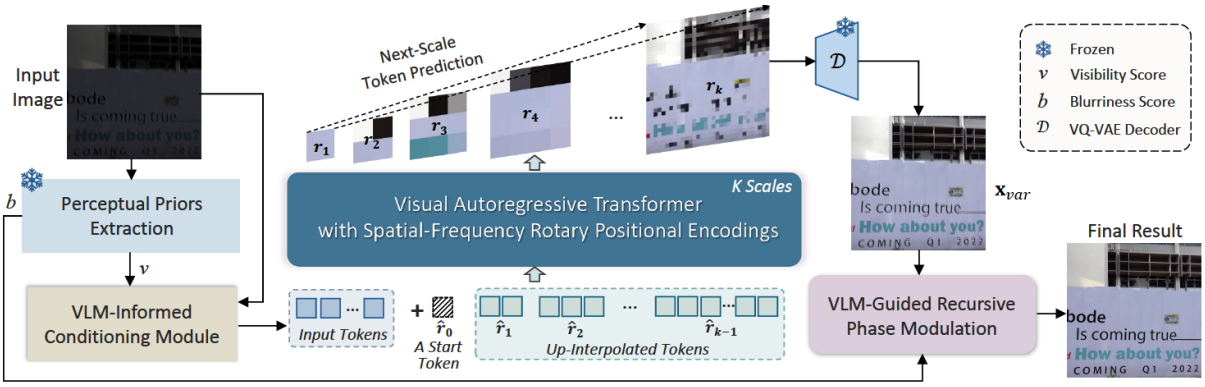


Figure 2: The overall framework of our proposed **VAR-LIDE** method, which adopts the pre-trained VAR model (Qu et al. 2025) as the backbone. We first leverage the perceptual priors extraction pipeline (Zhou et al. 2025) to acquire visibility-aware and blurriness-aware scores (v and b). Then, v is integrated into our VLM-Informed Conditioning Module (VICM) to adaptively improve the visibility and further support informative cues for VAR modeling. Moreover, to generate content-aware representations of positional embeddings, we develop the spatial-frequency rotary positional encodings (SF-RoPE) in VAR transformer blocks. Finally, guided by the VLM assessment b, we introduce a recursive modulation mechanism (VGPM) in the FFT phase domain to further mitigate blurriness and achieve visually compelling outputs.

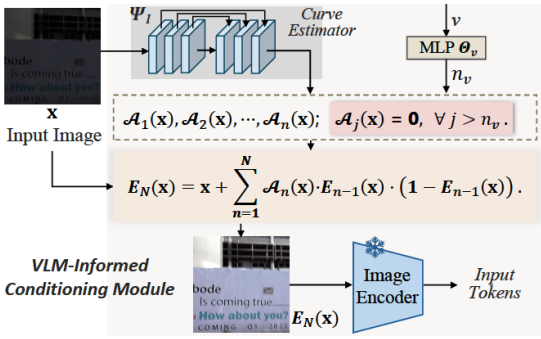


Figure 3: The overall framework of our VICM. It estimates illumination curves and adaptively truncates them based on a visibility-aware iteration count n_v.

alizable joint restoration remains challenging.

Visual Autoregressive Modeling VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017) encodes images as quantized tokens for autoregressive generation but lacks spatial awareness. VAR (Tian et al. 2024) improves this via next-scale token prediction, boosting quality and speed. Though applied to synthesis tasks (Chen et al. 2025; Han et al. 2025), VAR is underexplored in image restoration under degradations like low-light blur. Recent work (Wang et al. 2025; Rajagopalan, Narayan, and Patel 2025; Qu et al. 2025) demonstrates its potential for image restoration by leveraging multiscale priors. We employ the VAR backbone for the joint LLIE and deblurring task, augmented with modules for illumination modulation and blur suppression.

Method

The main focus of this work is to extend the capabilities of pre-trained Visual Autoregressive (VAR) models (Tian et al. 2024) to the challenging task of joint low-light image enhancement (LLIE) and deblurring. While the most relevant prior work, VARSR (Qu et al. 2025) explores scale-aligned rotary positional encodings (SA-RoPE) and diffusion refiners for image super-resolution, our method targets a different problem domain and proposes three novel modules tailored for real-world degradation: (i) a VLM-informed conditioning module for perceptual-aware illumination control (Sec.), (ii) a spatial-frequency RoPE mechanism that fuses FFT-phase guided rotation with scale-aligned spatial encoding (Sec.), and (iii) a recursive phase modulation module that explicitly targets blur-induced phase duplication artifacts (Sec.). Lastly, a reference-free optimization strategy is proposed to enable training without ground-truth supervision (Sec.). Our framework is illustrated in Fig. 2.

VLM-Informed Conditioning Module

In VAR-based restoration (Qu et al. 2025), the low-quality (LQ) input is embedded as prefix tokens to guide the generation process, making the informativeness of these conditional cues crucial for reconstruction fidelity. To strengthen the generative conditioning, we propose a VLM-Informed Conditioning Module (VICM) that adaptively modulates luminance based on perceptual cues.

Our design is motivated by the limitations of heuristic illumination adjustment strategies (*e.g.*, Zero-DCE) in serving as effective conditioning for generative restoration models. Although Zero-DCE improves brightness, it lacks adaptability across diverse lighting conditions. As shown in Fig. 1(b), shallow enhancement (*e.g.*, n=4) results in under-exposure in extremely dark scenes, while deeper enhancement (*e.g.*, n=8) causes overexposure in moderately lit inputs. These suboptimal adjustments (*e.g.*, Fig. 4(a)(ii)) degrade the conditioning quality and propagate artifacts in the generative output (Fig. 4(a)(iii)). Formally, Zero-DCE models enhance-

(i) Low-quality Input | (ii) Zero-DCE Enhanced Result | (iii) VARSR Output with Condition (ii) | (iv) Enhancement of Our Proposed VICM | (v) VARSR Output with Condition (iv) | (vi) GT Reference

(a) Our proposed VICM module provides more effective conditional cues for VAR, achieving more satisfactory results.

(i) Low-quality Input | (ii) w/ VICM only | (iii) + Vanilla RoPE | (iv) + Our SF-RoPE | (v) + VGPM (Full model) | (vi) GT

(b) Progressive Enhancement via VICM, SF-RoPE, and VGPM. Vanilla RoPE is adopted from VARSR (Qu et al. 2025)

Figure 4: Comparative visualization of conditioning effects and enhancement quality across modules in our framework.

ment as an iterative curve-based transformation:

$$\boldsymbol{E}_N(\mathbf{x}) = \mathbf{x} + \sum_{n=1}^{N} \boldsymbol{\mathcal{A}}_n(\mathbf{x}) \cdot \boldsymbol{E}_{n-1}(\mathbf{x}) \cdot (\mathbf{1} - \boldsymbol{E}_{n-1}(\mathbf{x})), \quad (1)$$

where $\boldsymbol{E}_0(\mathbf{x}) = \mathbf{x}$, $\boldsymbol{\mathcal{A}}_n(\mathbf{x})$ denotes the curve parameter at iteration n, and N is the total number of iterations. While this formulation captures nonlinear illumination trends, its fixed-step (N) nature fundamentally limits adaptiveness.

To address this limitation, we incorporate perceptual priors extracted from the vision-language model (VLM) following GPP-LLIE (Zhou et al. 2025). A visibility-aware score v is first computed via the *Global Perceptual Prior Extraction Pipeline* in GPP-LLIE, which is then processed by a lightweight MLP $\boldsymbol{\Theta}_v$ to estimate an optimal iteration count n_v. As illustrated in Fig. 2, this n_v is used to truncate the illumination adjustment process within our VICM. Specifically, the curve estimator $\boldsymbol{\Psi}_I$ produces illumination curves $\{\boldsymbol{\mathcal{A}}_n(\mathbf{x})\}_{n=1}^{N}$, and curve parameters beyond n_v are masked to ensure perceptual adaptiveness:

$$\boldsymbol{\mathcal{A}}_j(\mathbf{x}) = 0, \quad \forall j > n_v. \quad (2)$$

This adaptive truncation ensures that illumination enhancement remains within a perceptually valid range. The enhanced image $\boldsymbol{E}_N(\mathbf{x})$ is then embedded and tokenized as a conditioning input to the VAR model. Compared with fixed-iteration enhancement pipelines, our VICM provides more informative and spatially adaptive guidance (Fig. 4(a)(iv)), thereby improving downstream generation (Fig. 4(a)(v)). Nonetheless, some structural artifacts remain (e.g., Fig. 4(b)(iii)), motivating the design of complementary modules to better handle motion-related degradations.

Content-Aware Spatial-Frequency RoPE

To further alleviate structure-related artifacts observed in blur-degraded regions, we focus on enhancing the positional encoding mechanism within the VAR backbone. This refinement is motivated by the observation that existing rotary encoding schemes, such as those used in VARSR (Qu et al. 2025), rely solely on position-indexed rotation matrices that lack sensitivity to content degradation. While Rotary Positional Encoding (RoPE) enables efficient modeling of relative positions by applying fixed sinusoidal rotations to query-key pairs, its static nature limits adaptability in structurally complex or spatially degraded regions.

To address this limitation, we propose a content-aware spatial-frequency RoPE (SF-RoPE) that modulates attention weights based on both positional and frequency-domain cues. Our approach introduces token-wise rotation matrices whose parameters are dynamically inferred from local frequency-phase statistics, thereby enabling finer control over attention in blur-sensitive areas.

Frequency-domain RoPE At scale K, we extract frequency-phase information from the embedding $\mathbf{x}_{K-1}$ of token map $\mathbf{r}_{K-1}$ via $\texttt{FFT}$:

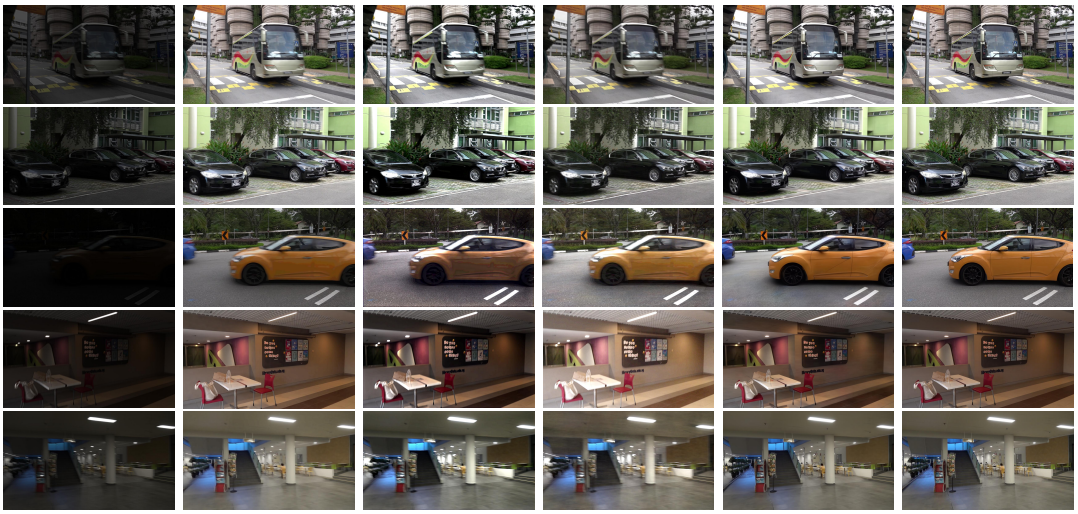
$$\begin{aligned} &\mathbf{F}(u,v) = \texttt{FFT}(\mathbf{x}_{K-1}), \Phi(u,v) = \arg(\mathbf{F}(u,v)), \\ &\mathbf{R}_{\Phi(u,v)} = \begin{bmatrix} \cos(\Phi(u,v)) & -\sin(\Phi(u,v)) \\ \sin(\Phi(u,v)) & \cos(\Phi(u,v)) \end{bmatrix}, \end{aligned} \quad (3)$$

where (u, v) denotes the frequency coordinate and $\Phi(u, v)$ captures the local phase. We then construct a token-specific frequency-based RoPE matrix as:

$$\texttt{RoPE}_{\text{freq}}(\mathbf{x}_K^{(u,v)}) = \begin{bmatrix} \mathbf{R}_{\Phi(u,v)}^{\frac{C}{2}} & \mathbf{0}^{\frac{C}{2}} \\ \mathbf{0}^{\frac{C}{2}} & \mathbf{R}_{\Phi(u,v)}^{\frac{C}{2}} \end{bmatrix}, \quad (4)$$

where C denotes the channel dimension. This enables the attention mechanism to be directly modulated by local blur-sensitive frequency variations.

Scale-aligned RoPE In parallel, we apply standard RoPE in spatial domain using scale-normalized token coordinates to ensure positional consistency across multiple resolutions.



LQ | Retinexformer | SSFlow | FourierDiff | **Ours** | Reference

Figure 5: Visual comparisons on the LOL-Blur dataset, which involves both severe low-light conditions and motion blur. Compared to existing methods, our approach better preserves fine details and improves perceptual quality across diverse scenes.

For each $\mathbf{x}_k^{(i,j)}$, the spatial rotary matrix is computed as:

$$\texttt{RoPE}_{\text{spa}}(\mathbf{x}_k^{(i,j)}) = \begin{bmatrix} \mathbf{R}^{\frac{C}{2}}_{\Theta,\left(\frac{ih_K}{h_k}\right)} & \mathbf{0}^{\frac{C}{2}} \\ \mathbf{0}^{\frac{C}{2}} & \mathbf{R}^{\frac{C}{2}}_{\Theta,\left(\frac{jw_K}{w_k}\right)} \end{bmatrix}, \quad (5)$$

where (i, j) denotes the spatial location, (h_k, w_k) are the dimensions at scale k, and (h_K, w_K) refer to the base scale.

Spatial-Frequency Fusion To adaptively leverage both structural and positional priors, we fuse frequency and spatial encodings using a learnable mixing coefficient λ:

$$\texttt{RoPE}_{\text{fused}} = \lambda \cdot \texttt{RoPE}_{\text{freq}} + (1 - \lambda) \cdot \texttt{RoPE}_{\text{spa}}. \quad (6)$$

This fusion balances global positional alignment with local content sensitivity, allowing the attention module to better capture distortions induced by motion blur and low resolution. As illustrated in Fig. 4(b)(iv), the incorporation of our SF-RoPE module yields sharper edge recovery and improved spatial coherence. Nevertheless, subtle distortions remain in highly cluttered areas (*e.g.*, bicycle), motivating further refinement toward structural consistency.

Recursive Phase Modulation

To address residual motion blur and structural degradation, we propose a VLM-guided recursive phase modulation module (VGPM) applied to the output $\mathbf{x}_{\text{var}}$ of our enhanced VAR backbone. Motivated by observations that blurry inputs often exhibit repeated edge artifacts in the FFT phase domain (see arrows in Fig. 6), we employ phase information as a structurally informative representation that is more robust to occlusion ambiguity and illumination noise than spatial-domain features. As shown in Fig. 6, we first compute the normalized phase map $\hat{\phi} = (\phi + \pi)/2\pi \in [0, 1]$ (ϕ: original phase). A recursive enhancement is then performed as:

$$\begin{aligned} \boldsymbol{M}_0(\hat{\phi}) &= \hat{\phi}, \\ \boldsymbol{M}_T(\hat{\phi}) &= \hat{\phi} + \sum_{t=1}^{T} \boldsymbol{\mathcal{F}}_t(\hat{\phi}) M_{t-1}(\hat{\phi}) \left(\mathbf{1} - \boldsymbol{M}_{t-1}(\hat{\phi})\right), \end{aligned} \quad (7)$$

where T is the total number of modulation steps (set to 8), and $\boldsymbol{\mathcal{F}}_t \in [0, 1]$ is a phase adjustment map predicted by the estimator $\boldsymbol{\Psi}_p$, which shares architecture with $\boldsymbol{\Psi}_I$ in VICM. Based on VLM-based blur assessment b, we further employ a MLP $\boldsymbol{\Theta}_b$ to adaptively guide the modulation strength . The final enhanced phase ϕ^* is obtained by inverting $\boldsymbol{M}_T(\hat{\phi})$ back to the original domain and applying an IFFT to produce the restored image $\mathbf{x}_{\text{out}}$.

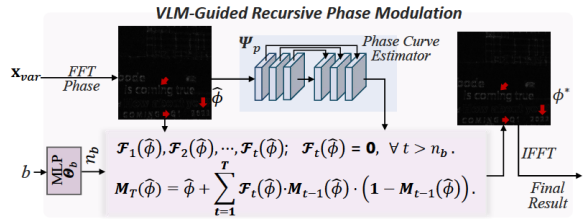


Figure 6: Our VGPM progressively refines the phase representation to mitigate ghosting artifacts introduced by blur.

Optimization

We jointly optimize all parameters ($\boldsymbol{\Theta}_b$, $\boldsymbol{\Psi}_I$, $\boldsymbol{\Theta}_v$, $\boldsymbol{\Psi}_p$, λ) in a reference-free manner using the following losses.

Adaptive Exposure Control Loss We adopt a visibility-aware exposure target, where the base level $E = 0.45$ is dynamically adjusted by $E_d \in [-0.1, 0.1]$ obtained from $\boldsymbol{\Theta}_v$, and Mean aims to calculate the mean intensity:

$$\mathcal{L}_{ex} = |\texttt{Mean}(I_{out}) - (E + E_d)|, \quad (8)$$

Methods	Type	Metrics					Methods	Type	Metrics				
		PSNR↑	NIQE↓	LPIPS↓	FID↓	CLIPIQA↑			PSNR↑	NIQE↓	LPIPS↓	FID↓	CLIPIQA↑
Cascaded Methods													
EnlightenGAN + BD_Noise	L+D	17.25	**4.98**	0.413	47.66	0.203	BD_Noise+ EnlightenGAN	D+L	17.11	**4.96**	0.419	**49.21**	0.201
Zero-DCE++ + BD_Noise	L+D	14.77	5.65	0.533	57.45	0.165	BD_Noise + Zero-DCE++	D+L	14.71	5.67	0.537	56.96	0.164
SCI + BD_Noise	L+D	14.34	5.48	0.541	60.28	0.163	BD_Noise + SCI	D+L	14.56	5.37	0.526	57.44	0.165
EnlightenGAN + Blur2Blur	L+D	**18.16**	5.02	**0.396**	**45.73**	**0.206**	Blur2Blur +EnlightenGAN	D+L	**17.38**	4.98	**0.395**	54.53	**0.226**
Zero-DCE++ + Blur2Blur	L+D	15.79	5.74	0.529	80.88	0.198	Blur2Blur + Zero-DCE++	D+L	14.55	5.85	0.543	60.06	0.183
SCI + Blur2Blur	L+D	16.10	5.75	0.507	69.39	0.169	Blur2Blur+ SCI	D+L	14.55	5.43	0.547	68.05	0.180
Joint LLIE and Deblurring Methods													
LEDNet	w R	24.36	5.37	0.227	25.19	0.207	LIEDNet	w R	26.25	5.40	0.133	13.18	0.292
Retinexformer	w R	24.76	6.07	0.219	22.58	0.214	LIEDNet-L	w R	**26.42**	5.17	**0.127**	**11.38**	**0.305**
JUDE	w R	25.26	5.87	0.186	22.11	0.247	SSFlow*	w/o R	19.24	5.93	0.307	42.05	0.183
DarkIR-M	w R	25.74	5.28	0.165	16.35	0.286	FourierDiff	w/o R	20.22	4.97	0.441	50.59	0.161
DarkIR-L	w R	26.14	**5.15**	0.146	14.27	0.291	**Ours**	w/o R	**23.39**	**4.80**	**0.191**	**26.04**	**0.262**

Table 1: Quantitative comparisons on LOLBlur Dataset. 'L+D' and 'D+L' indicate that the methods belong to **LLIE** $\rightarrow$ **Deblurring** and **Deblurring** $\rightarrow$**LLIE** methods, respectively; 'w R' or 'w/o R' represent that the methods are **Joint LLIE and Deblurring methods with or without GT reference**, respectively. The best results for each type are highlighted in **bold**. [Key: ↑ (↓): Larger (smaller) values leads to better performance, *: using GT mean intensity for illumination adjustment]

Structural Entropy Loss To promote phase-guided structural fidelity, we reconstruct $\mathbf{S}^*_\phi = |\texttt{IFFT}(e^{j\phi^*})|$ and compute Shannon entropy (Lin 2002) over its histogram:

$$\mathcal{L}_{en} = -\sum_{i=1}^{B} p_i \log(p_i), \tag{9}$$

where p_i is the probability of the i-th bin in the normalized histogram of $\mathbf{S}^*_\phi$, B is the total number of bins.

Structural Contrast Loss We improve local structural distinctiveness via negative variance over $N = 16$ patches:

$$\mathcal{L}_{con} = -\frac{1}{N}\sum_{k=1}^{N} \sigma^2(\mathbf{S}^*_{\phi,k}). \tag{10}$$

Total Variation Loss To suppress artifacts, we apply a total variation (TV) loss (Osher et al. 2005) on $\mathbf{x}_{out}$:

$$\mathcal{L}_{tv} = \sum_{x,y} |I_{x+1,y} - I_{x,y}| + |I_{x,y+1} - I_{x,y}|. \tag{11}$$

The overall optimization objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{ex} + \lambda_{en}\mathcal{L}_{en} + \lambda_{con}\mathcal{L}_{con} + \lambda_{tv}\mathcal{L}_{tv}, \tag{12}$$

where λ_{en}, λ_{con}, and λ_{tv} are weights of the losses.

Experiments

Experiment Settings

Training Details Our model is trained on 512×512 resolution images using the AdamW optimizer for 200 epochs. The initial learning rate is set to 10^{-4} and decayed by a factor of 0.5 at epochs 100, 150, 180, and 190. All experiments are conducted on a single NVIDIA 5090 GPU.

Datasets and Metrics We evaluate our method on the LOLBlur and Real-LOLBlur (Zhou, Li, and Change Loy 2022) datasets. LOLBlur comprises 12,000 image pairs with diverse illumination and motion blur. We utilize 10,200 low-blur, noisy images from the official training split for optimization, and the remaining 1,800 pairs for both quantitative and qualitative evaluation using reference-based (PSNR, LPIPS (Zhang et al. 2018), and FID (Heusel et al. 2017)) and no-reference metrics. Following LEDNet (Zhou, Li, and Change Loy 2022), we further assess generalization on Real-LOLBlur dataset, which contains 1,354 unpaired real-world low-light blurry images. Evaluation employs NIQE (Mittal, Soundararajan, and Bovik 2012), CLIP-IQA (Wang, Chan, and Loy 2023), MUSIQ (Ke et al. 2021), and MANIQA (Yang et al. 2022).

Performance on LOLBlur Dataset

We compare our method against four baseline categories: **LLIE** $\rightarrow$ **Deblurring**, **Deblurring** $\rightarrow$ **LLIE**, supervised joint frameworks, and unsupervised joint frameworks.

LLIE $\rightarrow$ **Deblurring** LLIE models (SCI (Ma et al. 2022), EnlightenGAN (Jiang et al. 2021), and Zero-DCE++ (Li, Guo, and Loy 2022)) are first trained. Their enhanced outputs are subsequently used to train BD_Noise (Lee et al. 2024) and Blur2Blur (Pham et al. 2024) for deblurring.

Deblurring $\rightarrow$ **LLIE** We begin by training deblurring methods. The outputs are then optimized by LLIE methods.

End-to-End Methods without Reference Our model, together with SSFlow (Li et al. 2024) and FourierDiff (Lv et al. 2024), is optimized without using any GT reference images.

Supervised End-to-End Baselines We also incorporate representative supervised models for benchmarking, including LEDNet (Zhou, Li, and Change Loy 2022), LIEDNet (Liu et al. 2025), RetinexFormer (Cai et al. 2023), JUDE (Vo and Park 2025), and DarkIR (Feijoo et al. 2025).

Quantitative and Qualitative Comparisons As summarized in Tab. 1, our proposed method significantly outperforms all cascaded pipelines, achieving over 5 dB PSNR gains and superior perceptual quality. Compared to existing unsupervised joint frameworks (SSFlow and FourierDiff), our method exhibits notable improvements across both pixel-level and perceptual metrics. Furthermore, despite optimized without ground-truth supervision, our model achieves performance competitive with fully supervised

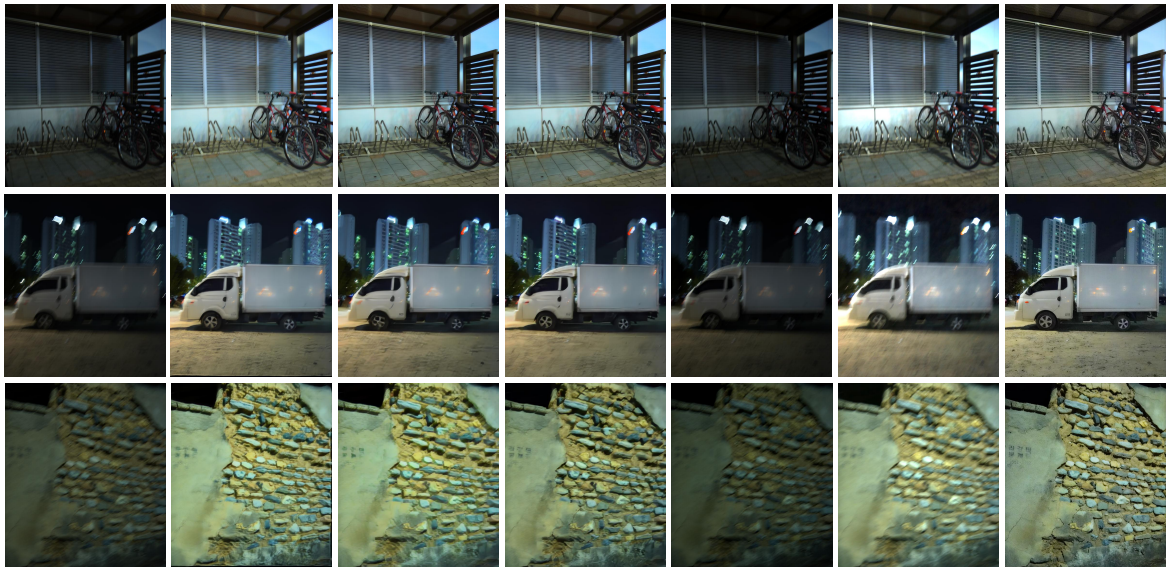


Figure 7: Visual comparisons on the Real-LOLBlur dataset. Our method restores natural illumination and achieves superior deblurring performance with sharper edges and clearer structures, showing strong generalization to complex real-world scenes.

baselines (*e.g.*, LEDNet and JUDE), particularly excelling in perceptual quality as indicated by the lowest NIQE. As illustrated in Fig. 5, our method delivers perceptually more faithful reconstructions compared to unsupervised baselines. In particular, it preserves fine-grained texture details and recovers natural illumination more effectively than SSFlow and FourierDiff. While supervised models such as Retinexformer yield enhanced brightness, they often fail to remove motion blur. In contrast, our approach achieves a more balanced restoration with fewer visual artifacts.

Performance on Unpaired Real-World Data

We conduct cross-dataset evaluations on unpaired real-world data. Specifically, we directly apply the model trained on the LOLBlur dataset to unseen samples from Real-LOLBlur dataset. Notably, FourierDiff involves an internal optimization process during inference. Quantitative comparisons and visual results are summarized in Tab. 2 and Fig. 7.

Comparison Results As shown in Tab. 2 and Fig. 7, our method demonstrates good generalization ability on the Real-LOLBlur dataset, despite being trained solely on LOLBlur without access to paired data. Quantitatively, it achieves the best NIQE, CLIPIQA, and MUSIQ scores among all unsupervised joint methods and even approaches or outperforms several supervised counterparts. Qualitatively, our outputs retain better structural integrity and perceptual fidelity, avoiding over-smoothing or illumination inconsistencies commonly observed in competing baselines.

Methods	**Type**	Metrics			
		NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
Cascaded Methods					
EnlightenGAN + BD_Noise	L+D	5.44	0.157	**40.05**	**0.171**
EnlightenGAN + Blur2Blur	L+D	5.49	0.160	39.24	0.169
Zero-DCE++ + BD_Noise	L+D	5.66	0.161	25.88	0.103
Zero-DCE++ + Blur2Blur	L+D	5.51	0.167	27.20	0.109
SCI + BD_Noise	L+D	5.22	0.181	34.28	0.126
SCI + Blur2Blur	L+D	**5.13**	**0.185**	33.87	0.129
BD_Noise+ EnlightenGAN	D+L	5.58	0.150	**40.36**	**0.166**
BD_Noise + Zero-DCE++	D+L	5.77	0.210	22.41	0.130
BD_Noise + SCI	D+L	5.64	0.180	26.37	0.109
Blur2Blur +EnlightenGAN	D+L	**5.43**	0.152	38.96	0.165
Blur2Blur + Zero-DCE++	D+L	5.92	**0.215**	22.47	0.127
Blur2Blur+ SCI	D+L	5.52	0.182	24.64	0.102
Joint LLIE and Deblurring Methods					
LEDNet	w R	5.07	0.256	49.46	**0.228**
Retinexformer	w R	5.69	0.208	40.47	0.173
JUDE	w R	4.92	0.236	**50.29**	0.223
DarkIR-M	w R	4.97	0.252	48.31	0.209
DarkIR-L	w R	**4.90**	**0.262**	48.72	0.216
SSFlow	w/o R	5.94	0.190	30.93	0.148
FourierDiff	w/o R	5.59	0.187	32.01	0.122
Ours	w/o R	**5.16**	**0.226**	**47.53**	**0.223**

Table 2: Quantitative comparisons on Real-LOLBlur Dataset. Our method achieves superior perceptual quality compared to unsupervised baselines.

Conclusion

In this work, we introduce **VAR-LIDE**, a fully unsupervised generative framework for joint LLIE and deblurring. By leveraging the autoregressive modeling capacity of VAR backbone and the perceptual guidance from VLMs, we design a VLM-informed conditioning mechanism that achieve adaptive illumination enhancement. To further strengthen structural fidelity under blur, we enhance the positional modeling of the VAR backbone via input-adaptive spatial-frequency RoPE. Additionally, a recursive phase-domain modulation module is developed to suppress blur-induced edge artifacts, with guidance from blur-aware VLM assessments. Extensive experiments confirm that VAR-LIDE achieves impressive quantitative and perceptual performance on several benchmarks.

References

Anger, J.; Facciolo, G.; and Delbracio, M. 2018. Modeling Realistic Degradations in Non-Blind Deconvolution. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 978–982.

Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; and Zhang, Y. 2023. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12504–12513.

Chen, Z.; Ma, X.; Fang, G.; and Wang, X. 2025. Collaborative decoding makes visual auto-regressive modeling efficient. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23334–23344.

Dong, J.; Roth, S.; and Schiele, B. 2022. DWDN: Deep Wiener Deconvolution Network for Non-Blind Image Deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9960–9976.

Dong, W.; Min, Y.; Zhou, H.; and Chen, J. 2025a. Towards Scale-Aware Low-Light Enhancement via Structure-Guided Transformer Design. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Dong, W.; Zhou, H.; Mousavi, S. A.; and Chen, J. 2025b. Retinex-guided histogram transformer for mask-free shadow removal. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Dong, W.; Zhou, H.; Tian, Y.; Sun, J.; Liu, X.; Zhai, G.; and Chen, J. 2024a. ShadowRefiner: Towards mask-free shadow removal via fast fourier transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Dong, W.; Zhou, H.; Wang, R.; Liu, X.; Zhai, G.; and Chen, J. 2024b. Dehazedct: Towards effective non-homogeneous dehazing via deformable convolutional transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Dong, W.; Zhou, H.; Zhang, Y.; Liu, X.; and Chen, J. 2024c. Ecmamba: Consolidating selective state space model with retinex guidance for efficient multiple exposure correction. *Advances in Neural Information Processing Systems*.

Feijoo, D.; Benito, J. C.; Garcia, A.; and Conde, M. V. 2025. Darkir: Robust low-light image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10879–10889.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1780–1789.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2024. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, 222–241. Springer.

Han, J.; Liu, J.; Jiang, Y.; Yan, B.; Zhang, Y.; Yuan, Z.; Peng, B.; and Liu, X. 2025. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15733–15744.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. EnlightenGAN: Deep Light Enhancement Without Paired Supervision. *IEEE Transactions on Image Processing*, 30: 2340–2349.

Jobson, D.; Rahman, Z.; and Woodell, G. 1997. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7): 965–976.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Kundur, D.; and Hatzinakos, D. 1996. Blind image deconvolution. *IEEE Signal Processing Magazine*, 13(3): 43–64.

Lee, C.; Kim, J.; Lee, S.; Jung, J.; Cho, Y.; Kim, T.; Jo, T.; Lee, M.; and Jang, M. 2024. Blind Image Deblurring with Noise-Robust Kernel Estimation. In *European Conference on Computer Vision*, 188–204. Springer.

Li, C.; Guo, C.; and Loy, C. C. 2022. Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4225–4238.

Li, L.; Zhu, C.; Chen, J.; Shi, B.; and Lian, Q. 2024. Self-supervised normalizing flow for jointing low-light enhancement and deblurring. *Circuits, Systems, and Signal Processing*, 43(9): 5727–5748.

Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Van Gool, L. 2023. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18278–18289.

Lin, J. 2002. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*.

Liu, M.; Cui, Y.; Ren, W.; Zhou, J.; and Knoll, A. C. 2025. LIEDNet: A Lightweight Network for Low-Light Enhancement and Deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7): 6602–6615.

Lv, X.; Zhang, S.; Wang, C.; Zheng, Y.; Zhong, B.; Li, C.; and Nie, L. 2024. Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25378–25388.

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5637–5646.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*.

Osher, S.; Burger, M.; Goldfarb, D.; Xu, J.; and Yin, W. 2005. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*.

Pham, B.-D.; Tran, P.; Tran, A.; Pham, C.; Nguyen, R.; and Hoai, M. 2024. Blur2blur: Blur conversion for unsupervised image deblurring on unknown domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2804–2813.

Pizer, S. M.; Amburn, E. P.; Austin, J. D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J. B.; and Zuiderveld, K. 1987. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3): 355–368.

Qu, Y.; Yuan, K.; Hao, J.; Zhao, K.; Xie, Q.; Sun, M.; and Zhou, C. 2025. Visual Autoregressive Modeling for Image Super-Resolution. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Rajagopalan, S.; Narayan, K.; and Patel, V. M. 2025. RestoreVAR: Visual Autoregressive Generation for All-in-One Image Restoration. arXiv:2505.18047.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Vo, T.; and Park, C. Y. 2025. Deep Joint Unrolling for Deblurring and Low-Light Image Enhancement (JUDE). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2696–2705.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*.

Wang, S.; Zheng, J.; Hu, H.-M.; and Li, B. 2013. Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images. *IEEE Transactions on Image Processing*, 22(9): 3538–3548.

Wang, S.; Zheng, N.; Huang, J.; and Zhao, F. 2025. Navigating Image Restoration with VAR's Distribution Alignment Prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7559–7569.

Xu, D.; Dong, W.; and Zhou, H. 2022. Sclera recognition based on efficient sclera segmentation and significant vessel matching. *The Computer Journal*.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Zhou, H.; Dong, W.; and Chen, J. 2025. LITA-GS: Illumination-Agnostic Novel View Synthesis via Reference-Free 3D Gaussian Splatting and Physical Priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Zhou, H.; Dong, W.; Liu, X.; Liu, S.; Min, X.; Zhai, G.; and Chen, J. 2024. Glare: Low light image enhancement via generative latent feature based codebook retrieval. In *European Conference on Computer Vision*.

Zhou, H.; Dong, W.; Liu, X.; Zhang, Y.; Zhai, G.; and Chen, J. 2025. Low-light image enhancement via generative perceptual priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhou, H.; Dong, W.; Liu, Y.; and Chen, J. 2023. Breaking through the haze: An advanced non-homogeneous dehazing method based on fast fourier convolution and convnext. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhou, S.; Li, C.; and Change Loy, C. 2022. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, 573–589. Springer.