MAPO-NET: MLLM-GUIDED ARTIFACT PERCEPTION AND FEEDBACK OPTIMIZATION NETWORK FOR HDR RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dynamic-range (HDR) reconstruction in dynamic scenes remains challenging due to the dual demands of exposure recovery and motion artifact removal. Existing methods mostly rely on implicit feature alignment without explicit semantic understanding, struggling to handle complex local degradations. To address this, we propose MAPO-Net, a novel framework that bridges high-level semantic reasoning with low-level reconstruction by reformulating artifact removal as a text-guided detection and feedback optimization process. MAPO-Net operates in two stages: 1) Cross-Modal Artifact Perception, which leverages a frozen MLLM (Qwen3-VL) and a lightweight detector (MPA-Detector) to generate semantic descriptions and localize artifact regions; and 2) Perception-Driven Feedback Optimization, which converts localized masks into adaptive weight maps to guide an artifact-aware loss, steering the HDR network toward repairing defective regions. Furthermore, we construct a new dynamic HDR dataset comprising 2,300 challenging real-captured LDR-HDR sample groups. Extensive experiments demonstrate that MAPO-Net effectively suppresses complex artifacts and consistently outperforms state-of-the-art approaches.

1 INTRODUCTION

High dynamic range (HDR) imaging is prone to ghosting artifacts, including duplicated structures, missing object regions, semi-transparent phantoms, and blurred edges. Jointly suppressing ghosting and motion blur remains a key challenge in dynamic HDR fusion.

HDR reconstruction has evolved from physics-based methods to CNN-based alignment and fusion O'shea & Nash (2015), Transformer-based context modeling Han et al. (2021), and diffusion and generation-based de-ghosting Ho et al. (2020); Xia et al. (2023). Mild motion calls for local alignment, larger displacements require broader contextual reasoning, and severe ghosting may require structure completion. These regional differences motivate us to introduce an artifact-aware feedback mechanism that evaluates reconstruction outputs and guides training toward challenging regions. To this end, we propose MAPO-Net, a network framework designed to address this challenge.

Existing dynamic HDR benchmarks often suffer from limited scene and motion diversity, synthetic LDR degradations, or imperfect HDR references. To address these limitations, we introduce *Hard-HDR*, a large-scale real-world dataset featuring controllable motion, diverse illumination, and high-quality HDR references.

Our contributions are summarized as follows:

1. We propose MAPO-Net, a feedback optimization framework that incorporates MLLM-based semantic reasoning into dynamic HDR reconstruction to suppress ghosting artifacts.
2. We design a perception-driven feedback mechanism that converts MLLM-generated textual diagnoses into pixel-wise artifact weights, guiding the reconstruction network to localize and restore challenging regions during training.
3. We construct and publicly release *Hard-HDR*, a dynamic HDR dataset comprising 2,300 challenging cases, each with three naturally captured LDR images and a high-quality HDR

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.

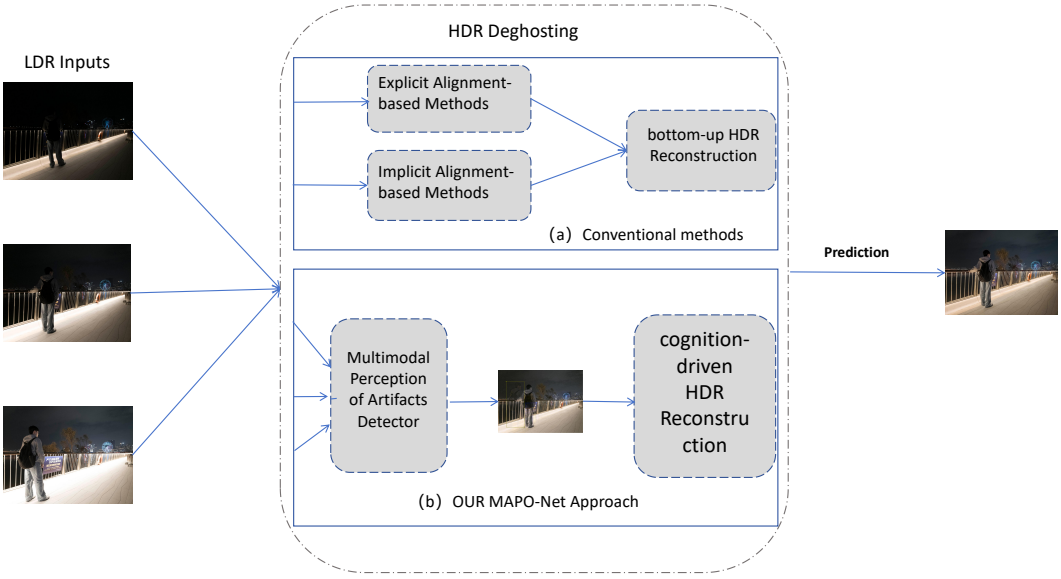

Figure 1: Paradigm comparison between existing multi-exposure HDR methods and MAPO-Net. Existing methods employ a fixed training and inference pipeline, rendering them incapable of handling diverse and complex scenes. Conversely, empowered by a Multimodal Large Language Model (MLLM), MAPO-Net introduces a novel anomaly diagnostic detector specifically for ghosting artifacts, enabling flexible and adaptive deghosting correction.

reference. Its diverse scenes, complex illumination, and severe motion provide a challenging benchmark for dynamic HDR reconstruction.

2 Related Work

2.1 Multi-Exposure HDR Reconstruction

Deep multi-exposure HDR methods have shifted from hand-crafted pipelines to implicit feature alignment. Early CNN-based methods Yan et al. (2019); Wang et al. (2018) are limited in handling complex motion, while Transformers Liu et al. (2022) improve alignment through global context modeling and non-local fusion. Recent diffusion models further exploit generative priors to restore fine details. Nevertheless, these methods largely rely on low-level feature matching without explicit semantic reasoning about local degradations, potentially confusing background bokeh with motion artifacts and leaving overexposure distortions or residual ghosting Perez-Pellitero et al. (2023); Wang & Yoon (2022). MAPO-Net addresses this limitation by introducing semantic diagnosis to guide dynamic HDR reconstruction.

2.2 Applications of Large Multimodal Models in Low-Level Vision

Multimodal large language models (MLLMs) Liu et al. (2023a); OpenAI (2023) have recently been applied to image restoration and enhancement. Existing approaches primarily use them to generate global guidance Yuan et al. (2024); Conde et al. (2024) or assess image quality Wu et al. (2024); You et al. (2023), with limited integration into reconstruction feedback and pixel-level artifact localization. In contrast, MAPO-Net embeds an MLLM as an artifact-diagnosis expert in a closed-loop HDR reconstruction framework. Its semantic diagnosis–pixel localization–penalization process translates degradation perception into spatially precise optimization guidance. To our knowledge, MAPO-Net is the first to integrate MLLM-based artifact diagnosis into an HDR feedback optimization loop.

3 HARD-HDR DATASET

3.1 DATASET COLLECTION PIPELINE

To capture the LDR images, we utilize an OPPO smartphone camera mounted on a tripod and adopt a three-exposure bracketing strategy with exposure intervals of ± 1, ± 2, and ± 3 EV (Exposure Value). The native resolution of the captured images is 2040×1528.

To simulate dynamic LDR scenarios with different difficulties, we design three controllable motion patterns:

- **Local motion:** the background remains static while foreground objects or people move;
- **Camera ego-motion:** the scene remains static while the camera translates, causing global misalignment;
- **Compound motion:** foreground motion and camera displacement occur simultaneously, forming the most challenging case.

3.2 HDR GROUND-TRUTH GENERATION

Since most dynamic HDR reconstruction methods take three LDR images as input, we generate candidate HDR images, denoted as HDR, from three-exposure LDR sequences. The fusion is performed using a triangular weighting function:

$$\hat{H}(p) = \frac{\sum_{n=1}^{3} w_n(p) H_n(p)}{\sum_{n=1}^{3} w_n(p)}, \tag{1}$$

where $\hat{H}(p)$ is the fused HDR value at pixel p, $w_n(p)$ is the triangular weight of the n-th exposure, and $H_n(p)$ denotes the corresponding linearized LDR image.

A total of 8,000 HDR samples were generated.

3.3 MANUAL SCREENING AND DATASET SPLIT

We conduct manual closed-loop quality control after raw sample generation. Eight annotators visually inspect all samples and remove images with blur, exposure failure, or residual artifacts. After screening, *Hard-HDR* contains 2,300 independent scenes, including 2,000 training scenes and 300 testing scenes. Furthermore, to provide ground-truth supervision for training the MPA-Detector, we simultaneously annotated corresponding artifact masks for all 2,000 training and 300 testing scenes. The training and testing sets are scene-disjoint to avoid data leakage.

3.4 COMPARISON WITH EXISTING BENCHMARKS

Figure 2 illustrates the comparison between *Hard-HDR* and the public datasets Kalantari17 Kalantari & Ramamoorthi (2017) and Tel Niu et al. (2021).

4 METHODOLOGY

4.1 OVERVIEW

Given a sequence of LDR images captured at different exposure levels in a dynamic scene $X = \{X_1, X_2, \ldots, X_K\}$ (typically $K = 3$), the goal of HDR reconstruction is to synthesize an artifact-free HDR image with complete dynamic range I_{HDR}.Existing methods usually directly learn the mapping from X to I_{HDR}.

4.2 MLLM-DRIVEN ARTIFACT PERCEPTION

This module is designed to leverage the powerful zero-shot reasoning capability of MLLMs Bai et al. (2025) to identify local degradations in images from a high-level semantic perspective. As

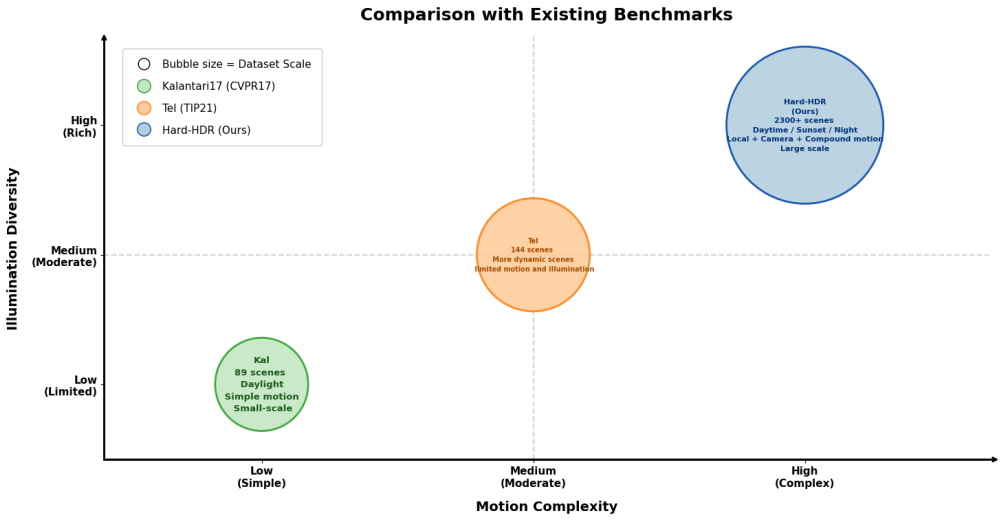

Figure 2: Quadrant-chart comparison between *Hard-HDR* and representative public datasets. Kalantari17 Kalantari & Ramamoorthi (2017) is limited to daytime scenes at a small scale (74 training / 15 testing), while Tel Niu et al. (2021) offers about 144 dynamic scenes with limited scene categories and motion patterns. *Hard-HDR* is built entirely from real-captured RAW images with minimal-artifact HDR ground truth, covering daytime, sunset, and nighttime scenes with a balanced mix of local, camera ego, and compound motion. Bubble size denotes dataset scale.

shown in Figure 3(b), this process consists of two sequential stages: visual feature extraction and textual semantic reasoning.

Visual Feature Extraction: We adopt the pre-trained and frozen CLIP ViT-L/14 Radford et al. (2021) as the visual encoder $\mathcal{E}_{vis}$. Given the initial reconstructed image $\hat{I}_{HDR}$, we extract its spatial visual features F_v:

$$F_v = \mathcal{E}_{vis}(\hat{I}_{HDR}) \in \mathbb{R}^{H_v \times W_v \times C}, \tag{2}$$

where H_v and W_v denote the spatial dimensions of the feature map, and C represents the channel capacity.

Textual Semantic Reasoning: We employ Qwen3-VL Bai et al. (2025) to function as a "diagnosis expert". To guide the MLLM to focus on specific issues in HDR reconstruction, we design a dedicated text prompt p:

> *"Locate and describe the following artifacts in this HDR image: ghosting, misalignment, saturation, exposure inconsistency, and other unnatural regions."*

The visual features F_v are first mapped into the language model's embedding space via a linear projection $\mathcal{P}_{v \to t}$. The MLLM then models the joint distribution of the artifact description sequence $T_a = \{t_1, t_2, \ldots, t_N\}$. The generation process is formulated as an autoregressive conditional probability maximization:

$$T_a = \arg\max_{T} \prod_{i=1}^{N} P_\theta \left(t_i \mid t_{<i}, \mathcal{P}_{v \to t}(F_v) \oplus \mathcal{E}_{txt}(p) \right), \tag{3}$$

where P_θ denotes the frozen probability distribution of the MLLM, $\oplus$ represents sequence concatenation, and N is the sequence length. This yields $T_a \in \mathbb{R}^{N \times C}$, a high-level semantic diagnosis explicitly identifying artifacts.

4.3 Cross-modal Artifact Detector (MPA-Detector)

While the MLLM excels at generating semantic descriptions, it lacks pixel-level localization. To bridge this gap, we design a lightweight cross-modal detector (MPA-Detector). As illustrated in Figure 3(c), it translates the high-level textual tokens T_a into a precise 2D artifact mask $\hat{M}$.

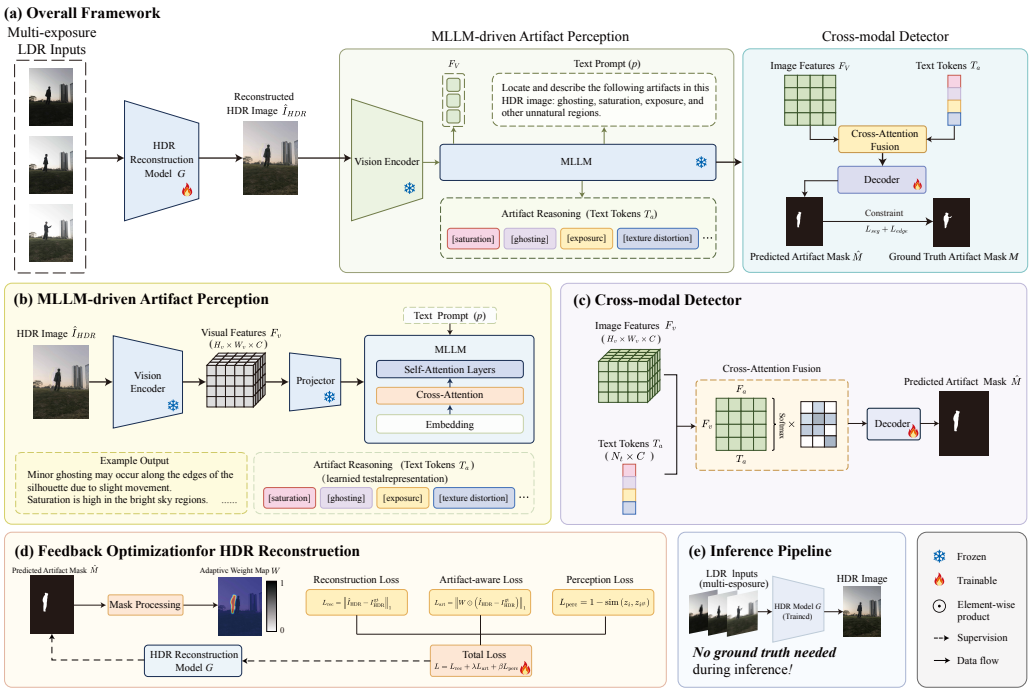

Figure 3: Overview of the proposed MAPO-Net framework. (a) Given multi-exposure LDR inputs, the HDR reconstruction network first produces an initial HDR prediction. A frozen MLLM performs semantic artifact perception and generates artifact-aware tokens. The cross-modal detector further aligns semantic and visual representations to produce pixel-level artifact masks. (b) Detailed illustration of the MLLM-driven artifact perception module. (c) Detailed illustration of the cross-modal detector. (d)Feedback optimization module for HDR reconstruction, where predicted artifact masks are transformed into adaptive weighting maps to construct artifact-aware losses and guide restoration of degraded regions. (e) Inference pipeline.

Cross-Attention Fusion: We leverage a cross-attention mechanism to inject textual priors into the visual features. The visual features F_v are linearly projected to form the Query (Q), while the textual tokens T_a serve as the Key (K) and Value (V):

$$Q = F_v W_Q, \quad K = T_a W_K, \quad V = T_a W_V. \tag{4}$$

The text-injected visual representation F_a is computed via scaled dot-product attention:

$$F_a = F_v + \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{5}$$

where $\{W_Q, W_K, W_V\}$ are learnable projection matrices, and d_k is the scaling factor corresponding to the channel dimension.

Mask Decoder: To achieve precise localization, a lightweight U-Net architecture $\mathcal{D}_{mask}$ is employed. By utilizing skip connections, it fuses high-level semantic features with low-level spatial details:

$$\hat{M} = \mathcal{D}_{mask}(F_a), \quad \text{s.t. } \hat{M}(\mathbf{u}) \in [0,1] \ \ \forall \mathbf{u} \in \Omega, \tag{6}$$

where $\Omega \subset \mathbb{R}^2$ denotes the continuous spatial domain of the image, and $\mathbf{u}$ represents the spatial coordinates.

To optimize the detector, we utilize the ground truth mask M. The detection loss $\mathcal{L}_{\text{det}}$ combines a Binary Cross-Entropy (BCE) loss and a boundary-aware Dice loss, utilizing the spatial gradient operator ∇ to explicitly penalize structural discrepancies at artifact boundaries:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{BCE}}(\hat{M}, M) + \mathcal{L}_{\text{Dice}}\left(\|\nabla \hat{M}\|_2, \|\nabla M\|_2\right), \tag{7}$$

where $\|\nabla(\cdot)\|_2$ computes the gradient magnitude, effectively extracting the structural edges.

4.4 Perception-driven Feedback Optimization

With the predicted artifact mask $\hat{M}$, MAPO-Net advances to the feedback optimization stage (Figure 3(d)). Crucially, we shift the optimization paradigm from conventional "blind global reconstruction" to a "targeted local penalty" strategy.

Adaptive Weight Map Generation: Since $\hat{M}$ may contain minor noise, we generate a smooth adaptive spatial weight map W through mathematical morphology and spatial filtering. Specifically, we apply a morphological dilation using a structuring element $\mathcal{B}$, followed by a convolution with a Gaussian smoothing kernel $\mathcal{K}_\sigma$:

$$W(\mathbf{u}) = \left(\sup_{\mathbf{v} \in \mathcal{B}} \hat{M}(\mathbf{u} - \mathbf{v}) \right) * \mathcal{K}_\sigma(\mathbf{u}), \tag{8}$$

where $*$ denotes the spatial convolution operation. In the resulting weight map $W \in [0,1]^{H \times W}$, values approaching 1 indicate regions with a higher probability of severe artifacts.

Customized Loss Function: To compel the base network (AFUNet) to prioritize artifact-heavy regions, we design a joint optimization objective for its parameters θ_G. The first term is a global reconstruction loss to preserve overall fidelity, formulated as the ℓ_1-norm over the spatial domain Ω:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\Omega|} \int_\Omega \left\| \hat{I}_{HDR}(\mathbf{u}) - I^{GT}_{HDR}(\mathbf{u}) \right\|_1 d\mathbf{u}. \tag{9}$$

The core of our objective is the artifact-aware loss. We spatially modulate the error using the weight map W, imposing amplified penalties on the degraded regions highlighted by the MLLM:

$$\mathcal{L}_{\text{art}} = \frac{1}{|\Omega|} \int_\Omega W(\mathbf{u}) \cdot \left\| \hat{I}_{HDR}(\mathbf{u}) - I^{GT}_{HDR}(\mathbf{u}) \right\|_1 d\mathbf{u}. \tag{10}$$

Furthermore, to enhance semantic consistency, we incorporate a perceptual loss within the MLLM's feature space. Let z_p and z_{gt} be the [EOS] token embeddings extracted from Qwen3-VL for $\hat{I}_{HDR}$ and I^{GT}_{HDR}, respectively. The loss is defined via cosine distance:

$$\mathcal{L}_{\text{perc}} = 1 - \frac{\langle z_p, z_{gt} \rangle}{\|z_p\|_2 \|z_{gt}\|_2}, \tag{11}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

Finally, AFUNet is trained end-to-end by optimizing the composite loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{art}} + \beta \mathcal{L}_{\text{perc}}, \tag{12}$$

where λ and β are hyperparameters balancing the artifact-aware and perceptual terms.

4.5 Inference Stage

MAPO-Net decouples the MLLM and MPA-Detector as training-only supervisors, allowing the lightweight AFUNet Yan et al. (2024) to perform HDR reconstruction independently at inference without additional latency or computational overhead.

5 Experiments

5.1 Experimental Settings

5.1.1 Evaluation Metrics

Following standard HDR evaluation protocols, we report PSNR and SSIM in both μ-law tone-mapped and linear domains, denoted as PSNR-μ, SSIM-μ, PSNR-L, and SSIM-L. We also use HDR-VDP-2 Mantiuk et al. (2011) to assess HDR perceptual quality.

Table 1: Quantitative comparison with state-of-the-art methods. All models are trained on our proposed Hard-HDR dataset and evaluated on the Kalantari17 Kalantari & Ramamoorthi (2017), Tel Niu et al. (2021), and Hard-HDR datasets. AFUNet Yan et al. (2024) serves as our baseline. The best results are highlighted in **bold**, and the second-best are underlined.

Methods	Venue	Kalantari17 Dataset					TEL Dataset					Our Hard-HDR Dataset				
		PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
DHDR Kalantari & Ramamoorthi (2017)	ECCV'18	41.92	0.9871	41.65	0.9892	70.85	40.25	0.9803	43.42	0.9931	68.12	40.90	0.9850	43.54	0.9887	68.77
NHDRR Yan et al. (2020)	TIP'20	41.88	0.9865	41.49	0.9883	68.55	37.78	0.9599	40.11	0.9832	68.21	39.90	0.9859	42.53	0.9882	68.10
HDRGAN Niu et al. (2021)	TIP'20	43.92	0.9908	41.51	0.9891	71.10	41.28	0.9842	45.44	0.9937	70.19	41.31	0.9888	43.66	0.9888	69.88
DiffHDR Liu et al. (2023c)	TCSVT'23	44.06	0.9912	41.68	0.9894	71.16	42.01	0.9865	45.63	0.9941	71.43	42.88	0.9898	44.82	0.9890	70.85
HyHDR Yan et al. (2023)	CVPR'23	44.47	0.9923	41.89	0.9898	71.84	42.32	0.9871	45.72	0.9944	71.68	43.73	0.9988	45.43	0.9921	71.66
SMHDR Ni et al. (2022)	ACM MM'22	44.62	0.9932	42.11	0.9902	71.92	43.19	0.9881	46.92	0.9951	72.38	44.29	0.9989	45.48	0.9922	71.87
HDR-Transformer Liu et al. (2022)	ECCV'22	44.82	0.9944	42.23	0.9881	70.58	42.44	0.9858	46.39	0.9944	71.22	44.64	0.9991	46.88	0.9948	72.37
AFUNet (Baseline) Yan et al. (2024)	TPAMI'24	44.97	0.9935	42.82	0.9910	71.98	43.65	0.9889	47.94	0.9983	73.47	44.70	0.9994	46.42	0.9952	72.78
MAPO-Net	**Ours**	**45.20** (+0.23)	**0.9958** (+0.0023)	**43.29** (+0.47)	**0.9920** (+0.0010)	**72.80** (+0.82)	**44.76** (+1.11)	**0.9896** (+0.0007)	**48.15** (+0.21)	**0.9991** (+0.0008)	**74.39** (+0.92)	**44.89** (+0.19)	**0.9995** (+0.0001)	**46.82** (+0.40)	**0.9971** (+0.0019)	**73.99** (+1.21)



Figure 4: Qualitative comparisons on Kalantari17 Kalantari & Ramamoorthi (2017) and Tel Niu et al. (2021). All methods are trained on our Hard-HDR dataset. Our framework outperforms state-of-the-art methods in suppressing motion artifacts and reconstructing dynamic scenes.

5.1.2 Implementation Details

Network Architecture. MAPO-Net consists of an AFUNet Yan et al. (2024)-based reconstruction backbone G and a cross-modal artifact perception module using frozen CLIP ViT-L/14 Radford et al. (2021) and Qwen3-VL Bai et al. (2025) for visual encoding and artifact diagnosis, respectively. We train for 200 epochs on two NVIDIA RTX 5090 GPUs with 128×128 patches, AdamW, an initial learning rate of 2×10^{-4}, cosine annealing, and loss weights $\lambda = 0.5$ and $\beta = 0.1$. The MLLM and MPA-Detector are removed at inference, preserving AFUNet's parameter count and inference speed.

5.2 Comparison with State-of-the-Art

5.2.1 Quantitative Comparison

As shown in Table 1, MAPO-Net consistently outperforms existing SOTA methods across nearly all metrics and datasets, delivering substantial gains in PSNR and SSIM. Notably,it improves HDR-VDP-2 by +0.82, +0.92, and +1.21 on the two public datasets and our proposed dataset, respectively, establishing a new SOTA benchmark.

5.2.2 Qualitative Comparison

As shown in Figures 4 and 5, visual comparisons on the Hard-HDR, Kal's, and Tel's datasets show that existing SOTA methods still suffer from noticeable motion-induced ghosting. In contrast, MAPO-Net consistently suppresses these artifacts across different datasets while preserving clear edge structures and sharp object contours, thereby producing HDR images with superior visual quality.

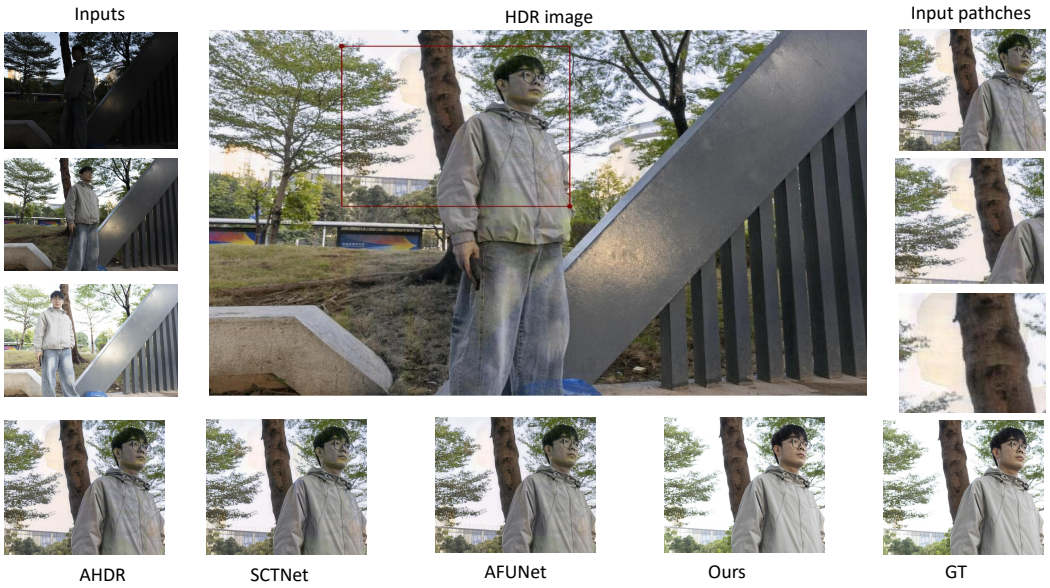

Figure 5: Qualitative comparison on the Hard-HDR dataset, with all methods trained on the same data. Our framework outperforms existing SOTA methods in suppressing motion artifacts and reconstructing dynamic scenes.

Table 2: Cross-Architecture Universality. Comparison between the original AHDR Yan et al. (2019)/SCTNet Liu et al. (2023b) architectures and those integrated with MAPO-Net. As observed, the inclusion of our framework substantially enhances the overall performance.

Base Architecture	Framework	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
AHDR Yan et al. (2019) (CNN-based)	Original	42.78	0.9928	42.72	0.9932	71.29
	+ MAPO (Ours)	**44.15** (+1.37)	**0.9942** (+0.0014)	**43.18** (+0.46)	**0.9958** (+0.0026)	**72.76** (+1.47)
SCTNet Liu et al. (2023b) (Transformer)	Original	44.85	0.9927	42.38	0.9921	71.92
	+ MAPO (Ours)	**46.30** (+1.45)	**0.9931** (+0.0004)	**43.21** (+0.83)	**0.9964** (+0.0043)	**72.83** (+0.91)

5.3 Cross-Architecture Universality

As shown in Table 2, MAPO-Net consistently improves PSNR and SSIM for both baseline architectures, while boosting HDR-VDP-2 by 1.47 for AHDR Yan et al. (2019) and 0.91 for SCTNet, demonstrating its effective plug-and-play capability.

5.4 MPA-Detector Mask Analysis

As shown in Table 3, compared with the Kal's and Tel's datasets, our dataset achieves the most significant improvements in Precision and IoU: Precision gains of 0.1311 over Kal's and 0.1251 over Tel's, and IoU gains of 0.1395 over Kal's and 0.1309 over Tel's, respectively. The improvements in F1 and Recall are relatively moderate, with gains of 0.1109 and 0.1064 over Kal's, and 0.0537 and 0.0835 over Tel's, respectively. These results demonstrate the effectiveness of the proposed cross-modal detector.

Table 3: Comparison of masks predicted by MPA-Detector across different datasets: Kal.'s Kalantari & Ramamoorthi (2017) dataset, Tel.'s Niu et al. (2021) dataset, and our custom dataset. MPA-Detector maintains strong performance across all three datasets and achieves its best performance on our dataset.

Test Dataset	F1 Score	Precision	Recall	IoU
Kal.'s Kalantari & Ramamoorthi (2017)	0.7192	0.7018	0.7736	0.6238
Tel.'s Niu et al. (2021)	0.7237	0.7078	0.7438	0.6324
Our Dataset	**0.8301**	**0.8329**	**0.8273**	**0.7633**
	(+0.1109)	(+0.1311)	(+0.0537)	(+0.1395)
	(+0.1064)	(+0.1251)	(+0.0835)	(+0.1309)

Table 4: Ablation study of our core components on the Hard-HDR dataset. (a) is the AFUNet Yan et al. (2024) baseline. (b) evaluates the perception loss ($\mathcal{L}_{perc}$). (c) is a vision-only variant that predicts the artifact mask using only visual features F_v (without MLLM). (d) utilizes our full cross-modal artifact perception ($\mathcal{L}_{art}$). (e) is our full MAPO-Net framework.

Model	Vision Mask	Text Tokens	Perception (L_{perc})	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
(a) Baseline (AFUNet)				44.97	0.9935	42.82	0.9910	71.98
(b) Baseline + Perception Loss			✓	44.73	0.9985	46.59	0.9960	72.33
(c) w/o MLLM Text Tokens (Vision-only)	✓			44.82	0.9937	42.88	0.9912	72.02
(d) Baseline + Cross-modal	✓	✓		44.77	0.9989	46.57	0.9983	72.85
(e) Full Model (Ours)	✓	✓	✓	**45.92**	**0.9996**	**46.92**	**0.9996**	**73.29**

5.5 Ablation Study

As shown in Table 4, we conduct a systematic ablation study on the core modules of our proposed framework.

1. Relying solely on the perceptual loss yields a slight improvement in metrics over the baseline model, indicating that the introduction of the perceptual loss is effective.
2. When only visual features are utilized in the cross-modal module (i.e., removing the text tokens from the MLLM), the improvements in PSNR and SSIM are marginal, with HDR-VDP-2 increasing by a mere 0.04.
3. The full cross-modal mechanism, which integrates visual and textual features, substantially outperforms the vision-only variant. While removing the MLLM yields only marginal gains over the baseline, its inclusion significantly improves PSNR and SSIM and increases HDR-VDP-2 by 0.44, confirming the indispensability of MLLM-derived semantic priors.
4. Finally, integrating all modules consistently surpasses the baseline and all single-module variants in PSNR and SSIM. HDR-VDP-2 also rises from 71.98 to 73.29 (+1.31), confirming the superior perceptual quality of the reconstructed images and the effectiveness of MAPO-Net.

6 Conclusion

We propose MAPO-Net, a closed-loop optimization framework that employs a frozen multimodal large language model (MLLM) as a diagnostic expert to identify ghosting artifacts and guide dynamic HDR reconstruction. Furthermore, we introduce MPA-Detector, a cross-modal detector that converts high-level semantic cues into pixel-level spatial masks. These masks enable the reconstruction process to focus on severely degraded regions, providing spatial guidance for targeted ghosting suppression during HDR reconstruction. In addition, we construct the Hard-HDR dataset, which contains 2,300 multi-exposure LDR image sequences and their corresponding ground-truth HDR images. This dataset addresses limitations of existing public datasets by providing paired LDR and HDR images spanning diverse scenes and capture devices. Extensive experiments on public benchmark datasets and Hard-HDR demonstrate that MAPO-Net achieves superior reconstruction quality.

AI USE STATEMENT In this work, we used generative AI tools solely to polish the language and improve the readability of the manuscript. We did not use generative AI tools for designing experiments, analyzing data, generating research ideas, or contributing to the scientific findings of this work. All AI-assisted text was reviewed by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts

ETHICS STATEMENT Our study does not involve human subjects, sensitive data, or personally identifiable information. The datasets we use are publicly available and commonly adopted in the community. We are not aware of potential ethical risks related to privacy, fairness, or misuse, and we have read and adhere to the ICLR Code of Ethics.

Reproducibility Statement We are committed to ensuring the reproducibility of our results. Implementation details, including the model architecture and training setup, are provided in Section 4 of the main paper. The dataset is introduced in Section 3, and its processing steps are described in the Experiments section. Furthermore, we provide the source code and scripts in the supplementary materials to facilitate reproduction.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Marcos V Conde et al. InstructIR: High-quality image restoration following human instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in neural information processing systems*, 34:15908–15919, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):144:1–144:12, 2017.

Haotian Liu, Chunyuan Li, Qinghe Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023a.

Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *European Conference on computer vision*, pp. 344–360. Springer, 2022.

Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Alignment-free HDR deghosting with semantics consistent transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12295–12304, 2023b.

Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Towards high-quality hdr deghosting with conditional diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2552–2566, 2023c.

Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

Zishun Ni, Dong-Liang Yang, Wen-Jie Wang, Si-Yuan Cao, and Song-Hai Zhang. Semantic masking with curriculum learning for robust hdr image reconstruction. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1726–1735, 2022.

Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021.

OpenAI. GPT-4V(ision): A first look at the system's capabilities. *arXiv preprint arXiv:2309.17421*, 2023.

Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Eduardo Perez-Pellitero, Sibi Bhardwaj, Radu Timofte, et al. NTIRE 2023 challenge on high dynamic range imaging: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1718–1732, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Lin Wang and Kuk-Jin Yoon. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8874–8895, 2022. doi: 10.1109/TPAMI.2021.3123686.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7794–7803, 2018.

Haoning Wu, Zicheng Zhang, Erli Zhang, et al. Q-Align: Teaching large language models for visual scoring. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13095–13105, 2023.

Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1751–1760, 2019.

Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020.

Qingsen Yan, Weiye Chen, et al. A unified hdr imaging method with pixel and patch level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14461–14470, 2023.

Qingsen Yan, Weiye Chen, Song Jiang, Hao Tang, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Afunet: Cross-iterative alignment-fusion synergy for hdr reconstruction via deep unfolding paradigm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zhiyuan You et al. DepictQA: Detailed prompt instruction for image quality assessment. *arXiv preprint arXiv:2312.08962*, 2023.

Fangyao Yuan et al. SUPIR: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. *arXiv preprint arXiv:2401.13627*, 2024.

A ADDITIONAL TABLES Table 1 is divided into three separate tables in the main text to provide a more detailed presentation of the results.

Table 5: Quantitative comparison on the Kalantari17 dataset. All models are trained on our proposed Hard-HDR dataset. The best results are highlighted in **bold**, and the second-best results are underlined.

Methods	Venue	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
DHDR Kalantari & Ramamoorthi (2017)	ECCV'18	41.92	0.9871	41.65	0.9892	70.85
NHDRR Yan et al. (2020)	TIP'20	41.88	0.9865	41.49	0.9883	68.55
HDRGAN Niu et al. (2021)	TIP'20	43.92	0.9908	41.51	0.9891	71.10
DiffHDR Liu et al. (2023c)	TCSVT'23	44.06	0.9912	41.68	0.9894	71.16
HyHDR Yan et al. (2023)	CVPR'23	44.47	0.9923	41.89	0.9898	71.84
SMHDR Ni et al. (2022)	ACM MM'22	44.62	0.9932	42.11	0.9902	71.92
HDR-Transformer Liu et al. (2022)	ECCV'22	44.82	0.9944	42.23	0.9881	70.58
AFUNet (Baseline) Yan et al. (2024)	TPAMI'24	44.97	0.9935	42.82	0.9910	71.98
MAPO-Net	**Ours**	**45.20**	**0.9958**	**43.29**	**0.9920**	**72.80**
		(+0.23)	(+0.0023)	(+0.47)	(+0.0010)	(+0.82)

Table 6: Quantitative comparison on the TEL dataset. All models are trained on our proposed Hard-HDR dataset. The best results are highlighted in **bold**, and the second-best results are underlined.

Methods	Venue	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
DHDR Kalantari & Ramamoorthi (2017)	ECCV'18	40.25	0.9803	43.42	0.9931	68.12
NHDRR Yan et al. (2020)	TIP'20	37.78	0.9599	40.11	0.9832	68.21
HDRGAN Niu et al. (2021)	TIP'20	41.28	0.9842	45.44	0.9937	70.19
DiffHDR Liu et al. (2023c)	TCSVT'23	42.01	0.9865	45.63	0.9941	71.43
HyHDR Yan et al. (2023)	CVPR'23	42.32	0.9871	45.72	0.9944	71.68
SMHDR Ni et al. (2022)	ACM MM'22	43.19	0.9881	46.92	0.9951	72.38
HDR-Transformer Liu et al. (2022)	TPAMI'24	42.44	0.9858	46.39	0.9944	71.22
AFUNet (Baseline) Yan et al. (2024)	TPAMI'24	43.65	0.9889	47.94	0.9983	73.47
MAPO-Net	**Ours**	**44.76**	**0.9896**	**48.15**	**0.9991**	**74.39**
		(+1.11)	(+0.0007)	(+0.21)	(+0.0008)	(+0.92)

Table 7: Quantitative comparison on the proposed Hard-HDR dataset. All models are trained on our proposed Hard-HDR dataset. The best results are highlighted in **bold**, and the second-best results are underlined.

Methods	Venue	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
DHDR Kalantari & Ramamoorthi (2017)	ECCV'18	40.90	0.9850	43.54	0.9887	68.77
NHDRR Yan et al. (2020)	TIP'20	39.90	0.9859	42.53	0.9882	68.10
HDRGAN Niu et al. (2021)	TIP'20	41.31	0.9888	43.66	0.9888	69.88
DiffHDR Liu et al. (2023c)	TCSVT'23	42.88	0.9898	44.82	0.9890	70.85
HyHDR Yan et al. (2023)	CVPR'23	43.73	0.9988	45.43	0.9921	71.66
SMHDR Ni et al. (2022)	ACM MM'22	44.29	0.9989	45.48	0.9922	71.87
HDR-Transformer Liu et al. (2022)	ECCV'22	44.64	0.9991	46.88	0.9948	72.37
AFUNet (Baseline) Yan et al. (2024)	TPAMI'24	44.70	0.9994	46.42	0.9952	72.78
MAPO-Net	**Ours**	**44.89**	**0.9995**	**46.82**	**0.9971**	**73.99**
		(+0.19)	(+0.0001)	(+0.40)	(+0.0019)	(+1.21)