MLLM-ITM: Multimodal Large Language Model Promotes Inverse Tone Mapping

Jingchao Peng[1,2], **Thomas Bashford-Rogers**[1] and **Haitao Zhao**[2*] and **Kurt Debattista**[1]
[1]Warwick Manufacturing Group, University of Warwick, Coventry, CV4 7AL, UK
[2]School of Information Science and Engineering, East China University of Science and Technology, Meilong Road 130, Shanghai, 200237, China
{Jingchao.Peng, Thomas.Bashford-Rogers, K.Debattista}@warwick.ac.uk, haitaozhao@ecust.edu.cn

Abstract

High dynamic range (HDR) imaging is crucial for capturing real-world lighting conditions. HDR imaging is traditionally achieved either by fusing multiple exposure frames or via inverse tone mapping from a single SDR image. However, the multi-exposure HDR method is prone to motion-induced artefacts and imposes demanding hardware requirements, limiting its practical applicability. Traditional inverse tone-mapping techniques primarily rely on pixel-wise regression methods, which ignore semantic scene contexts and thus frequently introduce halo artefacts and structural distortions. To address this limitation, this paper proposes MLLM-ITM, a novel inverse tone-mapping framework incorporating multimodal large language models (MLLMs). MLLM-ITM utilizes cross-modal features extracted from a frozen MLLM to simultaneously encode visual features and semantic understanding. These features are integrated into a downstream HDR reconstruction backbone through lightweight adapters, enabling content-aware dynamic-range expansion. Furthermore, the decoupled design between the frozen MLLM and the HDR backbone avoids costly MLLM fine-tuning and enables portability across compatible frozen MLLM backbones with minimal interface adaptation. Extensive experiments on public benchmarks demonstrate that the proposed MLLM-ITM achieves state-of-the-art performance compared with existing inverse tone-mapping methods, highlighting the effectiveness of cross-modal semantic priors in enhancing HDR imaging performance.

1 Introduction

High dynamic range (HDR) imaging addresses the limitations of 8-bit standard dynamic range (SDR) capture by significantly extending the luminance range that a camera can record [Wang and Yoon, 2022]. By preserving detail in both brightly illuminated and deeply shadowed areas, HDR imaging allows for a more accurate reproduction of real-world lighting conditions across professional photography, computer vision, remote sensing, and emerging mixed-reality scenarios. This capability has grown increasingly relevant as more and more vision-related applications appear, enabling significant improvements in visual recognition, semantic understanding, and immersive media experiences, particularly under challenging illumination conditions.

Figure 1: Visual comparison between traditional pixel-to-pixel ITMO and the proposed semantic-aware MLLM-ITM. (a) Generated HDR results of pixel-wise ITMO, which relies solely on the learned SDR-to-HDR transformation. (b) Generated HDR results of the proposed MLLM-ITM, which integrates semantic understanding and visual features extracted from MLLM.

Current HDR capture methods typically employ multi-exposure HDR or inverse tone mapping. Multi-exposure HDR imaging involves capturing multiple images at different exposure levels and subsequently fusing them to reconstruct HDR content [Lecouat *et al.*, 2022]. However, this approach is susceptible to ghosting artefacts due to object movement and requires longer capture times, constraining its usability in real-world applications [Tel *et al.*, 2023]. On

*Corresponding author.

Figure 2: Pink-shaded regions share the same luminance range.

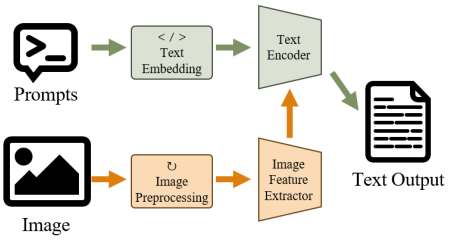

Figure 3: Typical architecture of MLLMs.

the other hand, inverse tone-mapping operators (ITMOs) reconstruct HDR content from a single SDR exposure, which reduces capture time compared with multi-exposure HDR. Nevertheless, ITMOs inherently struggle to restore information lost in severely clipped or underexposed regions, especially since most ITMOs lack semantic understanding, limiting their effectiveness. Recently, hybrid imaging methods fusing complementary sensor modalities, such as visible with infrared [Peng *et al.*, 2026a; Chen *et al.*, 2026] or with event-based cameras [Wang *et al.*, 2019], have emerged as promising solutions for addressing these limitations. Nonetheless, such hybrid approaches generally require specialized hardware and precise cross-sensor calibration, presenting practical challenges in widespread adoption.

When only SDR images are available (usually when the existing content needs to be converted into HDR content or the hardware is limited), the only viable approach is through ITMOs. However, most existing ITMOs learn an SDR-to-HDR mapping from paired HDR/SDR data, effectively performing a locally pixel-to-pixel operation that disregards the content-dependent nature of luminance expansion. Since identical luminance values can correspond to significantly different real-world radiances depending on scene semantics, such as whether the pink-shaded regions in Fig. 2 correspond to the sky, foliage, or an artificial light source, these pixel-to-pixel mappings frequently fail. Fig. 1 contrasts a pixel-wise ITMO with our proposed approach: the former demonstrates halo artefacts along high-contrast edges and structural distortions around artificial lighting sources (highlighted by red boxes), indicative of inappropriate mapping due to the absence of semantic context.

To address this limitation, multi-modal large language models (MLLMs) provide an effective means of incorporating semantic context. MLLMs have been trained on extensive image-text datasets, learning rich, joint embedding spaces capable of encoding object identities, materials, and illumination conditions. These embeddings have successfully functioned as versatile priors for various vision tasks, including object detection [Huang *et al.*, 2025], image style transfer [Peng *et al.*, 2026b], and so on. Notably, LLM-HDR [Barua *et al.*, 2025] used GPT-4o to generate a self-supervised guide during training, while lacking during inference.

Motivated by the above analysis, this paper introduces MLLM-ITM, an innovative framework integrating an MLLM front-end with an inverse tone-mapping backbone, as shown in Fig. 1. By injecting semantic priors, our pipeline achieves semantic-aware dynamic-range expansion, effectively eliminating halo artefacts and preserving structural integrity even in challenging areas where traditional pixel-wise ITMOs fail.

In summary, our contributions are summarized below:

1. To the best of our knowledge, MLLM-ITM is the first method that employs an MLLM as part of an end-to-end backbone, harnessing its joint visual-language reasoning during both training and inference to directly perform inverse tone mapping, rather than invoking an MLLM only as an external tutor.
2. The proposed framework adopts a decoupled integration strategy that treats the MLLM as an interchangeable, frozen module. This design allows other MLLMs to be plugged in without redesigning the framework, enhancing the portability.
3. Experiments show that by injecting semantic priors learned from billions of image-text pairs, the proposed MLLM-ITM achieves state-of-the-art performance on the public HDR dataset.

2 Related Work

This section first presents HDR imaging methods and then introduces MLLMs.

2.1 High Dynamic Range Imaging

Traditional HDR pipelines recover saturated highlights and suppressed shadows by capturing a bracketed stack of exposures and fusing them into a single radiance map, commonly known as **multi-exposure HDR** [Lecouat *et al.*, 2022]. Considerable effort has gone into suppressing ghosting artifacts that arise when objects or the camera itself move, using optical-flow-based alignment and exposure-time optimisation [Kalantari and Ramamoorthi, 2017]. For video, researchers have explored hardware and sampling variants: multiplying the number of sensors [Tocci *et al.*, 2011], cycling the exposure per frame [Kalantari and Ramamoorthi, 2019], spatially varying exposure across the sensor [Serrano *et al.*, 2016], or exploiting coded optics [Metzler *et al.*, 2020]. While these strategies broaden temporal or spatial coverage, they also increase capture latency, calibration complexity, and the probability of motion artifacts or spatial misregistration, limiting their practicality in dynamic scenes.

Inverse tone mapping, on the other hand, seeks to reconstruct an HDR image directly from a single SDR image, turning HDR reconstruction into a learned mapping. Early work applied global linear or non-linear expansions [Masia *et al.*, 2009; Bist *et al.*, 2017] and content-aware guidance maps [Banterle *et al.*, 2006; Rempel *et al.*, 2007;

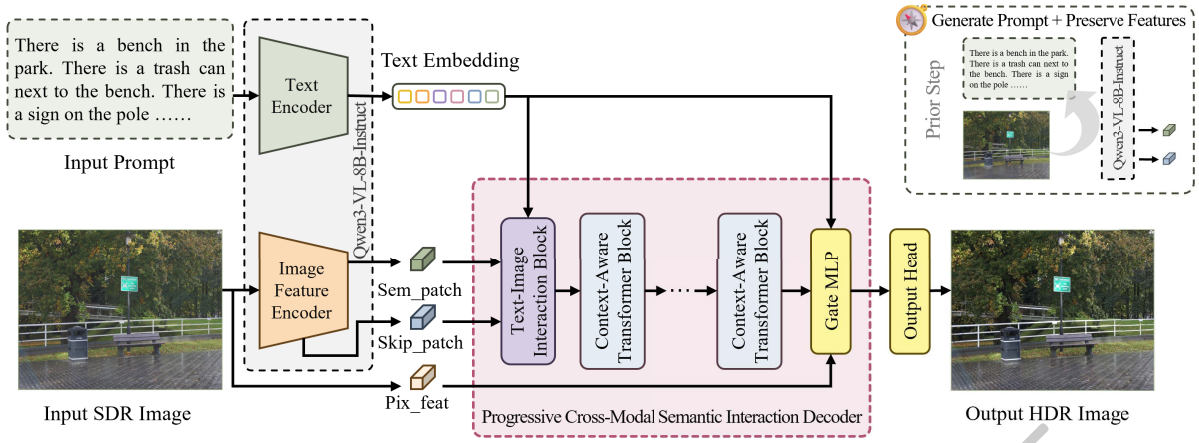

Figure 4: Overall architecture of the proposed MLLM-ITM.

Kovaleski and Oliveira, 2014], but struggled to reconstruct heavily clipped regions [Wang *et al.*, 2007; Didyk *et al.*, 2008]. Deep learning has markedly improved performance: U-Net-based models [Eilertsen *et al.*, 2017] augmented with attention mechanisms [Yu *et al.*, 2021a] and edge-aware decompositions [Zhang and Aydin, 2021], multi-branch architectures that balance quality and efficiency [Marnerides *et al.*, 2018], 3D-UNets that predict missing exposure stacks [Endo *et al.*, 2017], GAN-driven refinement [Lee *et al.*, 2018], and transformer-guided exposure selection [Zhang *et al.*, 2023] all advance reconstruction fidelity. Nevertheless, because of a lack of semantic understanding, ITMOs remain vulnerable to color shifts and unrealistic textures. This becomes more serious when severe over- or underexposure occurs, motivating research into auxiliary modalities or priors to stabilize the expansion.

2.2 Multimodal Large Language Models

Early MLLMs adopted a modular architecture that bridges a frozen vision encoder with a frozen or lightly tuned large language model (LLM) via a compact adapter module. BLIP-2 [Li *et al.*, 2023], for instance, aligns EVA-CLIP [Sun *et al.*, 2023] visual features with OPT [Zhang *et al.*, 2022] or Flan-T5 using a Query-Former mechanism, enabling zero-shot image captioning and visual question answering. Similarly, MiniGPT-4 demonstrates that even a single linear projection trained on instruction-style image-text pairs can improve conversational multimodal capabilities [Zhu *et al.*, 2024]. These bridge-based approaches significantly reduce training cost and resource demands but inherit the resolution limitations and modality gaps of their fixed backbone components.

To overcome these constraints, recent research has explored end-to-end fine-tuning of all parameters, giving rise to modern MLLMs. DeepSeek-VL initially tunes adapter layers locally before transitioning to joint optimization [Lu *et al.*, 2024]. A notable example is Qwen3-VL [Bai *et al.*, 2025], which provides a family of dense and mixture-of-experts vision-language models and supports strong multimodal perception, reasoning, and long-context understanding. These comprehensive strategies consistently improve performance across a wide range of tasks, including caption generation, image grounding, and video-based question answering. However, they also introduce significant computational demands.

Most modern MLLMs follow a two-stream architecture [Li *et al.*, 2025], as shown in Fig. 3. First, each modality is preprocessed independently: textual prompts are tokenized and converted to fixed-width vectors by a text embedding layer, while raw images undergo a preprocessor (resize, pad, normalize) before a vision backbone transforms them into a sequence of patch features. Then, the two streams are fed into the language stack, where cross-modal attention fuses semantics during the decoding process. This unified architecture enables the decoupling of the downstream model from MLLMs.

3 Methodology

This section first provides an overview of the proposed MLLM-ITM and then introduces its main modules.

3.1 Overall Architecture

The overall structure of the proposed MLLM-ITM framework is shown in Fig. 4. MLLM-ITM integrates an MLLM as an encoder, a progressive cross-modal decoder designed for dynamic-range expansion, and an output head responsible for HDR radiance reconstruction. We adopt Qwen3-VL-8B-Instruct [Bai *et al.*, 2025] as our off-the-shelf frozen MLLM encoder.

First, a brief textual description of the input scene is generated using the prompt template “Please describe this image” (gray box, upper right in Fig. 4). This combined caption is processed alongside the input SDR image by the MLLM backbone. The text encoder converts the prompt into text tokens. During this process, three parallel feature streams are extracted from the vision encoder and preserved: **Pix_feat**, low-level visual features extracted from the original input image while preserving full spatial resolution; **Sem_patch**, the feature from the high-level MLLM encoder that contains semantically rich information; and **Skip_patch**, the feature from the early-layer MLLM encoder preserving local detail. All parameters of the MLLM remain frozen, allowing the

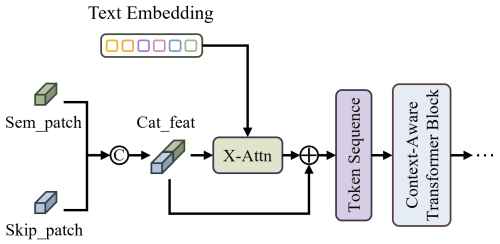

Figure 5: Architecture of the Text-Image Interaction Block.

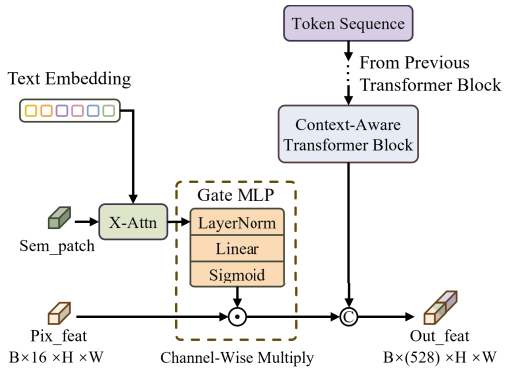

Figure 6: Architecture of the Gate-MLP.

HDR backbone to benefit from large-scale pretraining and remain compatible with other MLLM variants.

Then, three streams are fed to the progressive cross-modal semantic interaction decoder (PSCID), which is shown in the red box of Fig. 4. PSCID contains a **Text-Image Interaction Block**, which injects language tokens into *Sem_patch* and *Skip_patch* via cross-attention, aligning visual features with scene semantics; a cascade of **Context-Aware Transformer Blocks** that iteratively propagate information, enabling the fusion of long-range contextual cues with fine-grained local details for more faithful HDR image reconstruction; and a **Gate-MLP** that adaptively balances semantics-driven and pixel-driven cues, preventing either from dominating.

Finally, a shallow convolutional head upsamples the fused representation and predicts the HDR radiance map. Since task-specific learning is confined to the HDR decoder, the overall design delivers strong performance with minimal training while retaining plug-and-play flexibility for future MLLMs.

In our Qwen3-VL-8B implementation, the frozen Qwen backbone contains 8.77B parameters, while the PSCID decoder contains only 43.29M parameters, making the task-specific decoder about 202.5× smaller than the frozen MLLM backbone. The lightweight claim therefore refers to the module-level decoder/head overhead relative to the frozen backbone rather than the full end-to-end latency. To curb caption-driven artifacts, 20% of training captions are intentionally replaced with mismatched or nonsense text, desensitizing the model to potential hallucinations from the frozen MLLM.

3.2 Text-Image Interaction Block

The Text-Image Interaction Block serves as a key component for aligning semantic priors from the language modality with spatially resolved visual features, as illustrated in Fig. 5. Let $\mathbf{T} \in \mathbb{R}^{N_t \times d}$ represent the language tokens produced by the text encoder, and let $\mathbf{P}_s, \mathbf{P}_k \in \mathbb{R}^{N_p \times d}$ denote the *Sem_patch* and *Skip_patch* tokens emitted by the visual encoder. These two sets of tokens capture complementary visual cues: *Sem_patch* encodes high-level semantics, while *Skip_patch* retains low-level appearance and structural detail. To enable precise alignment without introducing modality imbalance, we employ an asymmetric cross-attention (X-Attn) mechanism in which the image tokens act as queries and the language tokens serve as keys and values:

$$\hat{\mathbf{P}} = \mathrm{Softmax}\left(\frac{\mathbf{Q}(\mathbf{P}_{s,k})\,\mathbf{K}^{\top}(\mathbf{T})}{\sqrt{d}}\right)\mathbf{V}(\mathbf{T}), \qquad (1)$$

where $\mathbf{P}_{s,k} = \mathbf{P}_s \copyright \mathbf{P}_k$ denotes the concatenated sequence of semantic and skip tokens, and $\mathbf{Q}(\cdot), \mathbf{K}(\cdot), \mathbf{V}(\cdot)$ are learnable linear projections. This step ensures that each visual token selectively attends to the most relevant textual cues in a spatially localized fashion, avoiding the global semantic drift that naive concatenation can introduce. The updated tokens are then computed as:

$$\mathbf{P}^{*} = \mathbf{P}_{s,k} + \hat{\mathbf{P}}, \qquad (2)$$

This semantically enriched visual representation is subsequently fed into a cascade of Context-Aware Transformer blocks, adapted from CA-ViT [Liu *et al.*, 2022], which have been shown to be effective for HDR reconstruction in their original work. By incorporating language-derived semantics at this early stage, the network becomes capable of performing content-aware dynamic range expansion. This results in more faithful reconstruction, particularly in challenging regions with high contrast or subtle texture, while effectively suppressing artifacts such as halos.

3.3 Gate-MLP

Following the Context-Aware Transformer Blocks, the model possesses two distinct but complementary token representations: (i) *context-aware tokens* $\mathbf{C}$, which incorporate global semantic reasoning and language-conditioned guidance, and (ii) *pixel-centric tokens* $\mathbf{P}$, which retain low-level spatial detail but may be susceptible to noise and artifacts. To reconcile these two modalities, we propose a lightweight gating mechanism termed Gate-MLP, designed to adaptively balance the semantic richness with the spatial fidelity of pixel features.

As shown in Fig. 6, the Gate-MLP module consists of a layer normalization $\mathbf{LN}$ followed by a linear projection $\mathbf{W}$ and a sigmoid activation σ:

$$\mathbf{g} \;=\; \sigma\big(\mathbf{W}(\mathbf{LN}(\mathbf{C}))\big), \qquad \mathbf{g} \in [0,1]^{N\times 1}, \qquad (3)$$

The resulting scalar gate determines the relative importance of the pixel-driven and semantic-driven streams on a per-token basis. To combine the two representations, $\mathbf{g}$ is broadcast across the feature dimension and used to modulate the fusion:

$$\mathbf{F} \;=\; \mathbf{g} \odot \mathbf{P} \copyright \mathbf{C}, \qquad (4)$$

with $\odot$ the Hadamard product. Because $\mathbf{g}$ depends solely on the MLLM output, the network can upweigh language-driven

Methods	PSNR	SSIM	VSI	HDR-VDP-3
DrTMO [Endo *et al.*, 2017]	10.934	0.707	0.978	7.131
Deep Recursive HDRI [Lee *et al.*, 2018]	15.439	0.902	0.982	7.739
HDRTVNet [Chen *et al.*, 2021b]	16.573	0.910	0.985	7.105
LaNet [Yu *et al.*, 2021b]	14.030	0.833	0.978	7.318
HDRCNN [Eilertsen *et al.*, 2017]	17.451	0.904	0.985	7.470
Deep-HDR Reconstruction [Santos *et al.*, 2020]	14.114	0.774	0.976	7.231
ICTCPNet [Huang *et al.*, 2023]	5.560	0.338	0.949	5.605
ExpandNet [Marnerides *et al.*, 2018]	13.691	0.811	0.972	7.036
HDRUNet [Chen *et al.*, 2021a]	9.427	0.609	0.972	6.482
HDRTNet [Peng *et al.*, 2025]	15.679	0.916	0.981	6.686
MLLM-ITM (Ours)	23.524	0.928	0.977	7.775

Table 1: Results on the HDRT dataset. The best results are highlighted in red, and the second-best results in orange.

guidance around critical scene elements while suppressing raw pixel features. This gating mechanism discourages trivial solutions (directly generating HDR from Pix_feat), yielding more faithful HDR reconstructions at a small parameter overhead of 0.49% of the frozen backbone size.

4 Experiments

This section first introduces the experimental settings, dataset, and evaluation metrics. We subsequently compare our MLLM-ITM with other methods. Finally, ablation studies are presented to highlight the contributions of the main components.

4.1 Implementation Details

We used the public HDRT dataset [Peng *et al.*, 2025] to evaluate the proposed MLLM-ITM. HDRT is a real-world dataset that includes paired HDR and SDR images. Since HDRT was originally proposed for infrared-guided HDR reconstruction, it contains poorly exposed SDR images that are difficult to convert into HDR content without supplementary infrared information. We only use the well-exposed SDR images to generate HDR images. The training set included 8,000 images, while the validation set contained 2,000 images. The dataset is partitioned following the official splits of the original HDRT benchmark.

The experiments were implemented and evaluated using PyTorch on an Intel Xeon W-2295 CPU paired with an NVIDIA Quadro RTX 8000 GPU. We trained 200 epochs using the Adam optimizer, with an initial learning rate of 4e-5, adjusted by a MultiStepLR scheduler that halved the learning rate every 20,000 steps. The loss function is the same as HDRT [Peng *et al.*, 2025]:

$$\mathcal{L}_{HDR} = \mathcal{L}_{pix} + \alpha\mathcal{L}_{per} + \beta\mathcal{L}_{GAN}, \tag{5}$$

where $\alpha = 1.0$ and $\beta = 1\mathrm{e}-5$. This loss formulation follows the benchmark framework provided by [Peng *et al.*, 2025], which has been extensively validated in prior work.

For evaluating image quality, we adopt perceptually aligned variants of three standard metrics, PSNR, SSIM, and VSI. In addition, we include HDR-VDP-3 [Mantiuk *et al.*, 2023], a perceptual metric specifically designed for HDR content. Since linear HDR color values do not correspond linearly to human visual perception, we encoded the HDR linear color values into perceptually uniform representations to compute PSNR, SSIM, and VSI.

4.2 Comparison with Other Methods

We benchmark our MLLM-ITM against representative ITMOs on the HDRT dataset. The competitors can be divided into three categories, early deep CNNs with direct SDR to HDR regression (DrTMO [Endo *et al.*, 2017], Deep Recursive HDRI [Lee *et al.*, 2018]); multi-branch or frequency-aware CNNs (LaNet [Yu *et al.*, 2021b], HDRCNN [Eilertsen *et al.*, 2017], Deep-HDR Reconstruction [Santos *et al.*, 2020], ExpandNet [Marnerides *et al.*, 2018], ICTCPNet [Huang *et al.*, 2023], HDRUNet [Chen *et al.*, 2021a]); and the cross-modal fusion model HDRTNet [Peng *et al.*, 2025], which leverages IR data. Quantitative results are shown in Tab. 1.

Our method attains the best score in PSNR, SSIM, and HDR-VDP-3, while remaining competitive in VSI. Specifically, MLLM-ITM delivers 23.524 dB PSNR, surpassing the strongest baseline (HDRCNN, 17.451 dB) by +6.073 dB. It also improves SSIM from 0.916 (HDRTNet) to 0.928 and achieves an HDR-VDP-3 score of 7.775, outperforming Deep Recursive HDRI by 0.036. Although the VSI score is slightly lower than the previous best value of 0.985, MLLM-ITM still achieves a competitive VSI of 0.977 while obtaining the best PSNR, SSIM, and HDR-VDP-3.

Traditional pixel-centric ITMOs struggle to disambiguate identical luminance values belonging to semantically different regions, leading to halo artifacts and color drift at strong contrast edges; this is reflected in their lower SSIM and HDR-VDP-3 scores. By contrast, our MLLM-ITM leverages semantic understanding priors learned from billions of image-text pairs, enabling content-aware dynamic range expansion and more accurate reconstruction of the entire image. The Gate-MLP further prevents over-smoothing, preserving fine texture that translates into higher VSI.

4.3 Visual Performance

Fig. 7 provides qualitative results that complement our quantitative findings, visually comparing the proposed MLLM-ITM against several state-of-the-art inverse tone-mapping methods across seven challenging scenes. For consistent visual comparison, every HDR result presented in this paper is tone-mapped using the Durand operator [Durand and Dorsey, 2002].

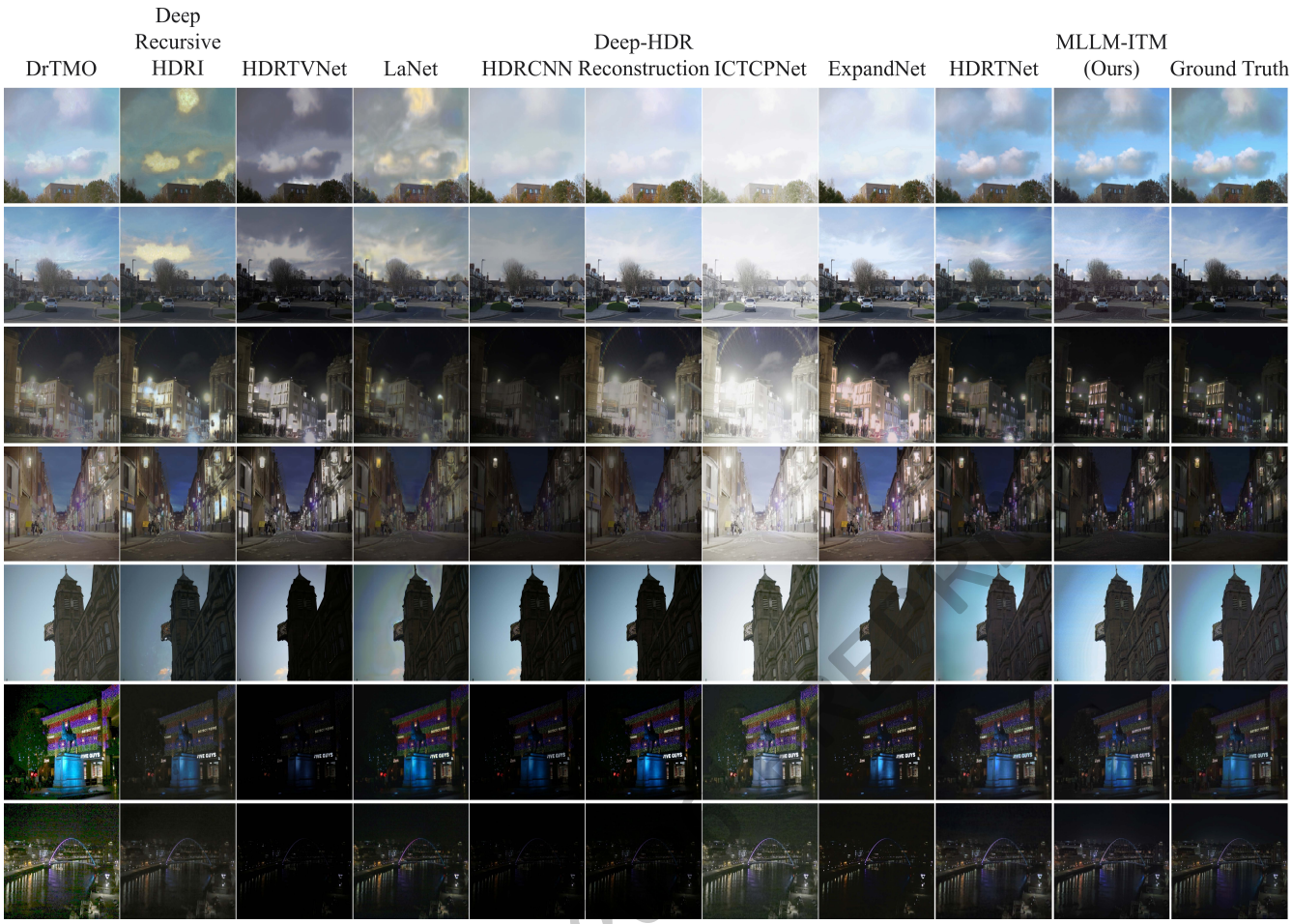

Figure 7: Qualitative results on the HDRT dataset.

In bright daylight scenes with high dynamic contrast (rows 1-2), existing methods such as ICTCPNet and ExpandNet typically clip highlights in the sky region, while methods like Deep Recursive HDRI produce pronounced halo artifacts around high-contrast edges. Conversely, our MLLM-ITM effectively reconstructs cloud textures and preserves structural details in building facades, yielding results visually indistinguishable from the ground truth.

Night-time urban scenes (rows 3-4) and the backlit building scenario in row 5 illustrate another common shortcoming of current ITMOs. Methods such as LaNet and HDRCNN produce overly dim and desaturated images, while ExpandNet and Deep Recursive HDRI excessively amplify streetlights and lamps, causing overexposed regions and characteristic "ring" artefacts. Similarly, DrTMO and ICTCPNet tend to overly brighten the entire scene, deviating substantially from the ground truth appearance. Benefiting from semantic priors, MLLM-ITM generates realistic lighting conditions, maintaining readable signage and accurate color fidelity without introducing artefacts.

Finally, in severely under-exposed scenarios (rows 6-7), MLLM-ITM better recovers mid-tone structural details, such as bridge trusses and neon patterns, along with subtle shadow

Figure 8: Visual comparison without Skip_patch on the left and with Skip_patch on the right.

textures. In contrast, HDRTNet continues to exhibit mild banding and residual noise under these challenging conditions. These visual examples strongly corroborate the quantitative performance improvements detailed earlier in Tab. 1, highlighting the superior effectiveness and general applicability of our proposed MLLM-ITM framework.

4.4 Ablation Study

Tab. 2 reports a step-wise ablation on the HDRT test set, where we progressively enable or disable the four feature streams introduced in Sec. 3.1. The comparison reveals the

Method	Pix_feat	Skip_patch	Sem_patch	Text	PSNR	SSIM	VSI	HDR-VDP-3
Baseline	√	×	×	×	18.396	0.846	0.929	5.805
v1	√	×	√	√	23.169	0.917	0.972	7.441
v2	√	√	×	√	22.858	0.907	0.964	7.112
v3	√	√	√	×	23.160	0.916	0.968	7.236
v4	×	√	√	√	14.842	0.617	0.743	0.379
MLLM-ITM (Ours)	√	√	√	√	**23.524**	**0.928**	**0.977**	**7.775**

Table 2: Ablation results on the HDRT dataset. "√" indicates that the feature is adopted, and "×" indicates that it is not.

Figure 9: Visual comparison without Sem_patch on the left and with Sem_patch on the right.

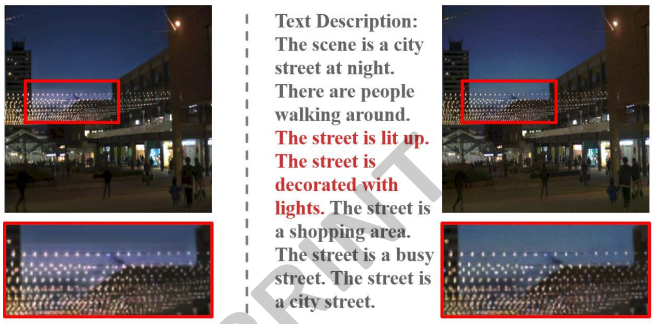

Figure 10: Visual comparison without text description on the left and with text description on the right.

complementary roles of the individual cues and their combined effect on reconstruction quality.

With only **Pix_feat**, the baseline model performs a purely local regression and attains 18.396 dB PSNR. Lacking any semantic guidance, it behaves as a typical pixel-wise ITMO.

No Skip_patch. Removing Skip_patch deprives the network of low-level, high-frequency structural details. As a result, fine textures in foliage and masonry become noticeably blurry and over-smoothed (boxes in Fig. 8), PSNR drops by 0.355 dB, and VSI drops by 0.005 relative to the full model. This confirms that Skip_patch is beneficial for recovering crisp local detail.

No Sem_patch. Suppressing the high-level semantic tokens forces the model to extrapolate tone mapping solely from local context and low-level pixels. The absence of semantic guidance manifests as pronounced halo-box artefacts at high-contrast boundaries, especially around lamp heads (boxes in Fig. 9). Structure score decreases (SSIM 0.907) despite the presence of language cues, indicating that high-level visual semantics are indispensable for robust brightness mapping.

No Text. When explicit language tokens are removed, the network retains visual semantics but loses high-level scene descriptors such as "street-light" or "decorative bulbs". This semantic gap causes inconsistent tone allocation: the sky surrounding the lamp inherits a warm cast taken from the light source, while the other regions remain cool (Fig. 10). The perceptual metric HDR-VDP-3 degrades, underscoring the role of language priors in disambiguating object categories with similar brightness profiles.

No Pix_feat. Pix_feat provides the direct low-level spatial anchor from the raw input image. Disabling Pix_feat forces the decoder to rely exclusively on patch- and text-level representations that have been resized and padded by the MLLM front-end. This leads to a severe degradation, with PSNR dropping from 23.524 dB to 14.842 dB, SSIM from 0.928 to 0.617, VSI from 0.977 to 0.743, and HDR-VDP-3 from 7.775 to 0.379. The loss of accurate pixel coordinates causes color bleeding, edge softness, and unstable local reconstruction. This result demonstrates that the direct convolutional features extracted from the raw input are indispensable as a model-independent spatial anchor for plug-and-play MLLM integration.

5 Conclusion

In this paper, we introduced MLLM-ITM, a novel inverse tone-mapping framework that leverages semantic priors provided by MLLMs. By injecting language and visual tokens into a task-specific HDR decoder, our approach achieves semantic-aware dynamic-range expansion, effectively reducing halo artefacts and preserving structural accuracy in areas where traditional pixel-based ITMOs struggle. The decoupled architecture, where the MLLM remains frozen and easily interchangeable, allows future advancements in MLLMs to be seamlessly incorporated with minimal redesign.

Extensive experiments conducted on the public HDRT dataset validate the effectiveness of our method: MLLM-ITM consistently outperforms current state-of-the-art ITMOs by substantial margins in terms of PSNR, SSIM, and HDR-VDP-3. Ablation studies further demonstrate the complementary benefits of the different features extracted from the MLLM.

Future work will explore further integration of larger-scale MLLMs to enhance HDR reconstruction capabilities. Additionally, future work will extend our framework to other image processing scenarios and integrate it into other computer vision tasks.

Acknowledgments

This work was supported by the Royal Society International Exchanges 2023 Cost Share under Grant IEC\NSFC\233809.

References

[Bai *et al.*, 2025] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

[Banterle *et al.*, 2006] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *GRAPHITE '06*, page 349–356, New York, NY, USA, 2006. ACM.

[Barua *et al.*, 2025] Hrishav Bakul Barua, Kalin Stefanov, Lemuel Lai En Che, Abhinav Dhall, KokSheik Wong, and Ganesh Krishnasamy. Llm-hdr: Bridging llm-based perception and self-supervision for unpaired ldr-to-hdr image reconstruction, 2025.

[Bist *et al.*, 2017] Cambodge Bist, Rémi Cozot, Gérard Madec, and Xavier Ducloux. Tone expansion using lighting style aesthetics. *Comput. Graph.*, 62:77–86, 2017.

[Chen *et al.*, 2021a] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 354–363, 2021.

[Chen *et al.*, 2021b] Xiangyu Chen, Zhengwen Zhang, Jimmy S. Ren, Lynhoo Tian, Yu Qiao, and Chao Dong. A new journey from sdrtv to hdrtv. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4480–4489, 2021.

[Chen *et al.*, 2026] Jingkun Chen, Jingchao Peng, Junde Wu, Tianlu Zhang, Thomas Bashford-Rogers, Kurt Debattista, Vicente Grau, and Jungong Han. Balancing light and shadow: Multi-modal image exposure restoration via frequency-phase-driven rgb-ir integration. *Information Fusion*, 127:103901, 2026.

[Didyk *et al.*, 2008] Piotr Didyk, Rafał Mantiuk, Matthias Hein, and Hans-Peter Seidel. Enhancement of bright video features for HDR displays. *Computer Graphics Forum*, 27(4):1265–1274, 2008.

[Durand and Dorsey, 2002] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics*, 21(3):257–266, jul 2002.

[Eilertsen *et al.*, 2017] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, RafalK. Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Trans. Graph.*, 36(6), nov 2017.

[Endo *et al.*, 2017] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6), nov 2017.

[Huang *et al.*, 2023] Peihuan Huang, Gaofeng Cao, Fei Zhou, and Guoping Qiu. Video inverse tone mapping network with luma and chroma mapping. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pages 1383–1391, New York, NY, USA, 2023. Association for Computing Machinery.

[Huang *et al.*, 2025] Feng Huang, Shuyuan Zheng, Zhaobing Qiu, Huanxian Liu, Huanxin Bai, and Liqiong Chen. Text-irstd: Leveraging semantic text to promote infrared small target detection in complex scenes. In *International Conference on Computer Vision*, 2025.

[Kalantari and Ramamoorthi, 2017] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4), jul 2017.

[Kalantari and Ramamoorthi, 2019] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep HDR video from sequences with alternating exposures. *Comput. Graph. Forum*, 38(2):193–205, 2019.

[Kovaleski and Oliveira, 2014] Rafael Pacheco Kovaleski and Manuel M. Oliveira. High-quality reverse tone mapping for a wide range of exposures. In *27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 49–56, New York, 2014. IEEE Computer Society.

[Lecouat *et al.*, 2022] Bruno Lecouat, Thomas Eboli, Jean Ponce, and Julien Mairal. High dynamic range and super-resolution from raw image bursts. *ACM Trans. Graph.*, 41(4), jul 2022.

[Lee *et al.*, 2018] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 613–628, Cham, 2018. Springer International Publishing.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[Li *et al.*, 2025] Xinyao Li, Jingjing Li, Fengling Li, Lei Zhu, Yang Yang, and Heng Tao Shen. Generalizing vision-language models to novel domains: A comprehensive survey, 2025.

[Liu *et al.*, 2022] 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*, pages 344–360. Springer, 2022.

[Lu *et al.*, 2024] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren,

Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

[Mantiuk *et al.*, 2023] Rafal K. Mantiuk, Dounia Hammou, and Param Hanji. Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content, 2023.

[Marnerides *et al.*, 2018] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. *Computer Graphics Forum*, 37(2):37–49, 2018.

[Masia *et al.*, 2009] Belen Masia, Sandra Agustin, Roland W. Fleming, Olga Sorkine, and Diego Gutierrez. Evaluation of reverse tone mapping through varying exposure conditions. *ACM Trans. Graph.*, 28(5):1–8, 2009.

[Metzler *et al.*, 2020] Christopher A. Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1372–1382. Computer Vision Foundation / IEEE, 2020.

[Peng *et al.*, 2025] Jingchao Peng, Thomas Bashford-Rogers, Francesco Banterle, Haitao Zhao, and Kurt Debattista. HDRT: A large-scale dataset for infrared-guided HDR imaging. *Information Fusion*, 120:103109, 2025.

[Peng *et al.*, 2026a] Jingchao Peng, Thomas Bashford-Rogers, Francesco Banterle, Haitao Zhao, and Kurt Debattista. Zero-shot infrared-guided hdr video deflickering. *Pattern Recognition*, 179:113759, 2026.

[Peng *et al.*, 2026b] Jingchao Peng, Thomas Bashford-Rogers, Zhuang Shao, Haitao Zhao, Aru Ranjan Singh, Abhishek Goswami, and Kurt Debattista. CapHDR2IR: Caption-driven transfer from visible light to infrared domain. *IEEE Transactions on Multimedia*, 28:2865–2875, 2026.

[Rempel *et al.*, 2007] Allan G. Rempel, Matthew Trentacoste, Helge Seetzen, H. David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: On-the-fly reverse tone mapping of legacy video and photographs. *ACM Trans. Graph.*, 26(3):39, 2007.

[Santos *et al.*, 2020] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Trans. Graph.*, 39(4), aug 2020.

[Serrano *et al.*, 2016] Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belén Masiá. Convolutional sparse coding for high dynamic range imaging. *Comput. Graph. Forum*, 35(2):153–163, 2016.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023.

[Tel *et al.*, 2023] Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, and Dominique Ginhac. Alignment-free hdr deghosting with semantics consistent transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12790–12799, 2023.

[Tocci *et al.*, 2011] Michael D. Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile hdr video production system. *ACM Trans. Graph.*, 30(4), jul 2011.

[Wang and Yoon, 2022] 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.

[Wang *et al.*, 2007] Lvdi Wang, Li-Yi Wei, Kun Zhou, Baining Guo, and Heung-Yeung Shum. High dynamic range image hallucination. In *SIGGRAPH '07: ACM SIGGRAPH 2007 Sketches*, page 72, New York, NY, USA, 2007. ACM.

[Wang *et al.*, 2019] Lin Wang, I.S. Mohammad Mostafavi, Yo-Sung Ho, and Kuk-Jin Yoon. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10073–10082, 2019.

[Yu *et al.*, 2021a] Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou, and Chunxia Xiao. Luminance attentive networks for hdr image and panorama reconstruction. *Computer Graphics Forum*, 40(7):181–192, 2021.

[Yu *et al.*, 2021b] Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou, and Chunxia Xiao. Luminance attentive networks for hdr image and panorama reconstruction. *Computer Graphics Forum*, 40(7):181–192, 2021.

[Zhang and Aydin, 2021] Yang Zhang and Tunç Ozan Aydin. Deep HDR estimation with generative detail reconstruction. *Comput. Graph. Forum*, 40(2):179–190, 2021.

[Zhang *et al.*, 2022] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[Zhang *et al.*, 2023] Ning Zhang, Yuyao Ye, Yang Zhao, and Ronggang Wang. Revisiting the stack-based inverse tone mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9162–9171, June 2023.

[Zhu *et al.*, 2024] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.