Adapting Large VLMs with Iterative and Manual Instructions for Generative Low-light Enhancement –Supplementary Material–

Xiaoran Sun[1,†], Liyan Wang[1,†], Yeying Jin[2], Kin-man Lam[3], Zhixun Su[1,*], Yang Yang[4]
Jinshan Pan[5], Cong Wang[4]
[1]Dalian University of Technology [2]National University of Singapore
[3]Hong Kong Polytechnic University [4]University of California, San Francisco
[5]Nanjing University of Science and Technology
{sunxiaoran, wangliyan}@mail.dlut.edu.cn; jinyeying@u.nus.edu; kin.man.lam@polyu.edu.hk
yang.yang4@ucsf.edu; zxsu@dlut.edu.cn; {sdluran, supercong94}@gmail.com

Abstract

In this supplementary material, we first provide additional controllable results in Sec. 1. Next, we provide more results on ablation studies, including more visual results about text instructions, the learnable instruction prior fusion (LIPF) module, and iterative instruction strategy, text encoder selection, the effect of ControlNet, the effect of each training module, and the failure cases of instructions from LLaVA in Sec. 2. Then, we present the efficiency comparison of various low-light image enhancement methods in Sec. 3. And, we also add a user study to assess whether non-expert text inputs reliably improve results in Sec. 4. Subsequently, we provide more visual comparisons with state-of-the-art approaches in Sec. 5. Finally, the limitations of the proposed approach are discussed in Sec. 6.

1. Additional Controllable Results

Figure 10 presents additional controllable enhancement results alongside corresponding Grad-CAM visualizations [14], demonstrating how different text instructions lead to distinct output variations. The Grad-CAM heatmaps reveal that the model primarily focuses on regions emphasized by the input instructions, such as facial areas or background elements. This indicates that our method enables precise and interpretable control in the denoising process, allowing for adaptive enhancement tailored to various light instructions.

2. Additional Results on Ablation Studies

2.1. More Results on Text Instructions (See Table 2 in the main manuscript)

Figures 11-12 and Figures 13-15 present more visual examples demonstrating the impact of text instructions on paired and real-world datasets, respectively. The comparison reveals that a lack of variation in text instructions affects the model's ability to effectively recover fine details and proper illumination in low-light images.

2.2. More Results on Learnable Instruction Prior Fusion (See Table 3 in the main manuscript)

Figure 16 and Figure 17 present more visual examples demonstrating the impact of Learnable Instruction Prior Fusion (LIPF) on paired and real-world datasets, respectively. As can be seen, our full model is able to produce more vivid and realistic results.

† Equal contributions.
∗ Corresponding author.

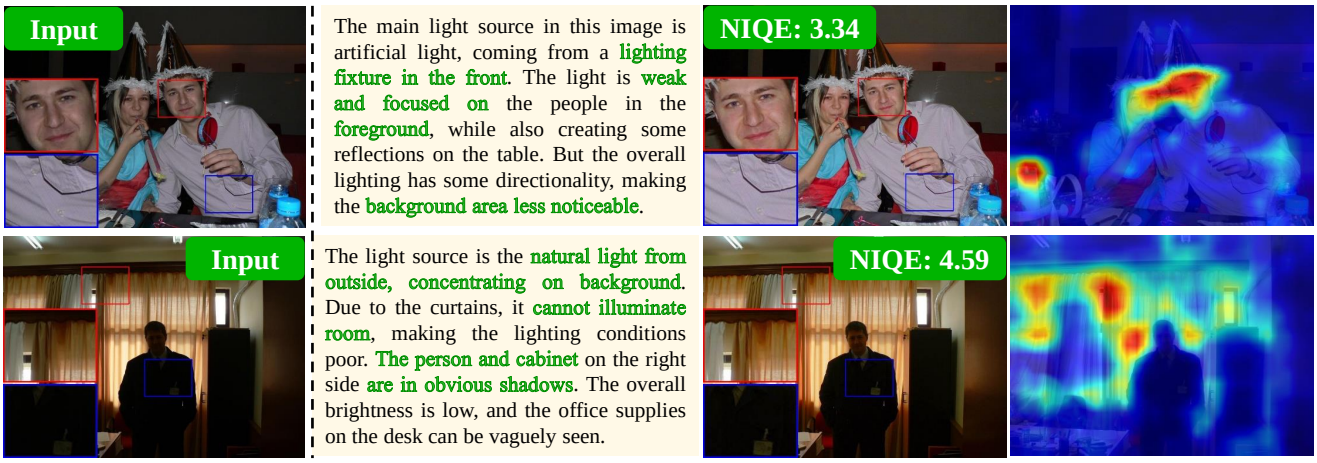


Figure 10. **Text instructions can control lighting enhancement**, producing results under different lighting conditions. The corresponding Grad-CAM heatmaps [14] highlight the model's attention areas influenced by the text instructions, such as faces or background regions, showing how instructions affect visual enhancement.

(a) Input (b) Instructions (k=2) (c) Tab. 2(a) (d) Tab. 2(b) (e) Tab. 2(c) **(f) Ours**

Figure 11. **Visual comparison of text instructions** on paired dataset LOL [20]. Results without and with LIPF are presented in the top and bottom rows, respectively. Our full model generates results with better naturalness according to visual quality and the realism metric, NIQE.

2.3. More Results on the Iterative Instruction strategy (See Figure 8 in the main manuscript)

Figure 18 provides more visualizations of different iterative instruction strategies applied to three real-world datasets. When $k = 2$, the contrast and brightness of the image in the shadowed areas improve significantly, revealing clearer details. The overall image appears more natural in subjective visual perception and exhibits a higher sense of realism.

2.4. Text Encoder Selection

Considering that the structure and parameter size of different large language models may lead to varying semantic understanding capabilities, we conduct experiments on SD2.1 using the T5 encoder [13], BERT [9], and Llama [15], with CLIP text encoder [12] selected as the baseline encoder. As shown in Table 4, T5 outperforms CLIP and BERT in semantic understanding, especially for lighting descriptions and scene details, due to its stronger ability to process diverse instructions. Although Llama achieves the best performance, its large parameter size (7B vs. 770M) significantly increases computational

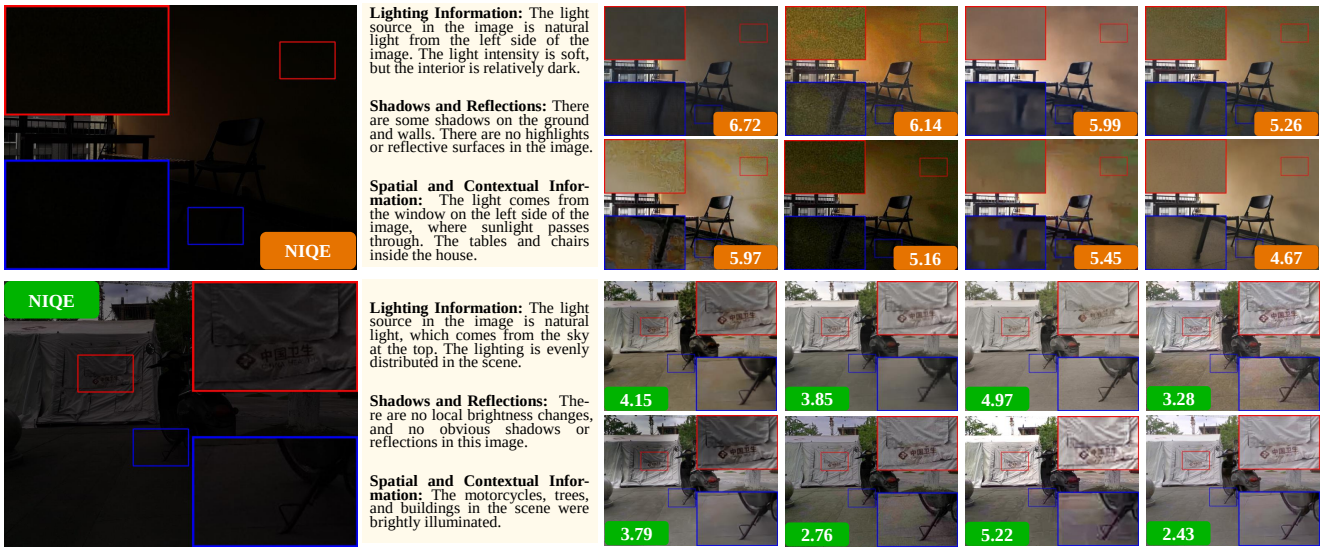


Figure 12. **Visual comparison of text instructions** on paired dataset LSRW [5]. Results without and with LIPF are presented in the top and bottom rows, respectively. Our full model generates results with better naturalness according to visual quality and the realism metric, NIQE.

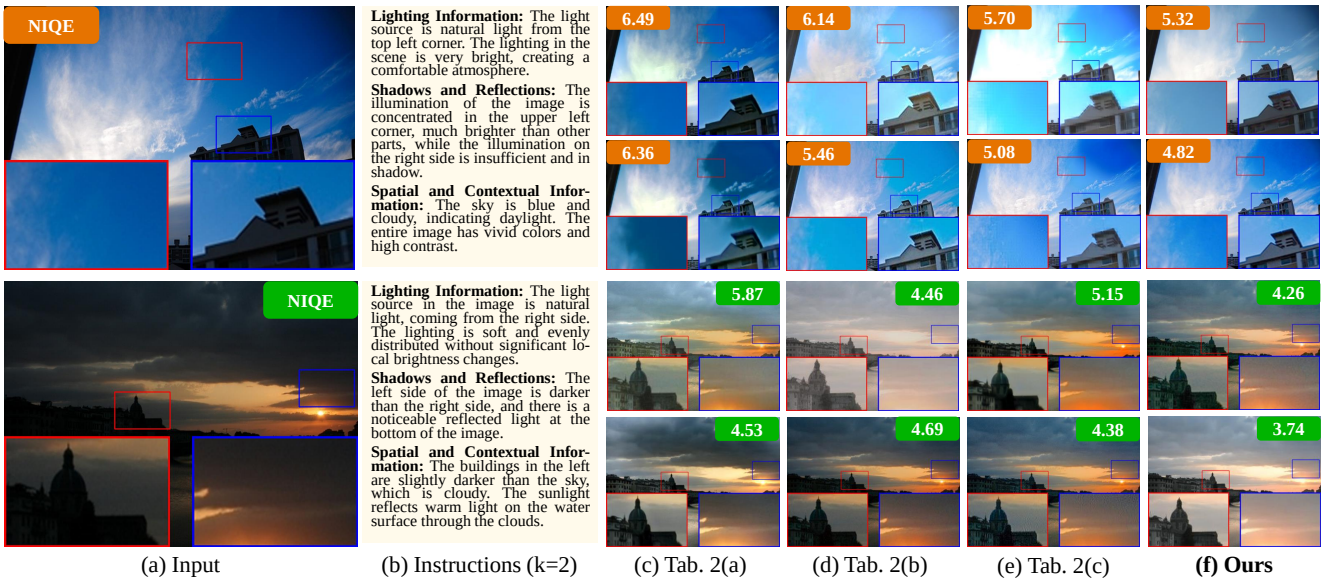
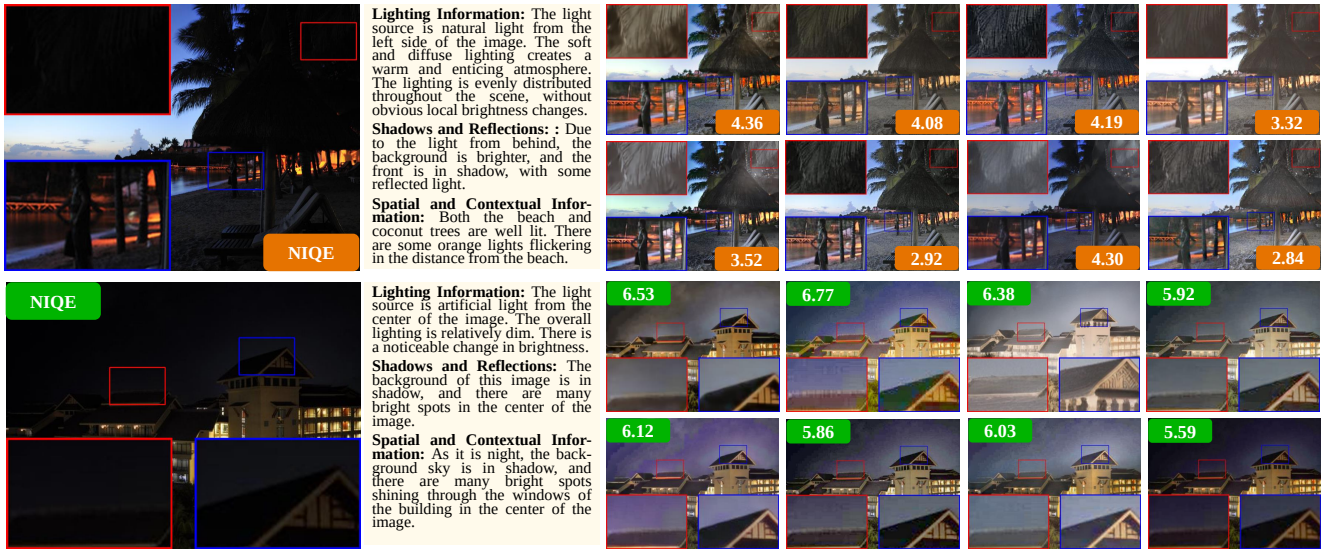


Figure 13. **Visual comparison of text instructions** on real-world dataset DICM [8]. Results without and with LIPF are presented in the top and bottom rows, respectively. Our full model generates results with better naturalness according to visual quality and the realism metric, NIQE.

costs. Balancing performance and efficiency, we select T5 as our final text encoder.

2.5. Effect of Each Training Module

For trainable modules, LIPF fuses semantic instructions with visual features at each diffusion stage. ControlNet ensures structural consistency with the low-light input. And image encoder provides low-light latent embeddings for ControlNet. Thus, we provide an ablation study of fixing each training module to illustrate their contributions (Note image encoder is required), as shown in Table 5. As can be observed, our final full model achieves better performance, demonstrating the



(a) Input (b) Instructions (k=2) (c) Tab. 2(a) (d) Tab. 2(b) (e) Tab. 2(c) **(f) Ours**

Figure 14. **Visual comparison of text instructions** on real-world dataset NPE [17]. Results without and with LIPF are presented in the top and bottom rows, respectively. Our full model generates results with better naturalness according to visual quality and the realism metric, NIQE.

(a) Input (b) Instructions (k=2) (c) Tab. 2(a) (d) Tab. 2(b) (e) Tab. 2(c) **(f) Ours**

Figure 15. **Visual comparison of text instructions** on real-world dataset VV [16]. Results without and with LIPF are presented in the top and bottom rows, respectively. Our full model generates results with better naturalness according to visual quality and the realism metric, NIQE.

effectiveness of each component used in our model.

2.6. Failure Cases of Instruction from LLaVA

LLaVA may generate inaccurate or misleading descriptions. For example, when both natural and artificial light sources coexist in an image, LLaVA generates an inaccurate instruction (e.g., the first row of Figure 19(c)): *The main light source for this image is artificial light, coming from the flickering ceiling lights at the top, illuminating the dim indoor train platform. The lighting in the scene is uniform, projecting clear figures below commuters.* However, LLaVA incorrectly identifies the

(a) Input (b) Instructions (c) Tab. 3(a) (d) Tab. 3(b) (e) Tab. 3(c) **(f) Ours**

Figure 16. **Visual comparison about LIPF** on LOL [20] (upper) and LSRW [5] (lower) datasets. Our VLM-IMI generates more natural and vivid results.

Table 4. Quantitative results of the ablation study on different text encoders using the LOL [20] test sets.

Text Encoder	**Param**	**LOL** [20]		
		PSNR ↑	SSIM ↑	LPIPS ↓
CLIP [12]	151M	17.791	0.663	0.265
BERT [9]	340M	19.864	0.749	0.217
Llama [15]	7B	22.083	0.815	0.162
T5 [13] (Ours)	770M	21.112	0.802	0.155

Table 5. Ablation study of fixing each training module on LOL [20], LSRW [5], DICM [8], NPE [17], and VV [16] datasets.

ID	**Module**	**LOL** [20]			**LSRW** [5]			**DICM** [8]		**NPE** [17]		**VV** [16]	
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ ↑	NIQE ↓	MUSIQ ↑	NIQE ↓	MUSIQ ↑	NIQE ↓
(a)	w/o LIPF & w/o ControlNet	12.513	0.362	0.516	10.112	0.287	0.533	55.903	5.104	56.568	6.511	38.615	5.861
(b)	w/o LIPF & w/ ControlNet	18.534	0.715	0.258	17.419	0.472	0.267	60.279	4.057	59.202	4.637	41.730	4.508
(c)	w/o ControlNet & w/ LIPF	16.779	0.597	0.338	15.283	0.359	0.406	58.428	4.633	57.005	4.372	43.128	4.457
(d)	**Full Model (Ours)**	**21.112**	**0.802**	**0.155**	**19.351**	**0.548**	**0.173**	**64.661**	**3.551**	**64.097**	**3.421**	**45.984**	**3.523**

main light source as artificial, whereas the actual scene is primarily illuminated by natural daylight, e.g., the first row of Figure 19(f). To better understand the effect, we first collect five real-world low-light images that produce the hallucination problem of VLM. We then conduct experiments on these five images to analyze the effect of LIPF. The results are summarised in Table 6 and Figure 19. The results demonstrate that our VLM-IMI with LIPF delivers better performance on both metrics than Diff-Plugin. Although our VLM-IMI without LIPF achieves slightly higher MUSIQ scores than the model with LIPF, it produces worse metrics in NIQE. Overall, the hallucination problem of VLM affects our final enhancement quality but has a slight influence, even still outperforming the state-of-the-art approach when our method faces the hallucination problem of VLM.

Figure 17. **Visual comparison about LIPF** on DICM [8] (upper), NPE [17] (middle), and VV [16] (lower). Our VLM-IMI generates more natural and vivid results.

Table 6. Failure cases of instruction from LLaVA. We first collect five real-world low-light images that produce the hallucination problem of VLM. We then conduct experiments on these five images to analyze the effect of LIPF. Note that "Right Case" denotes the enhanced results produced by using the correct instruction.

Methods	MUSIQ ↑	NIQE ↓
(a) Diff-Plugin [11]	55.290	3.891
(b) VLM-IMI (w/o LIPF)	56.736	3.917
(c) VLM-IMI (w/ LIPF)	55.977	3.809
(d) Right Case	58.249	3.562

3. Efficiency Comparison

We report the efficiency comparison of various methods at a resolution of 512×512 on an RTX A6000 GPU, as shown in Table 7. Existing vision-language model-based methods also face the challenges of running time. We acknowledge that higher computational demand makes VLM-IMI unsuitable for real-time or resource-constrained scenarios. Our design instead prioritizes enhancement quality and controllability, which are especially valuable for image enhancement applications.

Table 7. Efficiency comparison of various methods at a resolution of 512×512 on an RTX A6000 GPU.

Methods	Params (M)	FLOPs (G)	Inference Time (s)
PairLIE [4]	0.34	5.59	0.07
NeRCo [21]	45.73	1013.62	2.72
CLIP-LIT [10]	0.28	72.85	2.38
Retinexformer [1]	1.61	62.36	2.01
DiffLL [7]	22.08	87.83	5.17
GSAD [6]	17.43	9.42	4.81
QuadPrior [19]	859.52	1220.88	7.81
Diff-Plugin [11]	14.86	0.74	6.33
UniProcessor [3]	623.29	1357.04	5.24
InstructIR [2]	17.11	64.76	2.71
GPP-LLIE [22]	131.18	1425.79	4.10
GEFU [18]	1.70	913.46	8.11
VLM-IMI	481.11	1162.15	$k=1$: 3.75; $k=2$: 8.64 Manual: 3.91

4. User Studies

We conduct a user study to assess whether non-expert text inputs reliably improve results. Specifically, we select 10 real-world low-light images and generate enhanced outputs under both iterative instruction ($k = 1$, $k = 2$) and manual instruction modes. For each image, we create three types of manual instructions: *(a) The lighting in this image is sufficient, ensuring that the entire image is clear and visible. (b) The scene is lit by natural light from the distant sky. The foreground is relatively darker compared to the more illuminated background. (c) The lighting in the scene is soft and evenly distributed, providing consistent illumination without strong variations in brightness or shadows.* We then present the enhanced results alongside their corresponding instructions to five users, who rate whether the enhancement matched the intent of the instruction (e.g., user satisfaction). The results are summarized in Table 8. Manual instructions yield high user satisfaction rates (up to 92%), which demonstrates that non-expert textual inputs can reliably guide the enhancement process.

Table 8. User Studies. We compare iterative instructions (k=1, k=2) and three types of manual instructions: (a) *The lighting in this image is sufficient, ensuring that the entire image is clear and visible.* (b) *The scene is lit by natural light from the distant sky. The foreground is relatively darker compared to the more illuminated background.* (c) *The lighting in the scene is soft and evenly distributed, providing consistent illumination without strong variations in brightness or shadows.* Five users judge whether each enhanced result matched the intended instruction (e.g., user satisfaction).

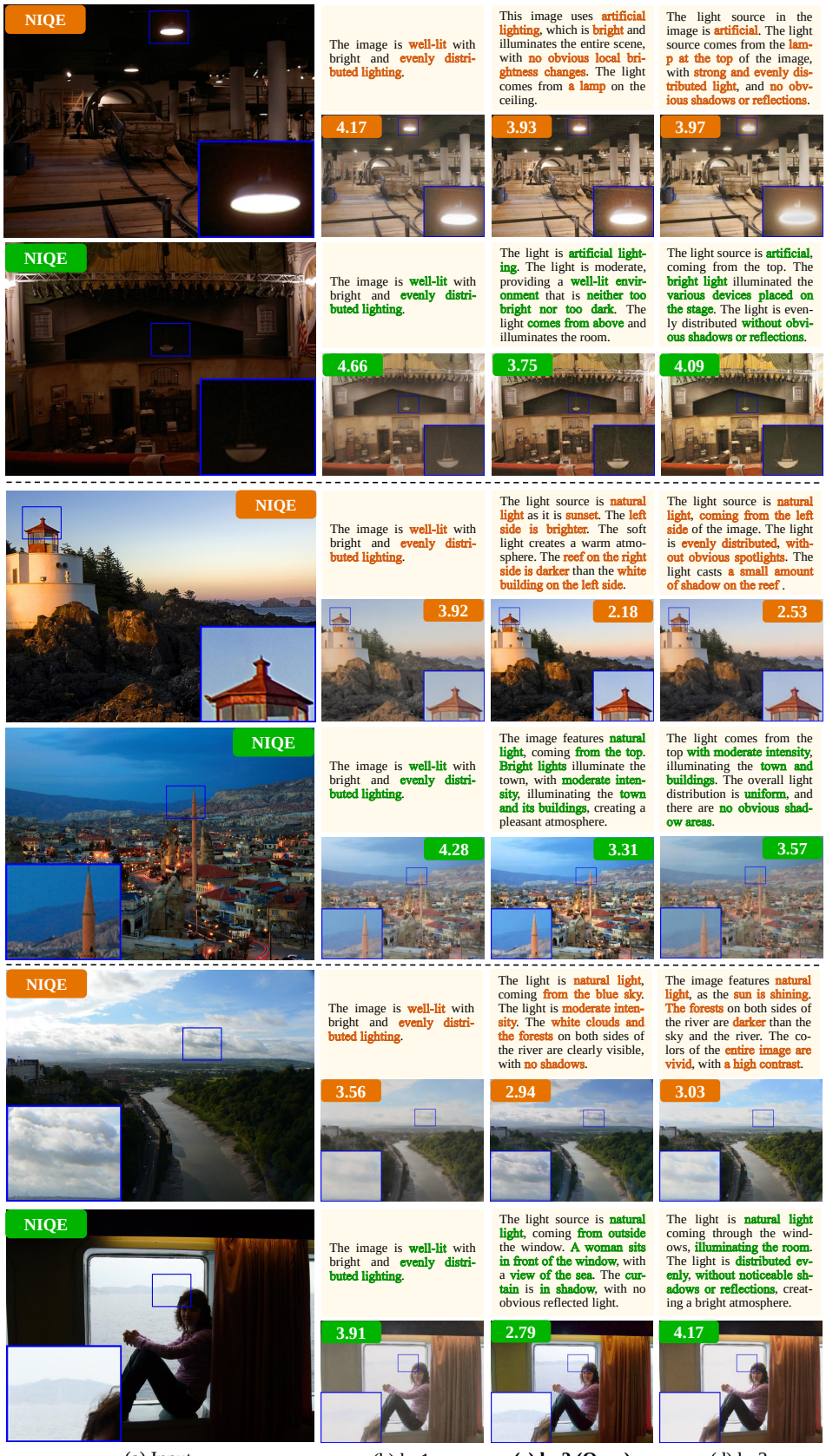
Options	Iterative Instructions		Manual Instructions		
	k =1	k = 2	(a)	(b)	(c)
User Satisfaction	76%	88%	90%	92%	86%

5. Additional Comparison Results with State-of-the-art Approaches

We provide more visualizations of various low-light image enhancement methods on five datasets. Figures 20-21 show more visual comparison results on paired datasets, LOL [20] and LSRW [5]. Figures 22-24 show present visual comparison results on real-world datasets, DICM [8], NPE [17], and VV [16]. It can be observed that our proposed VLM-IMI achieves the best visual results with more natural outputs. Moreover, our proposed VLM-IMI also obtains the realism metric, *i.e.*, NIQE.

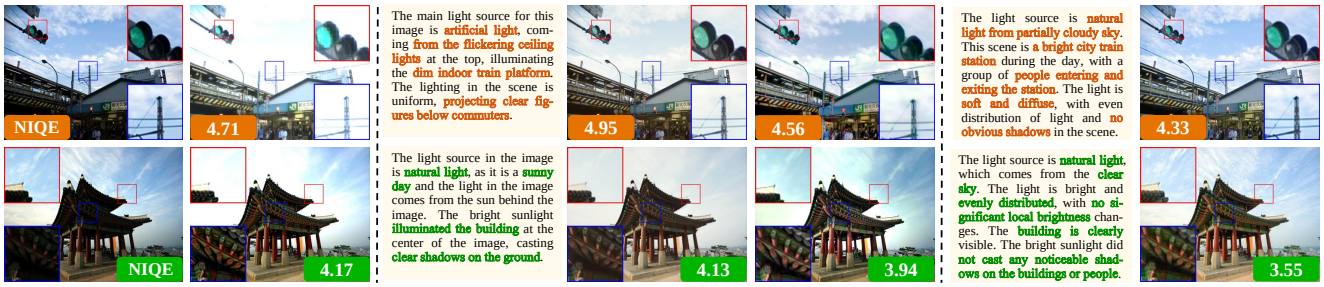
6. Limitations

Extreme low-light scenarios, where scenes are heavily degraded by darkness and noise with minimal useful signal remaining, exceed the capabilities of our method. Furthermore, the diffusion-based framework, although robust to noise, still relies on the presence of discernible features in the latent space to guide the generation. When those features are lost due to extreme lighting conditions, the model's ability to infer plausible outcomes diminishes. As one of the future directions, we aim to integrate auxiliary data (e.g., infrared signals or event-based cameras) to enhance performance in highly challenging scenarios.



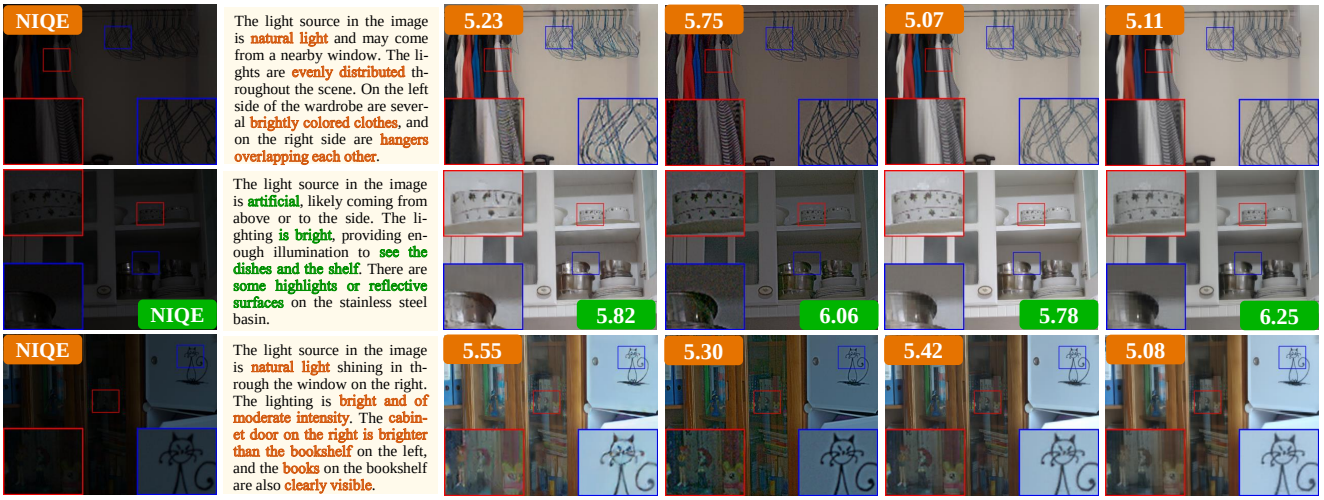
(a) Input (b) k=1 **(c) k=2 (Ours)** (d) k=3

Figure 18. **Visual comparison of the iterative instruction strategy** on real-world datasets about DICM [8] (upper), NPE [17] (middle), and VV [16] (lower).

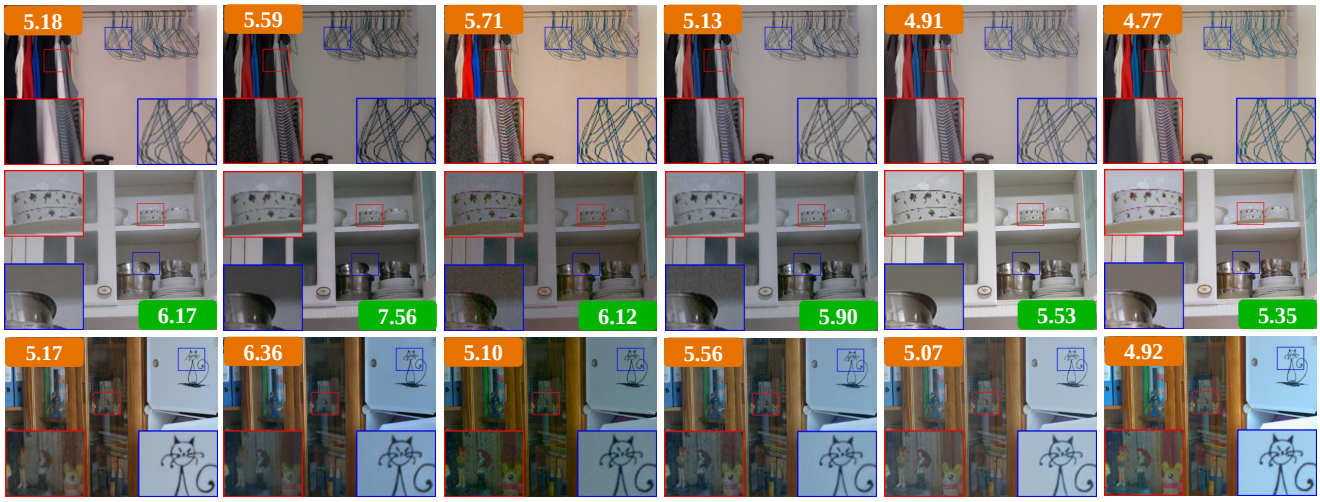


(a) Input (b) Tab. 6(a) (c) Instructions (d) Tab. 6(b) (e) Tab. 6(c) (f) Instructions **(g) Tab. 6(d)**

Figure 19. **Visual comparison about failure cases** of instruction from LLaVA.



(a) Input (b) Instructions (c) NeRCo (d) CLIP-LIT (e) DiffLL (f) Diff-Plugin



(g) GSAD (h) QuadPrior (i) GPP-LLIE (j) GEFU **(k) Ours** (l) Ground Truth

Figure 20. **Visual comparison** on the LOL [20] dataset. Our VLM-IMI produces more realistic results with sharper structures and textures.

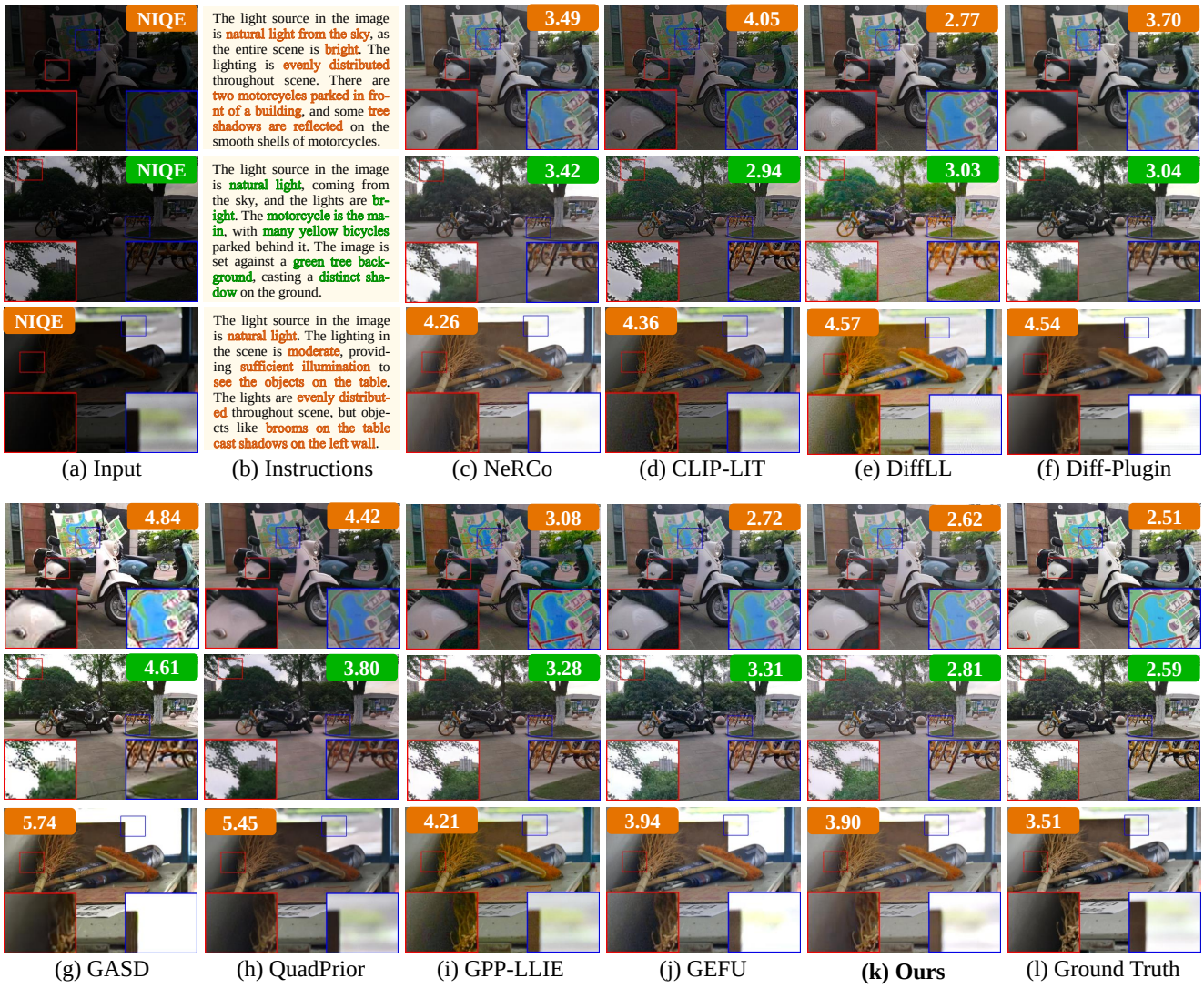
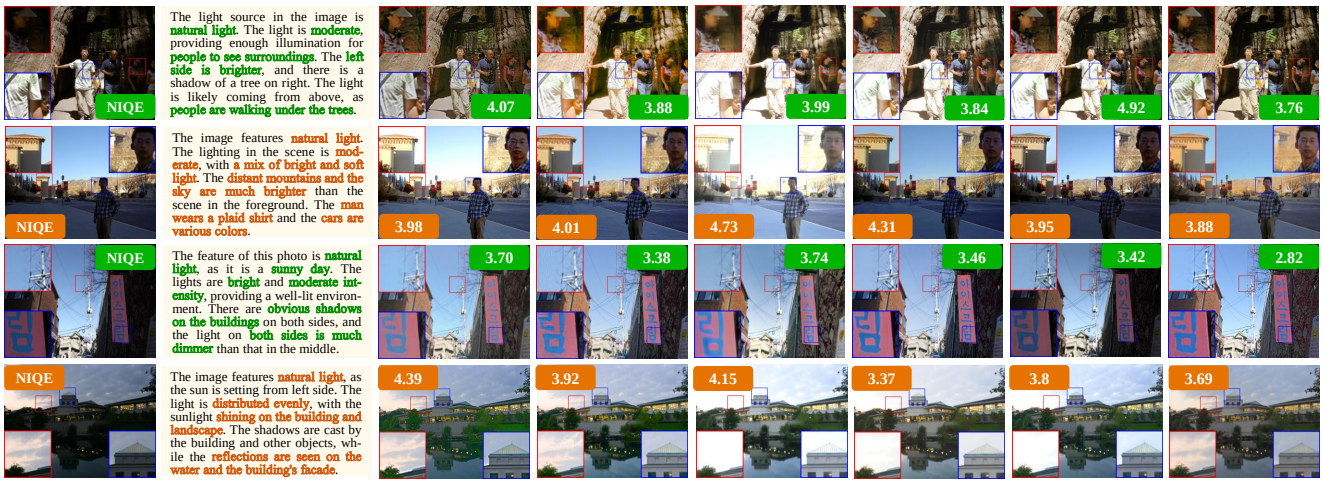
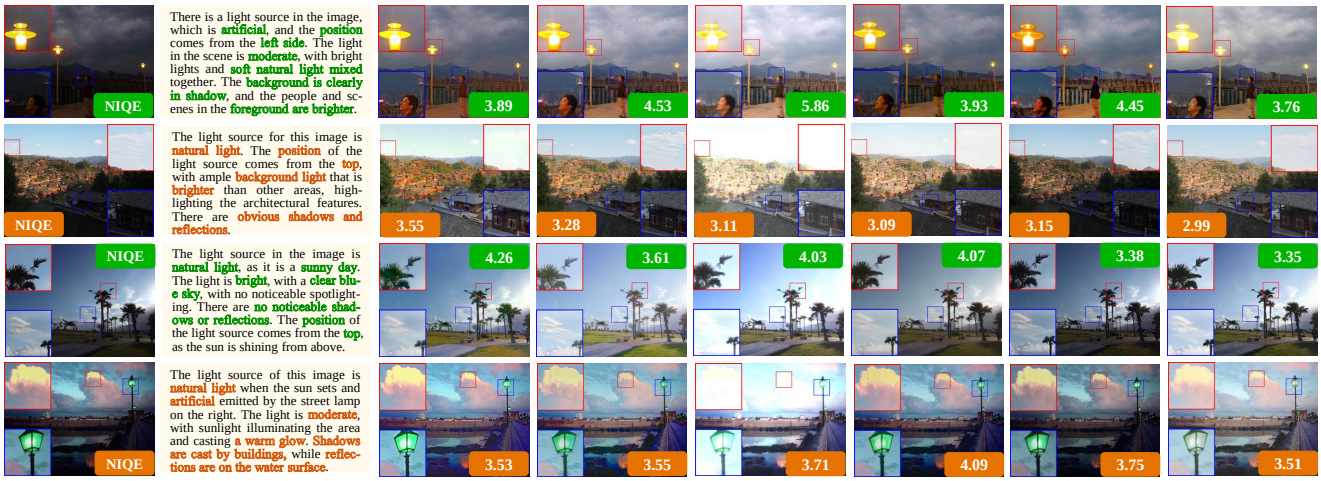


Figure 21. **Visual comparison** on the LSRW [5] dataset. Our VLM-IMI generates more realistic results with sharper structures and textures.



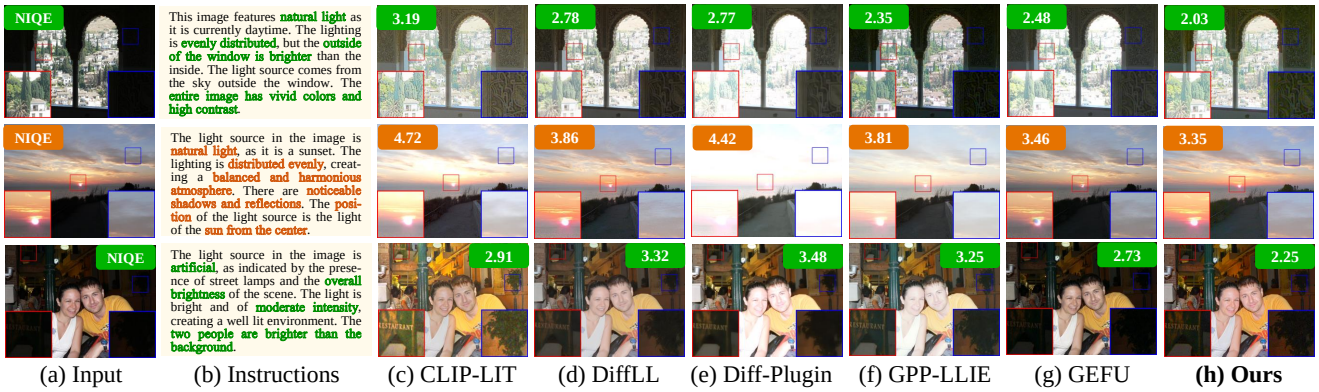
(a) Input (b) Instructions (c) CLIP-LIT (d) DiffLL (e) Diff-Plugin (f) GPP-LLIE (g) GEFU **(h) Ours**

Figure 22. **Visual comparison of real-world images** on the DICM [8] dataset. Our VLM-IMI generates results with better naturalness.



(a) Input (b) Instructions (c) CLIP-LIT (d) DiffLL (e) Diff-Plugin (f) GPP-LLIE (g) GEFU **(h) Ours**

Figure 23. **Visual comparison of real-world images** on the NPE [17] dataset. Our VLM-IMI generates results with better naturalness.



(a) Input (b) Instructions (c) CLIP-LIT (d) DiffLL (e) Diff-Plugin (f) GPP-LLIE (g) GEFU **(h) Ours**

Figure 24. **Visual comparison of real-world images** on the VV [16] dataset. Our VLM-IMI generates results with better naturalness.

References

[1] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023. 7

[2] Marcos V. Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *ECCV*, 2024. 7

[3] Huiyu Duan, Xiongkuo Min, Sijing Wu, Wei Shen, and Guangtao Zhai. Uniprocessor: A text-induced unified low-level image processor. In *ECCV*, 2024. 7

[4] Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *CVPR*, 2023. 7

[5] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *JVCIR*, 2023. 3, 5, 7, 10

[6] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *NeurIPS*, 2023. 7

[7] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM TOG*, 2023. 7

[8] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE TIP*, 2013. 3, 5, 6, 7, 8, 11

[9] JDMCK Lee and K Toutanova. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 5

[10] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *ICCV*, 2023. 7

[11] Yuhao Liu, Zhanghan Ke, Fang Liu, Nanxuan Zhao, and Rynson W. H. Lau. Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In *CVPR*, 2024. 6, 7

[12] 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 *ICML*, 2021. 2, 5

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 2, 5

[14] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 1, 2

[15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 5

[16] Vassilios Vonikakis, Rigas Kouskouridas, and Antonios Gasteratos. On the evaluation of illumination compensation algorithms. *MTAP*, 2018. 4, 5, 6, 7, 8, 11

[17] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE TIP*, 2013. 4, 5, 6, 7, 8, 11

[18] Sen Wang, Shao Zeng, Tianjun Gu, Zhizhong Zhang, Ruixin Zhang, Shouhong Ding, Jingyun Zhang, Jun Wang, Xin Tan, Yuan Xie, and Lizhuang Ma. From enhancement to understanding: Build a generalized bridge for low-light vision via semantically consistent unsupervised fine-tuning. In *ICCV*, 2025. 7

[19] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *CVPR*, 2024. 7

[20] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 2, 5, 7, 9

[21] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *ICCV*, 2023. 7

[22] Han Zhou, Wei Dong, Xiaohong Liu, Yulun Zhang, Guangtao Zhai, and Jun Chen. Low-light image enhancement via generative perceptual priors. In *AAAI*, 2025. 7