

Multimodal Large Language Model-Guided ISP Hyperparameter Optimization with Dynamic Preference Learning

Supplementary Material

Xinyu Sun^{1,2}, Zhikun Zhao^{1,2}, Congyan Lang^{1,†}, Bing Li^{2,3,4}, Juan Wang^{2,3}

1 School of Computer Science and Technology, Beijing Jiaotong University

Key Laboratory of Big Data & Artificial Intelligence in Transportation (Ministry of Education)

2 State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), CASIA

3 Beijing Key Laboratory of Super Intelligent Security of Multi-Modal Information, CASIA

4 PeopleAI Inc., Beijing

{xinyusun, zhikunzhao, cylang}@bjtu.edu.cn {bli@nlpr, jun.wang}@ia.ac.cn

This supplement adds more details about the Image Signal Processing (ISP) pipeline, the implementation details, more supplementary experiments and text information visualization to further complement the manuscript.

A. Image Signal Processing Pipeline

The hardware ISP consists of a series of image processing modules. These modules process in sequence to further optimize the image. We show the typical stages included in the synthetic (simulated) ISP:

(1) Optics and Sensor: The optical system directs incoming light onto the image sensor, which captures the scene and produces RAW images in Bayer format. The sensitivity and dynamic range of sensor play a crucial role in determining the quality of the captured image.

(2) Noise Reduction: Noise, often modeled as a random variation with zero mean, degrades image quality. Denoising [14] techniques help suppress unwanted noise but may also blur fine details, requiring a balance between noise removal and preserving image sharpness.

(3) White Balance: RAW pixel values undergo color correction and gain adjustments using predefined white balance [7] matrices for standard illuminations or dynamically estimated lighting conditions. It ensures color consistency and accuracy across different lighting environments.

(4) Demosaicking: Since Bayer sensors capture only one color per pixel, demosaicking [2] algorithms interpolate missing color values to reconstruct a full RGB image. Effective interpolation techniques are crucial for minimizing artifacts such as color fringing and moiré patterns.

(5) Color Space Transformation [12]: White-balanced RAW-RGB values are mapped to the standardized CIE XYZ color space using a 3×3 transformation matrix. This step ensures color fidelity and facilitates further color processing and rendering.

(6) Sharpening: Edge detection techniques identify significant structures in the image, and specialized filters enhance edge contrast to improve perceived sharpness. Proper sharpening [6] improves detail visibility but must be carefully

fully tuned to avoid over-enhancement artifacts.

(7) Color and Tone Correction [8]: The image’s overall appearance is refined by applying gamma correction and contrast adjustments. Histogram-based techniques help optimize brightness and tonal distribution, ensuring a visually appealing result.

The main modules of the synthetic ISP are as shown above. Qualcomm Spectra 580 ISP is a commercial ISP. In this paper, we mainly optimize its 114 ISP hyperparameters for IQA. Its processing stage includes noise reduction, OpticalFlow, sharpening, etc. Considering its commercial and special nature, we will not introduce each processing module in detail.

B. Implementation Details

B.1. Evaluation Metrics

We give a detailed introduction to the evaluation metrics of the downstream tasks involved in the manuscript, including mAP, SSIM, PSNR, etc.

B.1.1. mean Average Precision (mAP)

mAP is a common evaluation metric for object detection and segmentation tasks. It calculates the average precision (AP) for each class and then takes the mean over all classes. The precision-recall curve is used to compute AP, and different IoU thresholds (e.g., mAP0.5, mAP0.75) can be used to evaluate the model’s performance at different levels of localization accuracy. The formula is as follows:

• **Precision:**

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

• **Recall:**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

• **Average Precision (AP):**

$$\text{AP} = \int_0^1 \text{Precision}(r) dr \quad (3)$$

Table 2. Comparison of results on object detection between models trained on instance segmentation and models trained on object detection.

Methods	Model (Instance Segmentation)			Model (Object Detection)		
	mAP@0.5	mAP@0.75	mAP@0.5:0.95	mAP@0.5	mAP@0.75	mAP@0.5:0.95
YOLOv3 [10]	0.617	0.435	0.421	0.653	0.482	0.454
YOLOv5 [5]	0.634	0.467	0.435	0.668	0.509	0.480
YOLOv8 [13]	0.641	0.493	0.459	0.675	0.536	0.513

- **mAP:**

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \text{AP}_i, \quad (4)$$

where N is the number of classes, and AP_i is the average precision for the i -th class.

B.1.2. Structural Similarity Index Measure (SSIM)

SSIM [4] is a metric in Image Quality Assessment (IQA) to measure the similarity between two images. It considers luminance, contrast, and structure to provide a perceptual quality score. SSIM values range from -1 to 1 , where 1 indicates perfect similarity. The SSIM is defined as:

- **SSIM:**

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (5)$$

where μ_x and μ_y are the mean intensities of images x and y , σ_x^2 and σ_y^2 are their variances, and σ_{xy} is the covariance between x and y . The constants C_1 and C_2 are used to stabilize the division.

B.1.3. Peak Signal-to-Noise Ratio (PSNR)

PSNR [11] is a metric used to measure the quality of reconstructed or compressed images compared to the original image. It is based on the Mean Squared Error (MSE) [1] between the two images. Higher PSNR values indicate better quality. The formulation is as follows:

- **MSE:**

$$\text{MSE} = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (x(i, j) - y(i, j))^2, \quad (6)$$

where M and N are the dimensions of the image, and $x(i, j)$ and $y(i, j)$ are the pixel values of the original and reconstructed images, respectively.

- **PSNR:**

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right), \quad (7)$$

where MAX_I is the maximum possible pixel value.

B.2. Experiment Settings

Our method is trained on four NVIDIA GeForce RTX 3090 GPUs. The batch size is set to 4. The agent is trained using the Adam optimizer with a learning rate of 10^{-5} . Our models are trained about 10,000 iterations in object detection and instance segmentation, about 5,000 iterations in IQA. After that, we iterate 5,000 times on each task using our proposed dynamic pair generation refinement strategy based on DPO to obtain the final results. The values of some hyperparameters in our method are as follows: The discount factor γ is set to 0.99. The clipping parameter ϵ is set to 0.2. The hyperparameter η that controls the DPO loss and strategy optimization loss is set to 0.1. The hyperparameter β that controls the sensitivity of preference score differences is set to 1.

C. Supplementary Experiments

C.1. Ablation study on downstream task models (Instance Segmentation)

For instance segmentation, we also conduct experiments based on different downstream pretrained models. For a fair comparison with other approaches, we adopt the same pre-trained Mask R-CNN model [3] as the instance segmentation tool. Additionally, we test our results using pre-trained SOLO [15] model. The experimental results are shown in Table 1. The results based on SOLO outperforms those of the Mask R-CNN model. The experiment demonstrates that the performance of different pre-trained downstream models has a significant impact on the final results.

Table 1. Ablation studies for downstream task models on instance segmentation.

Methods	mAP0.5	mAP0.75	mAP0.5:0.95
Mask R-CNN [3]	0.632	0.401	0.373
SOLO [15]	0.653	0.426	0.401





	A man is fixing a motorcycle tire in the foreground, while a motorcycle and a bicycle are parked in the background.	A man in a green shirt and black pants is working on a motorcycle wheel.	A man in a green shirt and dark pants crouching on the ground while repairing a motorcycle wheel.
	A man sitting on a train with a laptop in front of him.	A man in a suit is sitting on a train and using a laptop.	A man with red hair and a dark suit, is working on a laptop.
	Two people are flying a large kite with a rocket design in a grassy field.	A group of people are flying kites in a park.	People in a park flying a colorful kite with a dragon-like design.
	A bus stop with a yellow bus parked in front of a building.	A bus stop with a bus parked on the side of the road.	A yellow bus parked on the side of the road. Another white bus was moving.
Images	Llava-7B	mPLUG-Owl-2	GPT-4V

Figure 1. Text visualization based on image semantics generated by MLLMs. The left column is the image information, and the three columns on the right are text descriptions generated based on different MLLMs.


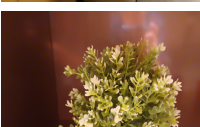
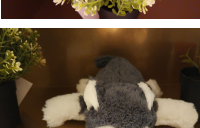
	The overall clarity of this image is very low. The main subject, the chair, has lost most of its texture details. The background is blurry.	The overall image quality is average. The table, chairs, and plants can be seen, and it is a well-composed image.	The image has moderate sharpness. The noise is present but not extreme. It is somewhat dark, which could reduce visibility.
	The main subject, the flower, retains most of its texture details. The background wall is slightly blurred, and the lighting is average.	The overall scene appears to be well-lit. The quality of the image is good, as it captures the beauty and details of the flowers and vase.	The image has high sharpness, indicating good detail, and moderate noise. Overall, decent image quality.
	The clarity of this image is acceptable. The texture details of the subject, a small dog, are rich and clear. The colors are also rich.	The overall quality of the image is good, with the stuffed dog and the potted plant being the main subjects.	The image has good sharpness but could be improved for finer details. Noise is slightly noticeable but not excessive.
Images	Llava-7B	mPLUG-Owl-2	GPT-4V

Figure 2. Text visualization based on low-level image quality by MLLMs. The left column is the image information, and the three columns on the right are text descriptions generated based on different MLLMs.

C.2. Verification of the generalization of the model in different tasks

To demonstrate the generalization of our method to other tasks, we also test results in the object detection using test images generated by the instance segmentation model. It should be noted that this model is an ISP hyperparameter optimization model based on instance segmentation trained on the synthetic COCO dataset using the pre-trained Mask-RCNN Segmenter. The experimental results are shown in Table 2. Although the detection results of images gener-

ated by the instance segmentation model are not as good as those generated by the detection model, they still outperform some other methods. It indicates the generalization and effectiveness of our model in optimizing ISP.

C.3. Visualization and analysis of text descriptions based on different tasks

C.3.1. Text description based on image semantics

To further compare the differences in text descriptions generated by different multimodal large language models

(MLLMs), we output the text content during the hyperparameter optimization process. The results are shown in Figure 1. It presents the text content generated by different models regarding image semantics in high-level vision tasks (*i.e.*, object detection and instance segmentation). The text generated by Llava-7B [16] focuses more on the instances present in both the foreground and background of the image. In contrast, the text generated by mPLUG-Owl-2 [16] and GPT-4V [9] pays more attention to the main characters or objects in the image and their details. For example, when describing the man in the first image, they include details such as his green T-shirt and dark pants. Since object detection and instance segmentation tasks focus more on the overall object rather than specific details, the experimental results in the main manuscript show that the text generated by the Llava-7B performs better.

C.3.2. Text description based on the low-level quality of the image

In addition to comparing the text descriptions of image semantics generated for the detection and segmentation tasks, we also present the text descriptions generated by different models regarding low-level image quality in IQA task. The results are shown in Figure 2. The low-level image quality descriptions generated by Llava-7B are more consistent with the actual quality of the images. In contrast, mPLUG-Owl-2 and GPT-4V struggle to distinguish between images with severe blurring and noise. For example, the first image has poor sharpness and contains a large number of blurry areas. However, these models rated the image quality as average, with moderate sharpness. The inconsistency between the generated text and the actual image quality may have a certain impact on the final experimental results.

C.4. Additional analysis on DPGM

Experiments show that the two agents produce different actions for the same input during the fine-tuning stage. These agents are frozen at this stage. To verify the differences and effectiveness of the outputs from two agents, we test their outputs with 1,000 images for detection and segmentation tasks, and 100 images for IQA, measuring which model’s generated images performed better. The results are shown in Table 3, which indicates that the different actions produced by two agents for the same image lead to various performance. The fine-tuning process facilitates the model to integrate these strengths and achieve further optimization.

Table 3. Number of superior images produced by agents.

	Detection	Segmentation	IQA
Agent A	236	371	61
Agent B	526	383	34
Equal	238	246	5

References

- [1] Mean Squared Error. Mean squared error. *MA: Springer US*, 5, 2010. 2
- [2] Bahadır K Gunturk, John Glotzbach, Yucel Altunbasak, Ronald W Schafer, and Russel M Mersereau. Demosaicking: color filter array interpolation. *IEEE Signal processing magazine*, 22(1):44–54, 2005. 1
- [3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2
- [4] A Hore and D Ziou. Image quality metrics: Psnr vs. ssim. international conference on pattern recognition. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 2366–2369. 2
- [5] Glenn Jocher, Alex Stoken, Jirka Borovec, Liu Changyu, Adam Hogan, Laurentiu Diaconu, Jake Poznanski, Lijun Yu, Prashant Rai, Russ Ferriday, et al. ultralytics/yolov5: v3. 0. *Zenodo*, 2020. 2
- [6] Baruch Lev. Sharpening the intangibles edge. *Harvard business review*, 6:109–116, 2004. 1
- [7] Yung-Cheng Liu, Wen-Hsin Chan, and Ye-Quang Chen. Automatic white balance for digital still camera. *IEEE Transactions on Consumer Electronics*, 41(3):460–466, 1995. 1
- [8] Radoslaw Mantiuk, Rafal Mantiuk, Anna Tomaszewska, and Wolfgang Heidrich. Color correction for tone mapping. In *Computer graphics forum*, pages 193–202. Wiley Online Library, 2009. 1
- [9] OpenAI. Gpt-4v(ision) system card, 2023. 4
- [10] Joseph Redmon and Ali Farhadi. Yolov3: an incremental improvement. arxiv e-prints. *arXiv preprint arXiv:1804.02767*, 2018. 2
- [11] De Rosal Igantius Moses Setiadi. Psnr vs ssim: imperceptibility quality assessment for image steganography. *Multimedia Tools and Applications*, 80(6):8423–8444, 2021. 2
- [12] Min C Shin, Kyong I Chang, and Leonid V Tsap. Does colorspace transformation make any difference on skin detection? In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pages 275–279. IEEE, 2002. 1
- [13] Mupparaju Sohan, Thotakura Sai Ram, Rami Reddy, and Ch Venkata. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*, pages 529–545. Springer, 2024. 2
- [14] Saeed V Vaseghi. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008. 1
- [15] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 649–665. Springer, 2020. 2
- [16] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25490–25500, 2024. 4