End-to-End RAW Synergy for Elevated Vision-Language Reasoning

Kepeng Xu , Tong Qiao , Zhenyang Liu , Li Xu , Gang He*
Xidian University
kepengxu11@gmail.com

Abstract

Visual Language Models (VLMs) typically rely on processed RGB images, leading to information loss that limits performance in challenging scenes like low-light or high dynamic range. Traditional Image Signal Processing (ISP) pipelines, optimized for human perception, discard crucial raw sensor data beneficial for machine understanding. To overcome this, we introduce Raw-VLM, an end-to-end model that enables VLMs to natively interpret raw image sensor data. Raw-VLM integrates a learnable ISP (GM-ISPNet) and a Raw-Tokenlizer module within its vision encoder (Raw-ViT). This differentiable frontend is jointly trained with the VLM, adaptively converting raw Bayer patterns into machine-centric representations that preserve vital semantic features while suppressing noise. Our approach addresses the information bottleneck, modality mismatch, and task-agnostic limitations of conventional RGB-based VLMs. Raw-VLM significantly improves performance on tasks such as raw image captioning (9% gain), visual question answering (5.4% gain), and reduces hallucinations (3.02% gain on POPE). By directly leveraging raw data, Raw-VLM enhances VLM capabilities in difficult scenarios, bridging the gap between sensor data and high-level semantic understanding.

1 Introduction

Image processing and analysis have long been fundamental tasks in computer vision. With the advent of large-scale datasets and the rapid development of deep learning techniques [LeCun *et al.*, 2015], significant progress has been made in various tasks such as object detection, segmentation, image restoration, and generation. However, most publicly available datasets consist of RGB images that have undergone compression and post-processing through the camera's image signal processing (ISP) pipeline. This process inevitably discards large amounts of original sensor information, which becomes a bottleneck for fine-grained visual analysis tasks.

*Corresponding author

In contrast to common compressed formats such as JPEG and PNG, Raw images retain the full fidelity of the data captured by the camera sensor before any ISP processing. Raw images preserve higher dynamic range, better color precision, and linear light intensity, offering a more physically accurate representation of the scene. These advantages have motivated researchers to utilize Raw images for tasks such as image restoration, denoising, low-light enhancement, and object detection under adverse lighting conditions, achieving notable results.

Meanwhile, large language models (LLMs) such as GPT, Claude, and Deepseek have shown remarkable capabilities in natural language understanding and generation, driven by advances in model architecture, training techniques, and computational power. However, LLMs primarily rely on textual inputs and lack the ability to directly process visual information. To bridge this gap, vision-language models (VLMs) have emerged, integrating vision (images, videos) and language to enable multimodal semantic understanding. VLMs extract visual features using vision encoders and align them with pretrained LLMs to perform tasks such as visual question answering (VQA), image captioning, and cross-modal retrieval. These models have demonstrated impressive performance in real-world applications such as autonomous driving, surveillance, robotic perception, and medical imaging.

Despite these advances, current VLMs suffer from a fundamental limitation: the vision encoders operate on RGB images that have undergone irreversible ISP processing. Traditional ISP is designed to enhance human visual perception, often at the cost of discarding high dynamic range details and suppressing high-frequency textures through non-differentiable operations such as demosaicing and tone mapping. These losses significantly hinder the model's ability to understand challenging visual scenes, such as low-light or HDR environments. In contrast, Raw images contain rich, unprocessed information that is crucial for robust semantic interpretation in such scenarios.

To address this issue, we propose a novel end-to-end Raw image-based vision-language model, **Raw-VLM**. Our key idea is to insert a learnable ISP module (**GM-ISPNet**) and a Raw feature tokenizer (**Raw-Tokenlizer**) before the standard vision encoder, forming a fully differentiable front-end (**Raw-ViT**) that directly processes Raw images. Unlike conventional ISP, which is optimized for human perception, our

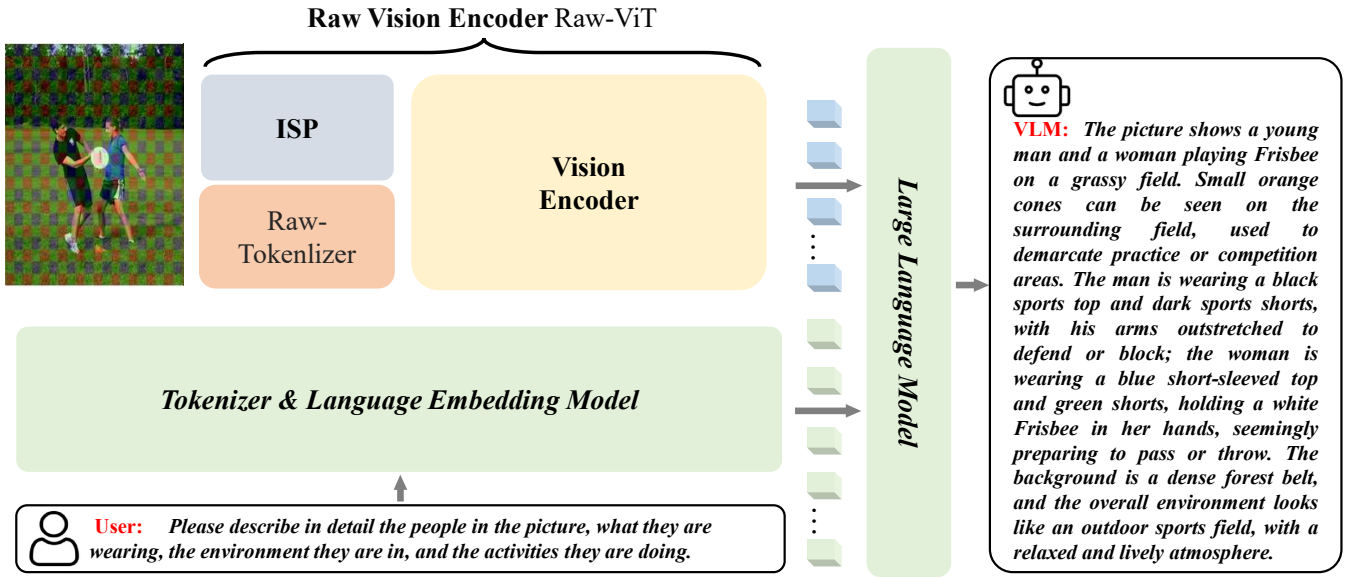

Figure 1: Framework of Raw-VLM.

learnable ISP is optimized jointly with the VLM for machine-level visual reasoning. This enables the model to preserve critical semantic features while suppressing irrelevant sensor noise, effectively transforming Raw photon data into features aligned with the language model's semantic space.

Furthermore, due to the lack of existing Raw-VLM datasets, we construct a large-scale Raw image vision-language dataset covering tasks such as Raw image captioning, Raw image OCR, and Raw image VQA. Our contributions can be summarized as follows:

- We present **Raw-VLM**, the first end-to-end vision-language model that natively processes Raw images, enabling direct semantic understanding from sensor data. We also propose a four-stage training strategy to jointly optimize the model components.
- We introduce **GM-ISPNet**, a differentiable ISP module deeply integrated with VLMs, featuring a global prior guidance mechanism and dynamic convolution-based mixture-of-experts for demosaicing, tone mapping, and denoising. It overcomes the non-differentiability and semantic limitations of traditional ISP.
- We design **Raw-Tokenlizer**, a module that extracts physically grounded visual priors from Raw data and collaborates with GM-ISPNet and the visual encoder to form the **Raw-ViT** front-end. Extensive experiments show that Raw-VLM significantly outperforms RGB-based VLMs on multiple tasks, especially in low-light and HDR scenarios.

2 Related Work

Recent advances in deep learning have led to significant breakthroughs across various domains, with notable impacts in fields like computer vision, natural language processing, and reinforcement learning[He *et al.*, 2022b; He *et al.*, 2022a; Xu and He, 2022; Xu *et al.*, 2024b; Xu *et al.*, 2024c; Xu *et al.*, 2023; Chen *et al.*, 2025; Xu *et al.*, 2025; He *et al.*, 2025; Deng *et al.*, 2025]. In particular, deep neural networks (DNNs) have enabled substantial improvements in performance, surpassing traditional algorithms in tasks such as image classification, object detection, and image generation. These achievements are driven by the increasing availability of large-scale datasets, the development of more sophisticated model architectures, and the growing computational power provided by modern GPUs and specialized hardware. Consequently, deep learning has become a cornerstone of modern artificial intelligence (AI), facilitating the creation of more intelligent systems capable of performing complex tasks that were once thought to be the domain of human expertise. In this section, we explore key developments that have propelled this rapid progress, focusing on three major areas: Image Signal Processing (ISP) pipelines, Vision-Language Models (VLMs), and the application of raw image data in computer vision tasks.

This section reviews prior works from three perspectives: (1) Image Signal Processing (ISP) pipelines, (2) Vision-Language Models (VLMs), and (3) Applications of Raw images in computer vision tasks.

2.1 Image Signal Processing Pipelines

The Image Signal Processing (ISP) pipeline [Ramanath *et al.*, 2005] transforms raw data captured by image sensors into RGB images that align with human visual perception. Traditional ISP typically involves a sequential set of operations, including black level correction, lens shading correction, bad pixel removal, denoising, white balance, demosaicing, color correction, tone mapping, gamma correction, sharpening, and compression.

Classical denoising methods in ISP include mean filtering, bilateral filtering [Tomasi and Manduchi, 1998], guided filtering [He *et al.*, 2012], non-local means, and BM3D [Dabov *et*

al., 2007]. With the rise of deep learning, CNN-based denoisers like DnCNN [Zhang *et al.*, 2017] and transformer-based models have achieved state-of-the-art performance. Brooks et al. [Brooks *et al.*, 2019] proposed an unprocessing pipeline to synthesize realistic Raw noise for learning-based denoising, followed by more accurate physical modeling [Wei *et al.*, 2020] and diffusion-based Raw denoisers [Yi *et al.*, 2024; Feng *et al.*, 2022].

White balance correction, another critical ISP task, historically relied on assumptions like Gray World and Perfect Reflector. However, these heuristics often fail in complex lighting. Deep learning-based methods [Afifi *et al.*, 2022; Afifi *et al.*, 2021] predict channel gains under diverse illumination and across sensors, improving generalization.

Recent research has moved towards end-to-end AI-based ISP networks. PyNet [Ignatov *et al.*, 2020] replaces the entire ISP using a pyramid CNN. Chen et al. [Zhang *et al.*, 2021] introduced global color mapping and flow-based alignment for Raw-to-RGB translation. Other works [A Sharif *et al.*, 2021] jointly optimize tasks like denoising and demosaicing. RealCamNet [Xu *et al.*, 2024a] introduces coordinate-aware modules for distortion correction and compression-aware mappings, making the ISP pipeline more practical for real-world deployment.

2.2 Vision-Language Models

VLMs have advanced rapidly by integrating visual understanding into large language models (LLMs). Models like ChatGPT-4V and LLaVA have enabled impressive capabilities in image captioning, visual question answering (VQA), and visual dialog.

VisualBERT [Li *et al.*, 2019] was among the first to align image regions and text using a unified Transformer. CLIP [Radford *et al.*, 2021] and DALL-E [Ramesh *et al.*, 2021] introduced contrastive and generative learning paradigms with massive multi-modal datasets. CLIP employs ViT [Dosovitskiy *et al.*, 2020] or CNN as image encoders and aligns them with text embeddings via contrastive learning.

FLAVA [Singh *et al.*, 2022] unified masked image modeling, masked language modeling, and contrastive learning to build representations across modalities. MiniGPT-4 [Zhu *et al.*, 2023] efficiently aligned visual and textual modalities using a frozen vision encoder and a lightweight projection layer.

LLaVA [Liu *et al.*, 2023] extended instruction tuning to VLMs by transforming image-text pairs into instruction-style prompts, training on a GPT-4V-curated dataset. MobileVLM [Chu *et al.*, 2023] optimized model architecture for mobile deployment, achieving high throughput on Snapdragon processors.

Qwen-VL [Bai *et al.*, 2023] introduced cross-attention between OpenCLIP visual features and a Qwen-based LLM decoder. Qwen2-VL [Wang *et al.*, 2024] further improved multimodal understanding with dynamic resolution and rotary position embeddings, achieving leading results and influencing the open-source VLM community.

2.3 Raw Images in Computer Vision

Raw images preserve rich sensor-level information that is often discarded by standard ISP. Early work [Zhou *et al.*, 2020; Buckler *et al.*, 2017] explored pedestrian detection directly from Raw gradient histograms, though noise modeling was often neglected, especially under low-light conditions.

The scarcity of large-scale Raw datasets (e.g., PASCAL Raw [Everingham *et al.*, 2010], LOD) compared to RGB datasets (e.g., ImageNet) has limited the development of Raw-based models. Recent trends focus on leveraging RGB-pretrained models while adapting them to Raw data.

VisionISP [Wu *et al.*, 2019] demonstrated that ISP optimized for human perception may not benefit machine vision, and proposed learnable ISP modules to enhance detection downstream. Rawgment [Yoshimura *et al.*, 2023] applied data augmentation directly in the Raw domain.

Furthermore, white balance errors and exposure control have been shown to significantly degrade performance in object detection tasks [Sayed and Brostow, 2021]. Cui et al. [Cui and Harada, 2024] proposed Raw-Adapter, a joint training framework that aligns Raw images with RGB-pretrained models, achieving state-of-the-art results on PASCAL and LOD datasets.

3 Methodology

We propose **Raw-VLM**, an end-to-end vision-language model that directly consumes Raw images for multimodal understanding. Instead of relying on fixed ISP-processed RGB images, Raw-VLM introduces a learnable image signal processing pipeline and Raw-aware feature extraction to better preserve high dynamic range and fine-grained details critical for challenging scenarios such as low-light and HDR conditions.

Figure 1 shows the overall architecture of Raw-VLM. It consists of two main components: (1) the Raw visual encoder **Raw-ViT** and (2) a large language model (LLM) decoder. The Raw-ViT module integrates a learnable ISP network (GM-ISPNet), a Raw-aware tokenizer (Raw-Tokenizer), and a vision encoder (NaViT).

3.1 Raw-ViT: Raw Visual Encoder

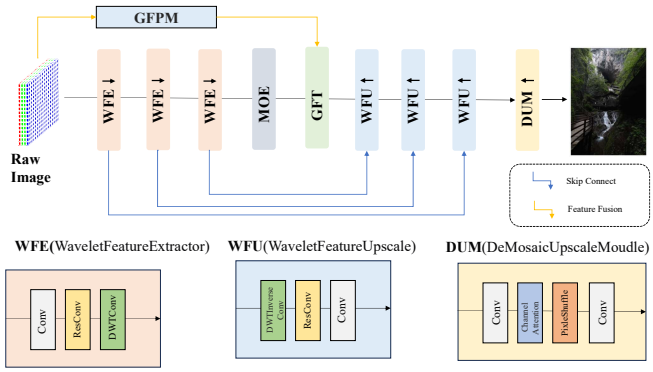

Figure 2: Framework of GM-ISPNet.

The **Raw-ViT** module transforms low-level Raw image signals into high-level semantic features. It is composed of three submodules:

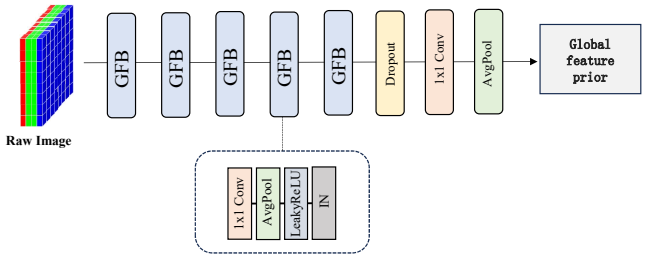

Figure 3: Global feature prior guidance module.

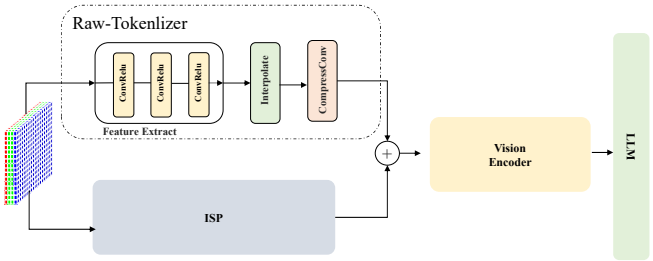

Figure 5: Framework of Raw-Tokenizer.

(1) GM-ISPNet: Learnable ISP Pipeline. To replace the fixed, human-centric ISP pipeline, we introduce GM-ISPNet, a differentiable and trainable module that maps Raw Bayer images to semantically meaningful RGB representations. It includes two novel components:

- **Global Feature Prior Module (GFPM):** Extracts global semantic, brightness, and color distribution priors from Raw images to guide RGB reconstruction, improving color fidelity and structural preservation (Fig. 3).
- **Mixture-of-Experts Module (MoE):** Combines several expert branches (both dynamic and static) using attention-based weighting, enabling adaptive feature selection based on scene content and noise profiles (Fig. 4).

(2) Raw-Tokenizer: Raw-Aware Feature Extractor. This module captures physical priors from Raw input such as linear photon response, noise distribution, and dynamic range. It aligns spatially with the ViT encoder using patch-based tokenization. The module is initialized with zero weights to ensure stable cold-start training and gradually learns to enhance semantic feature quality (Fig. 5).

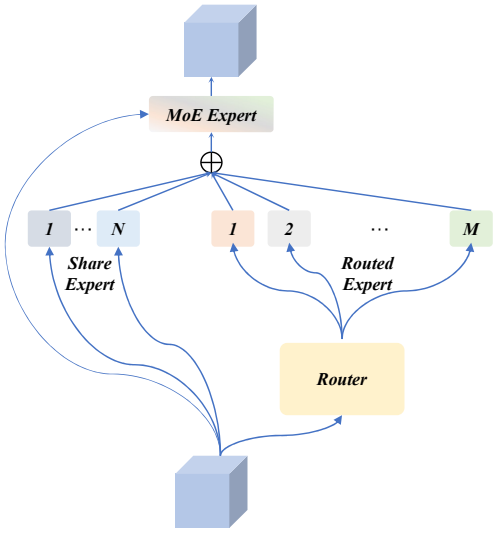

Figure 4: MoE hybrid expert feature selection module.

(3) Vision Encoder: NaViT. We adopt NaViT, a flexible ViT variant that supports variable resolution and aspect ratios using masked attention and local pooling. It avoids input distortion and improves performance on visually structured content like text and charts.

3.2 LLM Decoder and Multimodal Alignment

The output features from Raw-ViT are fed into a frozen LLM decoder (Qwen2-7B [Wang *et al.*, 2024]), which performs multimodal reasoning and language generation. The model leverages:

1. **Multimodal Semantic Bridging:** Raw visual features are aligned to the language space via cross-attention, enabling the model to generate context-aware, semantically aligned responses.
2. **Knowledge Transfer:** The pretrained LLM transfers world knowledge and commonsense reasoning to visual tasks, enhancing performance on complex VQA and descriptive prompts.

3.3 Training Strategy

To ensure effective cross-modal alignment, we propose a four-stage progressive training strategy (Fig. 6):

- **Stage 1: GM-ISPNet Pretraining.** We pretrain GM-ISPNet on a Raw-RGB paired dataset using L1 loss to learn a good initialization for Raw-to-RGB mapping.
- **Stage 2: Feature Alignment.** Freeze the Raw-Tokenizer and ViT encoder. Fine-tune GM-ISPNet to align its output distribution to the pretrained vision encoder's RGB domain.
- **Stage 3: Joint Raw-ViT Optimization.** Fine-tune GM-ISPNet, Raw-Tokenizer, and the vision encoder jointly. This enables feature fusion and improves Raw domain robustness.
- **Stage 4: LLM Instruction Tuning with LoRA.** Freeze Raw-ViT and fine-tune the LLM using Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022] to support downstream VQA and captioning tasks with minimal additional parameters.

3.4 Loss Functions

Different stages use specialized loss functions:

- **Stage 1:** Uses L1 loss $\mathcal{L}_1 = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ for Raw-to-RGB regression.

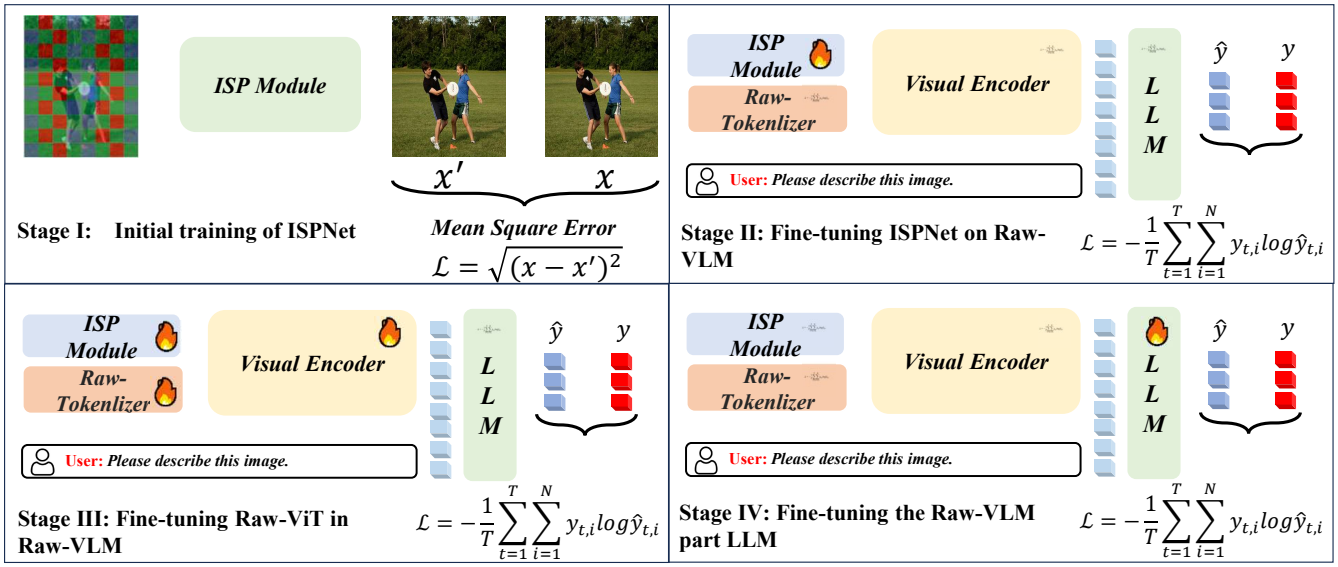

Figure 6: Four training stages of Raw-VLM.

- **Stages 2–4:** Use standard supervised fine-tuning with cross-entropy loss:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} -\log P(y_i|x_i;\theta)$$

where (x_i, y_i) are instruction-answer pairs.

3.5 Raw-VLM Dataset Construction

To support training and evaluation, we construct a large-scale Raw-VLM dataset by back-converting high-quality RGB images into Raw format using the unprocessing pipeline [Brooks *et al.*, 2019]. To simulate realistic degradation, we apply:

1. **Brightness Attenuation:** Multiply Raw values by a scalar $r \in [0.1, 0.3]$.
2. **Poisson-Gaussian Noise:** Model shot noise and read noise as:

$$\text{variance} = I \cdot \text{ShotNoise} + \text{ReadNoise}$$

and generate samples from $\mathcal{N}(0, \sqrt{\text{variance}})$.

The dataset includes 340,873 image-text pairs across 8 diverse VLM tasks: COCO-Caption, VG, GQA, TextVQA, OCR-VQA, A-OKVQA, ScienceQA, and ChartQA. Each sample includes a degraded Raw image and a corresponding textual annotation, enabling evaluation of Raw-VLM under challenging conditions (Fig. 7).

3.6 Evaluation Protocol

We adopt standard VLM benchmarks spanning:

- **Answer Matching:** Evaluate open-ended predictions using exact match, BLEU, and ROUGE. For long answers, we use LLM-based scoring to assess semantic equivalence.
- **Multiple Choice:** Select the correct answer from distractors; accuracy is computed directly.
- **Hallucination Detection:** Assess alignment between visual input and textual output using contradiction detection metrics.

4 Experiments

In this section, we evaluate the proposed **GM-ISPNet** and **Raw-VLM** across multiple benchmarks and application scenarios. We first present experiments for GM-ISPNet on the Raw-to-RGB reconstruction task, including quantitative comparison, qualitative results, and ablation studies. Then, we validate the effectiveness of Raw-VLM on various vision-language tasks using Raw images, including zero-shot and hallucination evaluations.

4.1 Experimental Setup

Hardware and Environment. All experiments are implemented using PyTorch 2.6.0 with CUDA 11.8 and Python 3.10. Training is conducted on an Ubuntu 20.04 server with Intel Xeon Gold 6148 CPUs (80 cores), 512GB RAM, and dual NVIDIA GeForce RTX 4090 GPUs.

Datasets. For Raw-to-RGB reconstruction, we use the Raw-RGB dataset described in Sec. 3.7, containing 613,945 training and 20,000 testing pairs. Input resolution is resized to 512×512. For vision-language evaluation, we use the Raw-VLM dataset (Sec. 3.7), comprising 340,873 image-text pairs across image captioning, VQA, and knowledge-based VQA.

4.2 GM-ISPNet: Raw-to-RGB Reconstruction

Quantitative Results. Table 1 compares GM-ISPNet with state-of-the-art end-to-end ISP methods. Our model achieves the best performance across PSNR, SSIM, FSIM, LPIPS,

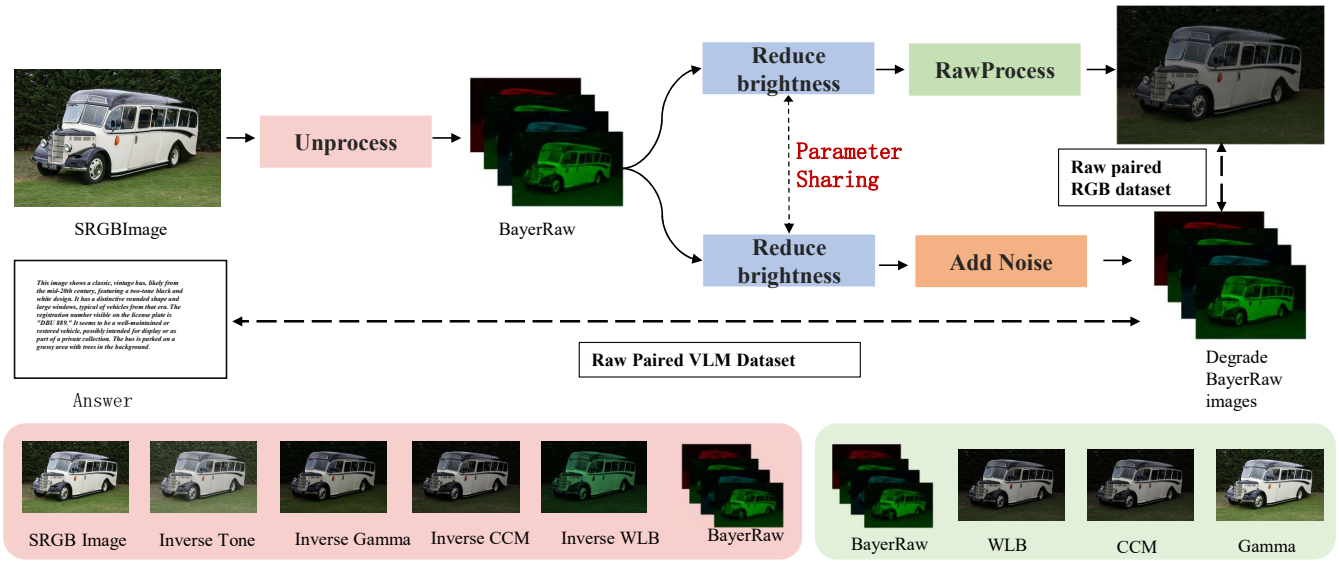

Figure 7: Raw-VLM dataset production.

DISTS, DeltaE, and MDSI metrics, outperforming LiteISP-Net by +1.52 dB in PSNR and achieving a new state-of-the-art score of 38.24 dB.

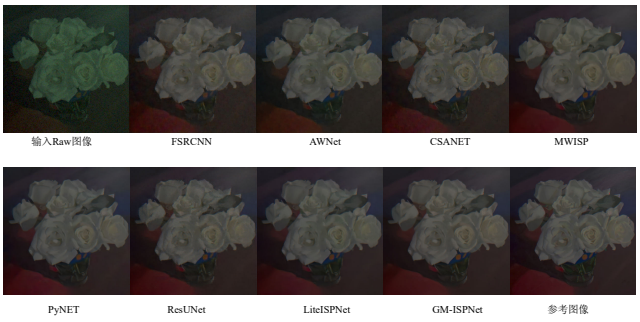

Figure 8: GM-ISPNet qualitative comparison results.

Qualitative Results. Figures 8 show visual comparison results. GM-ISPNet produces sharper textures, more natural color transitions, and better denoising, especially in low-light and high dynamic range regions. It also outperforms LiteISP and PyNet in color fidelity and detail preservation.

4.3 Raw-VLM: Vision-Language Understanding

Quantitative Evaluation. We evaluate Raw-VLM on multiple VLM benchmarks: ChartQA, ScienceQA, TextVQA, OCR-VQA, LLaVA-Bench, and POPE hallucination detection. We compare Raw-VLM with 7 baselines formed by combining existing ISP models (e.g., FSRCNN, PyNet, LiteISP) with Qwen2VL-7B.

Tables 3 show that Raw-VLM consistently outperforms all baselines. Notably, it improves ChartQA by +4.2 points and achieves higher POPE and LLaVA-Bench scores, indicating better factuality and reduced hallucination.

Zero-Shot Generalization. To test generalization, we evaluate Raw-VLM on Raw-DocVQA and Raw-InfoVQA without any fine-tuning. Table 4 shows Raw-VLM achieves the best zero-shot accuracy, demonstrating strong transferability and semantic alignment capabilities.

4.4 Ablation Study for Raw-VLM

To validate the contribution of individual components and training stages, we conduct a comprehensive ablation study on Raw-VLM using ChartQA, TextVQA, and POPE metrics.

Results in Table 2 show that:

- Adding the MoE module improves fine-grained feature extraction in challenging regions.
- Global priors from GFPM enhance color consistency and scene-level understanding.
- Progressive training (stages 1–4) significantly improves performance and stability.
- Raw-Tokenizer enhances representation by incorporating physical priors.
- LoRA fine-tuning improves LLM reasoning ability with minimal additional parameters.

Conclusion of Results. Raw-VLM achieves superior performance on both quantitative and qualitative metrics across multiple vision-language tasks. It effectively leverages Raw image information and demonstrates strong generalization, hallucination resistance, and task adaptability compared to traditional ISP+VLM pipelines.

5 Conclusion

In this paper, we propose **Raw-VLM**, the first end-to-end vision-language model capable of reasoning directly from Raw images. To bridge the gap between sensor-level data and semantic understanding, we introduce a learnable ISP module (GM-ISPNet) and a Raw-aware tokenizer, forming the Raw-ViT encoder that integrates seamlessly with large language models. We also construct a large-scale synthetic Raw-VLM

Table 1: Quantitative comparison results of different RAW2RGB methods on the test dataset

Method	PSNR↑	SSIM↑	FSIM↑	LPIPS↓	DISTS↓	DeltaE↓	MDSI↓
FSRCNN(ECCV'17)	34.3441	0.9166	0.9623	0.2618	0.1432	4.6904	0.2493
PyNet(CVPR'20)	35.7916	0.9463	0.9756	0.2055	0.0896	4.1979	0.2249
AWNet(CVPR'21)	33.6232	0.9239	0.9680	0.2595	0.1298	5.6193	0.2467
CSANet(CVPR'22)	34.9142	0.9301	0.9706	0.2218	0.1212	4.5287	0.2379
MwISP(CVPR'21)	35.3660	0.9431	0.9734	0.2019	0.0937	4.1157	0.2288
ResUNet(CVPR'21)	35.7885	0.9338	0.9705	0.2281	0.1129	3.7302	0.2338
LiteISPNet(ICCV'21)	36.7222	0.9526	0.9774	0.1617	0.0828	3.5090	0.2130
GM-ISPNet(Ours)	**38.2424**	**0.9626**	**0.9819**	**0.1291**	**0.0678**	**3.1767**	**0.2005**

Table 2: Ablation experiment quantitative results table

Method	TextVQA↑	ChartQA↑	POPE↑
Basic model + Qwen2VL-7B	74.682	69.04	85.193
Basic model + MoE module + Qwen2VL-7B	74.924	69.78	85.423
Raw-VLM first stage	74.917	71.12	85.319
Raw-VLM second stage	74.992	71.42	85.518
Raw-VLM third stage	75.012	72.119	86.142
Raw-VLM Phase 4	75.248	73.24	87.039

dataset to facilitate training and evaluation. Extensive experiments demonstrate that Raw-VLM significantly outperforms RGB-based VLM pipelines, especially under low-light and noisy conditions. Furthermore, ablation studies confirm the effectiveness of each component and our progressive training strategy. Our work highlights the potential of leveraging Raw data for robust multimodal reasoning and opens new directions for vision-language research beyond conventional RGB inputs.

References

[A Sharif *et al.*, 2021] SM A Sharif, Rizwan Ali Naqvi, and Mithun Biswas. Beyond joint demosaicking and denoising: An image processing pipeline for a pixel-bin image sensor. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 233–242, 2021.

[Afifi *et al.*, 2021] Mahmoud Afifi, Jonathan T Barron, Chloe LeGendre, Yun-Ta Tsai, and Francois Bleibel. Cross-camera convolutional color constancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1981–1990, 2021.

[Afifi *et al.*, 2022] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Auto white-balance correction for mixed-illuminant scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1210–1219, 2022.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arxiv 2023. *arXiv preprint arXiv:2308.12966*, 1(8), 2023.

[Brooks *et al.*, 2019] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11036–11045, 2019.

[Buckler *et al.*, 2017] Mark Buckler, Suren Jayasuriya, and Adrian Sampson. Reconfiguring the imaging pipeline for computer vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 975–984, 2017.

[Chen *et al.*, 2025] Zheng Chen, Jingkai Wang, Kai Liu, Jue Gong, Lei Sun, Zongwei Wu, Radu Timofte, Yulun Zhang, Jianxing Zhang, Jinlong Wu, et al. Ntire 2025 challenge on real-world face restoration: Methods and results. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1536–1547, 2025.

[Chu *et al.*, 2023] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.

[Cui and Harada, 2024] Ziteng Cui and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images. In *European Conference on Computer Vision*, pages 37–56. Springer, 2024.

[Dabov *et al.*, 2007] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[Deng *et al.*, 2025] Shangwei Deng, Qianwen Ma, Bincheng Li, Liaoran Jin, Kepeng Xu, Shangqi Deng, Xiaobo Li, and Haofeng Hu. Iovarnet: Inner-outer variation synergy

Table 3: Quantitative comparison of VLM performance on multiple benchmarks

Method	**ChartQA-Test↑**	**ScienceQA-TEST↑**	**TextVQA↑**	**LLaVA-Bench↑**	**POPE-Score↑**	**OCRBench↑**
FSRCNN + Qwen2VL-7B	67.84	0.840	72.206	76.7	84.240	629
PyNet + Qwen2VL-7B	66.80	0.843	73.304	11.0	84.487	669
MwISP + Qwen2VL-7B	67.80	0.840	73.918	78.3	84.780	650
AwNet + Qwen2VL-7B	69.52	0.846	73.112	74.5	84.442	689
ResUNet + Qwen2VL-7B	69.00	0.844	73.480	26.7	84.028	663
CSANet + Qwen2VL-7B	69.32	0.837	73.300	76.8	83.989	649
LiteISP + Qwen2VL-7B	69.04	0.845	74.682	75.8	85.192	664
Raw-VLM (Ours)	**73.24**	**0.863**	**75.248**	**78.5**	**87.039**	**721**

Table 4: VLM zero-shot performance quantitative comparison results table on different test data sets

Method	DocVQA↑	Infovqa↑
FSRCNN + Qwen2VL-7B	72.7	68.854
PyNet + Qwen2VL-7B	73.6	69.268
MwISP + Qwen2VL-7B	73.4	69.541
AwNet + Qwen2VL-7B	73.1	70.055
ResUNet+ Qwen2VL-7B	72.9	69.207
CSANet + Qwen2VL-7B	73.6	69.546
LiteISP+ Qwen2VL-7B	73.8	69.947
Raw-VLM(Ours)	**74.6**	**72.423**

network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[Feng *et al.*, 2022] Hansen Feng, Lizhi Wang, Yuzhi Wang, and Hua Huang. Learnability enhancement for low-light raw denoising: Where paired real data meets noise modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1436–1444, 2022.

[He *et al.*, 2012] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012.

[He *et al.*, 2022a] Gang He, Kepeng Xu, Chang Wu, Zijia Ma, Xing Wen, and Ming Sun. Hybrid video coding scheme based on vvc and spatio-temporal attention convolution neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1791–1794, 2022.

[He *et al.*, 2022b] Gang He, Kepeng Xu, Li Xu, Chang Wu, Ming Sun, Xing Wen, and Yu-Wing Tai. Sdrtv-to-hdrtv via hierarchical dynamic context feature mapping. In *Proceedings of the 30th ACM international conference on multimedia*, pages 2890–2898, 2022.

[He *et al.*, 2025] Gang He, Siqi Wang, Kepeng Xu, and Lin Zhang. Realrep: Generalized sdr-to-hdr conversion with style disentangled representation learning. *arXiv preprint arXiv:2505.07322*, 2025.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Ignatov *et al.*, 2020] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 536–537, 2020.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Radford *et al.*, 2021] 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. pages 8748–8763, 2021.

[Ramanath *et al.*, 2005] Rajeev Ramanath, Wesley E Snyder, Youngjun Yoo, and Mark S Drew. Color image processing pipeline. *IEEE Signal processing magazine*, 22(1):34–43, 2005.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. pages 8821–8831, 2021.

[Sayed and Brostow, 2021] Mohamed Sayed and Gabriel Brostow. Improved handling of motion blur in online object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1706–1716, 2021.

[Singh *et al.*, 2022] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022.

[Tomasi and Manduchi, 1998] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv e-prints*, pages arXiv–2409, 2024.

[Wei *et al.*, 2020] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020.

[Wu *et al.*, 2019] Chyuan-Tyng Wu, Leo F Isikdogan, Sushma Rao, Bhavin Nayak, Timo Gerasimow, Aleksandar Sutic, Liron Ain-kedem, and Gilad Michael. Visionisp: Repurposing the image signal processor for computer vision applications. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4624–4628. IEEE, 2019.

[Xu and He, 2022] Kepeng Xu and Gang He. Dnas: A decoupled global neural architecture search method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1985, 2022.

[Xu *et al.*, 2023] Kepeng Xu, Li Xu, Gang He, Xianyun Wu, Zhiqiang Zhang, Wenxin Yu, and Yunsong Li. Dual inverse degradation network for real-world sdrtv-to-hdrtv conversion. *arXiv preprint arXiv:2307.03394*, 2023.

[Xu *et al.*, 2024a] Kepeng Xu, Zijia Ma, Li Xu, Gang He, Yunsong Li, Wenxin Yu, Taichu Han, and Cheng Yang. An end-to-end real-world camera imaging pipeline. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2804–2813, 2024.

[Xu *et al.*, 2024b] Kepeng Xu, Li Xu, Gang He, Wenxin Yu, and Yunsong Li. Beyond alignment: Blind video face restoration via parsing-guided temporal-coherent transformer. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, {IJCAI-24}*, pages 1489–1497, 2024.

[Xu *et al.*, 2024c] Kepeng Xu, Li Xu, Gang He, Zhiqiang Zhang, Wenxin Yu, Shihao Wang, Dajiang Zhou, and Yunsong Li. Beyond feature mapping gap: Integrating real hdrtv priors for superior sdrtv-to-hdrtv conversion. *arXiv preprint arXiv:2411.10775*, 2024.

[Xu *et al.*, 2025] Kepeng Xu, Li Xu, Gang He, Wei Chen, Xianyun Wu, and Wenxin Yu. Unleashing the potential of transformer flow for photorealistic face restoration. In *34th International Joint Conference on Artificial Intelligence*, 2025.

[Yi *et al.*, 2024] Mingxin Yi, Kai Zhang, Pei Liu, Tanli Zuo, and Jingduo Tian. Diffraw: leveraging diffusion model to generate dslr-comparable perceptual quality srgb from smartphone raw images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6711–6719, 2024.

[Yoshimura *et al.*, 2023] Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, and Takeshi Ohashi. Rawgment: Noise-accounted raw augmentation enables recognition in a wide variety of environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14007–14017, 2023.

[Zhang *et al.*, 2017] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[Zhang *et al.*, 2021] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4348–4358, 2021.

[Zhou *et al.*, 2020] Wei Zhou, Shengyu Gao, Ling Zhang, and Xin Lou. Histogram of oriented gradients feature extraction from raw bayer pattern images. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(5):946–950, 2020.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.